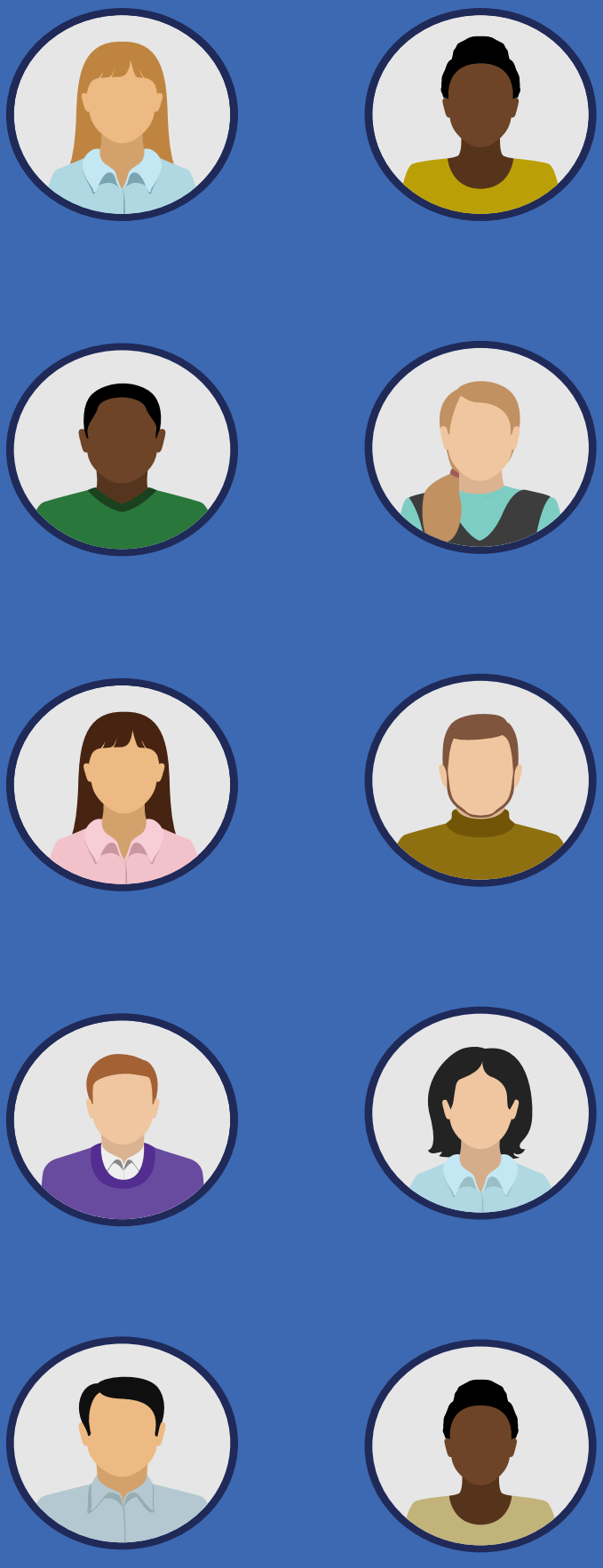


March 2022

How to de-identify personal information when sharing data

Knowledge Institute
on Child and Youth Mental Health and Addictions

Suggested citation

Knowledge Institute on Child and Youth Mental Health and Addictions. (2022). *How to de-identify personal information when sharing data*. Ottawa, ON: Knowledge Institute on Child and Youth Mental Health and Addictions. www.cymha.ca/equity

For information about this report, please contact Dr. Evangeline Danseco at edanseco@cymha.ca

Table of Contents

What is de-identification?

De-identification of data is the removal of personal identifiable information that could be used to identify individuals. This includes personal health information, among other things.

Why de-identify data?

When two or more agencies share client data containing personal information, certain legislative and ethical requirements need to be considered. De-identification ensures individuals' privacy is protected when data is shared between organizations. The organization receiving the de-identified data can analyze it and use it for strategic purposes, such as supporting decision-making or making improvements to programs and services.

This document provides suggestions on how your organization can collect, use and share data.

Personal identifiable information

"Personal identifiable information" is any information that can be used to identify individuals, either alone or in combination with other information. The types of identifiable information are grouped into two categories: 1) direct; and 2) indirect, or quasi-identifiers. Here are examples of both types.

Direct identifiers:

- Name
- Address
- Telephone number
- Email address
- Health card number
- Medical record/chart number
- IP address
- Social insurance number
- Any other unique identifying number

Indirect or quasi-identifiers:

- Date of birth
- Gender
- Postal code
- Ethnic origin
- Country of birth
- Profession
- Diagnosis
- Any unique or uncommon characteristic
- Event dates (admission, clinic visit, discharge)

Information from one indirect identifier might not be enough to identify a person (postal code, for example). However, using that identifier in combination with others, such as country of birth, might make it possible to identify an individual. It is also important to note that in some cases, even one rare indirect identifier could make it easy to identify a person. For instance, in a community with only one family that immigrated from Iceland, data on country of birth could make it possible to link a client to that family.

Step 1: Select relevant data for sharing

The data your organization chooses to receive and use should be limited to information relevant to the analysis you are undertaking. For instance, if the purpose of your agency's data analysis is focused solely on reducing wait times at walk-in clinics, any information not related to wait times — for example, health card number — should not be included in the dataset to be analyzed.

Any identifying data essential to your analysis needs to be classified into direct and quasi-identifiers for subsequent de-identification.

Step 2: Choose a de-identification framework

Two types of frameworks can be used to de-identify data: heuristic and risk-based.

- "Heuristic de-identification" applies general de-identification rules to data to remove or generalize the information.
- "Risk-based de-identification" assesses the overall risk in sharing data by evaluating the context risk and data-related risk factors.

The risk-based approach is recommended by the Information and Privacy Commissioner of Ontario (IPC). Although this framework is considered the best practice for de-identifying data, assessing the level of risk when sharing data may be difficult without using de-identification software or expert assistance. Guidelines for the risk-based approach can be found on the IPC's website.

Heuristic de-identification

Heuristic de-identification consists of removing or generalizing identifying information. With this approach, direct identifiers are typically not required for analyses and can be removed from the dataset.

"Generalizing" is a process to aggregate data by grouping information into broader categories, rendering it less specific. The goal is to reduce the number of unique records in the data to limit the likelihood that an individual can be identified. If generalizing the data fails to completely de-identify certain fields, then those values or records should be removed. Careful attention should be paid to how much the data is generalized, as some forms of de-identification may limit the use of the data for future analyses.

Data records having the same quasi-identifiers are referred to as an "equivalence class." Data is aggregated into larger categories to increase the size of each equivalence class. The minimum size of each equivalence class to ensure privacy protection is five records.[1]

Example

Consider Table 1 below, looking only at the columns for age and gender. In this table, individuals sharing the same age and gender would form an equivalence class. Only two individuals have the same age (35 years) and gender (M) values. These two individuals would form an equivalent class (35-year-old males) whose group size is 2.

If there were another row with age and gender equal to 35 and M respectively, then the size of the 35-year-old male equivalence class would be 3. As all the other individuals in the table do not share the same age and gender, they each form their own unique equivalence class whose size is 1.

Table 1 illustrates how data can be generalized to modify equivalence class sizes. When age is transformed into age category, the equivalence class sizes change. However, a unique record remains despite categorizing the age variable (90-year-old female). In such cases, the age field for that record can be generalized further (>50), the value for age can be deleted, or the record (that is, the entire row) can be deleted.

1 Privacy Analytics. (2020). *The definitive guide to de-identification* [White paper].

Please note: the example provided is to demonstrate how the size of the equivalence classes can be modified. For the data to be considered de-identified in this example, the smallest equivalence class size would need to be five.

Table 1: Example of generalizing data into categories

Age	Age category	Gender
35	30 – 40	M
35	30 – 40	M
42	40 – 50	F
38	30 – 40	M
48	40 – 50	F
50	40 – 50	F
90	> 50	F

Using age and gender variables:

- for 35-year-old males, class size n=2
- all other records are unique, class size is 1 for each row

Using age category and gender variables:

- for 30-40-year-old males, class size n=3
- for 40-50-year-old females, class size n=3
- for > 50-year-old female, class size is n=1

Risk-based de-identification[2]

This method assesses the overall risk in sharing data by evaluating contextual and data-related factors. It aims to determine the amount of de-identification required based on the risk of re-identifying individuals in the dataset. Re-identification is when information from a de-identified data set is linked back to a particular individual or individuals. The steps in the risk-based de-identification are outlined here.

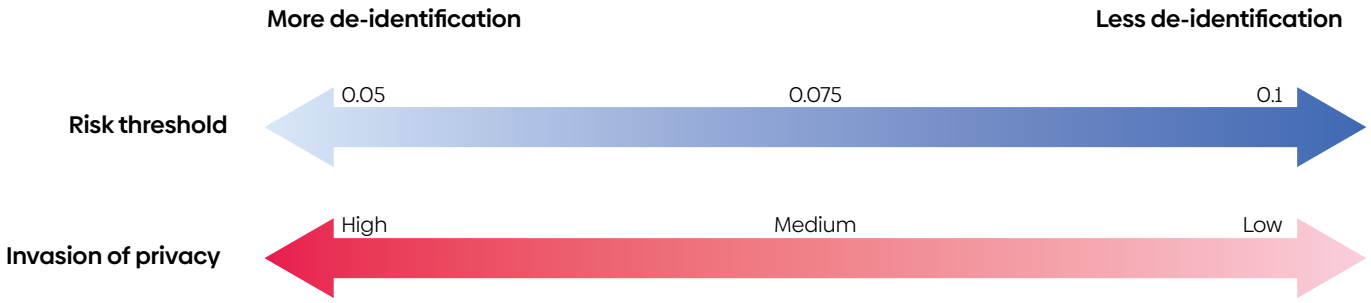
1. Identify the variables that are direct and quasi-identifiers.
2. Determine the acceptable re-identification risk threshold.
3. Measure the overall risk (data risk x context risk).
4. De-identify the data, then re-assess the overall risk and ensure it is below the set re-identification risk threshold.

2 This section has been adapted from the Information and Privacy Commissioner's *De-identification guidelines for structured data*.

Re-identification risk threshold

Prior to de-identification, the organization sharing the data needs to determine an acceptable re-identification risk threshold. This is defined as "the minimum amount of de-identification that must be applied to a dataset in order for it to be de-identified."[3] The re-identification threshold is the baseline level of risk that the overall risk is measured against.

The IPC recommends setting the re-identification risk threshold based on an invasion of privacy assessment.[4] This assessment is "the extent to which the release of the data set would invade an individual's privacy" and is categorized as low, medium or high. The figure below provides a guideline of risk thresholds for each invasion of privacy level and its corresponding re-identification risk. Scenarios in which an individual's invasion of privacy is high should aim for a smaller re-identification risk threshold (less than 0.1). This would require more data de-identification.



Overall risk

The overall risk is calculated by multiplying the context risk by the data risk. The overall risk is then compared to the risk threshold to guide the amount of de-identification required.

Context risk

Context risk assesses the likelihood of a re-identification attack on the data. A re-identification attack is an attempt to re-identify individuals in a de-identified dataset (e.g. hacking, cyber threats). There are different types of attacks that can occur, and these vary according to the type of data release[6] (public or private). Measuring the context risk can be challenging without expert assistance or the use of de-identification software.

3 Information and Privacy Commissioner of Ontario. (2016). *De-identification guidelines for structured data*.

4 Please refer to the Information and Privacy Commissioner's *De-identification guidelines for structured data* for the list of factors to consider in the invasion of privacy assessment.

To simplify the context risk assessment explained in this document, the highest context risk of a re-identification attack will be used, which has a value of 1. For a more detailed description of the context risk and the types of re-identification attacks, please see the IPC's De-identification Guidelines for Structured Data.

Data risk

Data risk is determined by measuring the probability of correctly re-identifying each row in the data. The re-identification risk for each row in the data is calculated by finding the inverse proportion of the equivalence class size (1 / equivalence class size). For instance, each row in an equivalence class with a size of 10 will have a 1 in 10 likelihood of being correctly identified. Rows in smaller equivalence classes have the greatest probability of re-identification compared to those in larger equivalence classes. Their re-identification risk should be applied to all the rows in the data when determining the data risk.

For example, a dataset with equivalence class sizes of 5, 10 and 20 will have corresponding re-identification risks of 0.20, 0.10 and 0.05, respectively. The equivalence class with a size of 5 rows has the largest re-identification risk (0.20) compared to the other equivalence classes. Therefore, the data risk for this dataset is 0.20, as that is the re-identification risk of the smallest equivalence class.

Example

Two organizations have agreed to share data containing personal health information and have signed a data-sharing agreement. As this is a private data release, it can be assumed the invasion of privacy would be low, which would require a lesser amount of de-identification. The custodian organization (the organization that houses the data) sets a re-identification risk threshold of 0.2.

- The overall risk prior to de-identification was measured to be 0.33 (data risk 0.33 x content risk 1.0).
- After de-identification, the resulting overall risk is now 0.1.
- As this is below the set risk threshold of 0.2, the data has been de-identified to an acceptable level and can be shared between the organizations.

Step 3: De-identify the data

The final step in the process is to de-identify the data. Direct identifiers can be suppressed. Quasi-identifiers can be supressed and generalized to modify the equivalent class sizes, along with other techniques. Specialized software can also be used to de-identify the data.

Once the de-identification process is complete, the overall risk is re-assessed and compared against the re-identification risk threshold determined at the outset.

Conventions for de-identifying common quasi-identifiers

Here, we explore generalization examples for three commonly used quasi-identifiers: dates, postal codes and uncommon characteristics.

Dates

Dates are traditionally formatted as date, month and year. Generalized date data can be presented as month and year (12/2000), year (2000) or an interval (2000–2010). Alternatively, dates can be reported as time intervals between dates.

For instance, rather than reporting the admission and discharge dates, the data can be presented as length of stay in days. The usefulness of the transformed date data needs to be considered to ensure it can be used in subsequent analyses. Birth dates can also be generalized to age, which can then be further generalized into age categories.

Postal code

Postal code data is generalized to the first three alphanumeric characters. The average population size in the 3-character Canadian postal codes is approximately 22,000, according to 2016 Census data, and can range from 0 to over 100,000 based on the geographic area.[5] If the population size within the geographic area is too small, then the postal code should be further generalized to a larger area, such as the first two characters or first character only.

5 Statistics Canada. (2019, February 20). *Population and dwelling count highlight tables, 2016 Census*.

For example, the population size for the P0Y postal code is 30. Generalizing the postal code to the first two characters (P0) increases the total population size in the geographic area to over 200,000.

Uncommon or unique data

Fields that have uncommon or unique data elements should be modified and grouped into larger, more general categories. For instance, a unique profession such as "Premier" should be generalized to "Politician." Other examples can include generalizing country of birth to region (for example, Somalia to Sub-Saharan Africa) and generalizing diagnoses into larger diagnostic categories (for example, depressive disorder to mood disorder).

Summary

- When organizations share data, removing personal identifiable information, or de-identifying, is required to protect individuals' privacy.
- Before de-identifying data, organizations should select relevant data and choose one of two possible de-identification frameworks: heuristic or risk-based.
- Heuristic de-identification uses general de-identification rules to remove or generalize data. Risk-based de-identification mathematically assesses the overall risk in sharing data to ascertain how much de-identification is required.
- For both frameworks, de-identification consists of suppressing and generalizing data to reduce the likelihood of the data being re-identified.

References

Canadian Institutes of Health Research Privacy Advisory Committee. (2005). *CIHR best practices for protecting privacy in health research*. Canadian Institutes of Health Research.

Cavoukian, A., & El Emam, K. (2014). *De-identification protocols: Essential for protecting privacy*. Information and Privacy Commissioner of Ontario.

El Emam, K., Jabbouri, S., Sams, S., Drouet, Y., & Power, M. (2006). Evaluating common de-identification heuristics for personal health information. *Journal of Medical Internet Research, 8*(4), e28.

Fraser, R., & Willison, D. (2009). *Tools for de-identification of personal health information*. Electronic Health Information Laboratory.

Information and Privacy Commissioner of Ontario. (2016). *De-identification guidelines for structured data*.

Information and Privacy Commissioner of Ontario. (2018). *General data protection regulation*.

Nelson, G. S. (2015). *Practical implications of sharing data: A primer on data privacy, anonymization, and de-identification*. Thotwave Technologies.

Office of Civil Rights. (2012). *Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) privacy rule*.

Privacy Analytics. (2020). *A privacy governance framework to support de-identification* [White paper].

Privacy Analytics. (2020). *De-identification 201* [White paper].

Privacy Analytics. (2020). *De-identification 301* [White paper].

Privacy Analytics. (2020). *The definitive guide to de-identification* [White paper].

CYMHA_ON

cymhaon

cymha_on

695 Industrial Avenue, Ottawa, Ontario K1G 0Z1

– 1 613 737 2297

– 1 613 738 4894

EN – cymha.ca

FR – smdej.ca