De–identification Guidelines for Structured Data

June 2016

Information and Privacy Commissioner of Ontario

Commissaire à l'information et à la protection de la vie privée de l'Ontario

Acknowledgments

We would like to thank Dr. Khaled El Emam for reviewing an earlier version of these guidelines and providing helpful comments.

CONTENTS

Introduction... 1

Scope of Guidelines 2

Overview of De–identification.. 3

Uses of De–identification 5

Open Data 5

Access to Information Requests... 5

Data Sharing within and among Institutions.. 6

Process for De–identifying Structured Data 6

Step 1: Determine the Release Model.. 7

Step 2: Classify Variables.. 8

Step 3: Determine an Acceptable Re–identification Risk Threshold 9

Step 4: Measure the Data Risk.. 11

Step 5: Measure the Context Risk 13

Step 6: Calculate the Overall Risk 17

Step 7: De–identify the Data 17

Step 8: Assess Data Utility 19

Step 9: Document the Process.. 20

De–identification Governance.. 20

Protecting Against Attribute Disclosure.. 21

Ongoing and Regular Re–identification Risk Assessments...................................... 22

Conclusion.. 22

INTRODUCTION

As the demand for government-held data increases, institutions require effective processes and techniques for removing personal information. An important tool in this regard is de-identification.

“De-identification” is the general term for the process of removing personal information from a record or data set. De-identification protects the privacy of individuals because once de-identified, a data set is considered to no longer contain personal information. If a data set does not contain personal information, its use or disclosure cannot violate the privacy of individuals.[1] Accordingly, the privacy protection provisions of the *Freedom of Information and Protection of Privacy Act* (*FIPPA*) and the *Municipal Freedom of Information and Protection of Privacy Act* (*MFIPPA*) would not apply to de-identified information.

It is important to note that de-identification does not reduce the risk of re-identification of a data set to zero. Rather, the process produces data sets for which the risk of re-identification is very small.

These guidelines will introduce institutions to the basic concepts and techniques of de-identification. They outline the key issues to consider when de-identifying personal information in the form of structured data and they provide a step-by-step process that institutions can follow when removing personal information from data sets.

De-identification can be a complex and technically challenging process. These guidelines take a conservative approach to risk in order to simplify the calculations involved in measuring it. However, some degree of complexity in the process is unavoidable.

When dealing with issues that may arise in de-identification, it is important that you seek advice from technical staff, or other experts in the field (such as your freedom of information and privacy coordinator, or legal counsel). The information contained in these guidelines can serve as a starting point for discussions with those individuals.

Some of the complexity and challenges of de-identification can be addressed through the use of automated tools. While it is possible (and may be appropriate in certain circumstances) to de-identify data sets manually, there are many software tools available that can automate some aspects of the process. When seeking to de-identify a data set, you may wish to consider using de-identification software.

[1] Note, however, that the same cannot be said with respect to the rights of groups of individuals. For a discussion of how to protect against harms relating to groups of individuals when de-identifying data sets, see the section on “De-identification Governance” below.

TERMINOLOGY

Some of the technical terms used in these guidelines are defined below.

adversary: individual or entity attempting to re-identify one or more individuals in the data set

brute force attack: trial-and-error attack that involves attempting all possible combinations to decode an encrypted value

masking: the process of removing a variable or replacing it with pseudonymous or encrypted information

one-way hash function: cryptographic mapping function that is practically impossible to reverse, that is, to recreate the input data from its encrypted value

re-identification: any process that re-establishes the link between identifiable information and an individual

release model: manner in which recipients of a data set are provided access to it

structured data (data set): collection of data in tabular form where every column represents a variable and every row represents a member or individual

target individual: individual targeted by an adversary for re-identification

variable: column of values in a data set representing a set of attributes

SCOPE OF GUIDELINES

Approaches to de-identification range from simple "cookie cutter" lists of variables to be removed or modified, to general loosely defined techniques such as the "cell size of five" rule,[2] to systematic risk-based methodologies. While it may be possible to de-identify data sets in different ways, these guidelines offer direction on taking a *risk-based approach* to de-identification.[3]

Risk-based de-identification involves calculating an acceptable level of re-identification risk for a given data release. The calculation requires the consideration of a number of factors, including whether an adversary can know if a target individual is in the data set. If an adversary knows that a target individual is in the data set, this is called "prosecutor risk." For example, if a teenager's parents know that their child has participated in a survey and the results are to be released in de-identified form, the risk of the parents attempting to re-identify their child's responses would qualify as prosecutor risk. If an adversary does not, or cannot, know if a target

[2] The cell size of five rule is the practice of releasing aggregate data about individuals only if the number of individuals counted for each cell of the table is greater than or equal to five.

[3] The approach to de-identification presented in these guidelines is based largely on the risk-based de-identification methodology developed by Dr. Khaled El Emam. For a select list of books and articles written and co-authored by Dr. El Emam on the topic of de-identification, see Appendix A: Resources.

individual is in the data set, this is called "journalist risk."[4] For example, if only a sample of de-identified rows from an original data set is released, this would qualify as journalist risk.

While some de-identification methodologies support both of the above types of risk—that is, prosecutor and journalist risk—these guidelines support prosecutor risk only—that is, they assume an adversary knows or can know whether a target individual is in the data set. Because prosecutor risk is always equal to or greater than journalist risk,[5] a consequence of this approach is that these guidelines err on the conservative side when it comes to calculating levels of re-identification risk.[6]

De-identification also involves a range of techniques, such as sub-sampling, randomization or swapping. While a number of techniques may be used to remove personal information from data sets, for simplicity these guidelines only discuss the application of the most commonly used techniques, namely *masking, generalization* and *suppression.* Therefore, when using these guidelines to de-identify data sets with a large number of variables, or "high-dimensional" data, the utility of the data sets may be lower than if other techniques were used.

OVERVIEW OF DE-IDENTIFICATION

As noted above, de-identification is the process of removing personal information from a record or data set. "Personal information" is defined in *FIPPA* and *MFIPPA* as "recorded information about an identifiable individual." The Office of the Information and Privacy Commissioner of Ontario (IPC) and the courts have elaborated on this definition, specifically on the meaning of "identifiable," in various orders and reviews.[7] Based on these, de-identification may be defined more precisely as the process of removing any information that (i) identifies an individual, or (ii) for which there is a reasonable expectation that the information could be used, either alone or with other information, to identify an individual.

Throughout these guidelines, the term "de-identification" will be used to convey different aspects of this definition. The term may be used when referring to the *process* of de-identification, which involves a series of steps, considerations and possible outcomes. The term may also be used when referring to the *removal* of identifiable information. From the context, it should be clear in which sense the term is being used.

Applying a "reasonableness standard" to the definition of personal information means that you must examine the context to de-identify information. When de-identifying a data set, you must navigate and consider a number of issues, including:

- *Different release models.* In de-identification, a data set may be released publicly, semi-publicly (also called "quasi-public") or non-publicly. In a public data release, the

[4] See Khaled El Emam, *Guide to the De-identification of Personal Health Information* (Boca Raton, FL: CRC Press, 2013), 182.

[5] See ibid., 195.

[6] Additional guidance on how to de-identify data sets under journalist risk may be found in El Emam, *Guide to the De-identification of Personal Health Information*.

[7] See the test for whether a record can reveal personal information in the judicial review of Order P-1880 at *Ontario (Attorney General) v. Pascoe,* 2002 CanLII 30891 (ON CA), para. 14–15.

data set is available to anyone for download or use without any conditions. This kind of release provides the greatest availability, but the least amount of protection.

In contrast, a non-public data release limits the availability of the data set to a select number of identified recipients. As a condition of receiving the data, recipients must agree to terms and conditions regarding the privacy and security of the data (typically set out in a data sharing agreement). This kind of release provides the least availability but can provide a higher amount of protection.

A data set may also be released semi-publicly, which involves elements of both the public and non-public options. In a semi-public data release, the data set is available to anyone for download; however, as a condition of receiving the data, the recipient must register with the organization releasing the data set and agree to restrictions regarding the processing and sharing of the data (typically in the form of a terms-of-use agreement).

While additional privacy and security measures may be included in terms-of-use agreements for semi-public data releases, these are difficult to enforce due to the open nature of the release. Accordingly, data sets released in this way are limited in terms of the amount of protection they can provide. Depending on the release model used, the required amount of de-identification may vary.

- *Different types of identifiers.* In de-identification, you need to remove information that directly identifies an individual and information for which there is a "reasonable expectation" that the information could be used, either alone or with other information, to identify the individual. The first type of identifier is known as a "direct identifier," and the second type is called an "indirect-" or "quasi-identifier."
- *Different re-identification attacks.* The amount of de-identification that needs to be applied to a data set is determined by how likely it is that an adversary will attempt to re-identify one or more individuals in the data set. Different types of adversaries need to be considered and different types of re-identification attacks need to be analyzed, depending on the release model used. For example, for public data releases, you should assume that someone will attempt a demonstration attack on the data set. For non-public data releases, you should evaluate the threat posed by insiders and data breaches.
- *Different de-identification techniques.* Once you know the level of re-identification risk and have calculated the required amount of de-identification, a corresponding amount of information must be removed from the data set. This can be done in various ways—through techniques such as masking, generalization and suppression.

- *Different types of disclosures.* De-identification techniques protect against the disclosure of individuals' identities and linking information to them. They do not, however, protect against the disclosure of attributes relating to groups of individuals that may be stigmatizing. While you must protect against the disclosure of individuals' identities when releasing de-identified data sets, as a best practice, you should also consider protecting against attribute disclosures. To do this, you may be required to develop a governance model that includes an ethics review of data sets.

USES OF DE-IDENTIFICATION

The primary objective of de-identification is protecting the privacy of individuals. If a data set contains any amount or kind of personal information, it cannot be considered de-identified.

At the same time, one of the main reasons for releasing de-identified data sets is to provide others with an opportunity to study the values and properties of the raw data for research purposes. De-identification, therefore, should also seek to preserve as much utility in the information as possible, while protecting the privacy of individuals.

This dual purpose of de-identification makes it an important tool to consider for use in a number of contexts, including open data, access to information requests and data sharing within and among institutions.

OPEN DATA

De-identification may be used to enable data sharing in situations where an institution does not have the authority to disclose personal information. An example of such a situation is the growing number of "open data" initiatives in Ontario. Open data initiatives seek to increase government transparency and accountability by proactively releasing data sets and making them freely available to anyone for use and republishing. Given the increased amount and availability of information these initiatives provide, it is important that institutions release their data sets in a way that protects the privacy of individuals.

Open data initiatives also seek to promote research, innovation and the development of new applications and services. The greater the utility of open data sets, the better the chances of success for researchers, start-up companies and entrepreneurs seeking to use public data.

ACCESS TO INFORMATION REQUESTS

De-identification may also be useful in responding to access to information requests for structured data or data sets. Under sections 10(2) of *FIPPA* and 4(2) of *MFIPPA*, institutions are required to "disclose as much of the record as can reasonably be severed" without disclosing any exempt information. By using de-identification, institutions can respond to requests in a privacy-protective manner while preserving the utility of the information. De-identification is an innovative tool that may present institutions with an opportunity to further the transparency purposes of *FIPPA* and *MFIPPA* in ways that were not possible before.

DATA SHARING WITHIN AND AMONG INSTITUTIONS

While access to information requests and open data initiatives provide information to the public, there is also a growing desire in government services for institutions to break down their "silos" and share more information within—and among—themselves. This may happen for a number of reasons. For example:

- information from one institution or program area may be relevant to the planning of a program or service in another institution or area
- one institution may have expertise in data processing or software development that another institution requires, but does not have
- an institution that funded a program or service that was delivered by another institution may want to evaluate the effectiveness of the program or service

Data sets that contain personal information may be shared within and among institutions only if the disclosure is permitted under section 42(1) of *FIPPA* or section 32 of *MFIPPA*. If the disclosure is not permitted and the institutions still wish to share data sets, then (similar to an access to information request or open data release) any personal information must be removed.

However, even if disclosure is permitted under *FIPPA* or *MFIPPA*, there may still be important privacy issues to consider. While information sharing among institutions can play an important role in providing better, more efficient services, the practice may also have the unintended consequence of undermining the privacy of individuals by diminishing the amount of control individuals have over their personal information. Therefore, as a best practice, institutions should always consider de-identifying data sets before sharing them.

PROCESS FOR DE–IDENTIFYING STRUCTURED DATA

To protect the privacy of individuals while preserving as much utility in the information as possible, the amount and types of de-identification need to be determined through a systematic analysis of the level and kinds of re-identification risk involved in the release of a data set. When attempting to de-identify a data set, you should consider the following process:

1. determine the release model
2. classify variables
3. determine an acceptable re-identification risk threshold

4. measure the data risk
5. measure the context risk
6. calculate the overall risk
7. de-identify the data
8. assess data utility
9. document the process

STEP 1: DETERMINE THE RELEASE MODEL

As noted above, a de-identified data set may be released publicly, semi-publicly or non-publicly. Each release model allows for different levels of availability and protection of information. Depending on the purposes and/or legislative requirements of the data release, the suitability of each model may vary.

The release model plays an important role in the de-identification process because the amount of de-identification required may vary, depending on the model. For example, because public data releases provide the greatest availability, but the least amount of protection, you may require a significant amount of de-identification to protect individual privacy. Non-public data releases provide the least availability but can provide a higher amount of protection, requiring a smaller amount of de-identification.

Access requests should be handled as though they are public data releases because *FIPPA* and *MFIPPA* do not require the person requesting information to agree to terms or conditions regarding the processing, privacy or security of the information.

Similarly, when publishing open data, it is common practice to place as few restrictions as possible on the information, including who can access it and how. Requirements for individuals to register and identify themselves to the organization publishing the data are considered barriers to access, use and the ability of individuals to find the information. As such, when individuals who download the data set cannot be identified, these disclosures should be handled as public data releases.

However, there may be instances where registration of individuals and verification of their identities is required. For example, a government- or university-sponsored programming competition, or "hackathon," may involve the release of a de-identified data set to the public or student body, but restrict participants from using the data set in certain ways (including re-identifying any individuals in it and disclosing the information to third parties, through a terms-of-use agreement). If the terms-of-use agreement does not require participants to have in place additional privacy and security measures or such measures are not enforceable, these kinds of disclosures should be handled as semi-public data releases.

Finally, when sharing information among institutions, because access to the data set is limited to the receiving program area or institution, requirements regarding the privacy and security of the information can be set and enforced through a data sharing agreement. In these cases, such disclosures may be handled as non-public data releases.

For a data release to be treated as non-public, there must be a data sharing agreement in place between the parties. The data sharing agreement is an important part of the risk mitigation strategy in these releases.

STEP 2: CLASSIFY VARIABLES

If a data set is about individuals, then each row in the file represents an individual, and each column represents a variable of information collected about the individuals. Depending on the type of information, some variables may be used to identify individuals, either directly or indirectly, while others may not. De-identification is only concerned with variables that may be used to identify individuals. As noted above, there are two kinds of such variables: direct identifiers and indirect or quasi-identifiers.

DIRECT IDENTIFIERS

Direct identifiers consist of one or more variables that can be used to identify a single individual, either by themselves or in combination with other readily available sources of information.[8] Examples include name, address, email address, telephone number, fax number, credit card number, license plate number, vehicle identification number, social insurance number, health card number, medical record number, device identifier, biometric identifiers, internet protocol (IP) address number and web universal resource locator (URL).

Typically, direct identifiers are not useful for the purposes of data analysis. For example, the email addresses of individuals will likely not be relevant to a study of work commutes. However, if the values of a direct identifier are relevant, then you should classify it as a quasi-identifier and allow the variable to be de-identified. However, if a variable is not useful for data analysis it should be classified as a direct identifier and flagged for removal or replacement with a pseudonym regardless of its characteristics (see step 7).

QUASI–IDENTIFIERS

Quasi-identifiers are variables with two important characteristics: (1) an adversary is assumed to have background knowledge of them, and (2) they can be used, either individually or in combination, to re-identify an individual in the data set.[9] A variable can be a quasi-identifier only if an adversary has background knowledge of it. A challenge with classifying quasi-identifiers

[8] Khaled El Emam and Bradley Malin, "Appendix B: Concepts and Methods for De-identifying Clinical Trial Data," *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk* (Washington D.C.: National Academies Press, 2015), **http://www.ncbi.nlm.nih.gov/books/NBK285994/**.

[9] See ibid.

is in anticipating the possible sources of background knowledge. An adversary may obtain background knowledge about one or more individuals in the data set in different ways, including:

- information about individuals may be available in public registries (such as voter lists or court records), in the media (e.g., obituaries), from professional organizations (e.g., member lists) or employers (e.g., staff directories or biographies)
- the adversary may know one or more individuals (e.g., neighbour, co-worker or ex-spouse)
- one or more individuals may be a celebrity and there is publicly available information about them
- the adversary may have access to additional sources of information about individuals (e.g., data sets from other research projects)
- individuals may post information about themselves online (e.g., on social networking sites or personal blogs)[10]

Examples of quasi-identifiers include gender, date of birth or age, event dates (e.g., death, admission, procedure, discharge, visit), locations (e.g., postal codes, building names, regions), ethnic origin, country of birth, languages spoken, aboriginal status, visible minority status, profession, marital status, level of education, total years of schooling, criminal history, total income and religious denomination.

The value of a quasi-identifier may also be predicted from one or more variables in the data set that share a correlation with it. For example, an individual's age may be predicted from the date or year of their graduation. Because such variables may reveal the value of a quasi-identifier, you should classify them as quasi-identifiers.

STEP 3: DETERMINE AN ACCEPTABLE RE-IDENTIFICATION RISK THRESHOLD

De-identification protects the privacy of individuals by removing information that identifies an individual or for which there is a reasonable expectation that it could be used, either alone or with other information, to identify an individual. To protect personal privacy, the amount of de-identification that is required to be applied is proportional to the level of re-identification risk involved in the release of the data set. The higher the re-identification risk of a data release, the greater the amount of de-identification required.

To determine an acceptable level of re-identification risk (or threshold) for a data set, you must assess the extent to which the release of the data set would invade an individual's privacy. The result of your assessment should be a qualitative value typically in the range of "low," "medium" or "high."

[10] See "What is a quasi-identifier?" *Electronic Health Information Laboratory,* **http://www.ehealthinformation.ca/faq/quasi-identifier/.**

When assessing the level of potential privacy invasion of individuals, assume that the information in the data set is identifiable and no de-identification has taken place. Under this assumption, the level of invasion of privacy is a function of different factors, including:

- the sensitivity of the information
- the scope and/or level of detail of the information
- the number of individuals
- the potential harms or injuries to individuals in the event of a breach or inappropriate use
- whether the disclosure of the information is permitted under *FIPPA* or *MFIPPA* without the consent of the individuals
- whether the information was unsolicited or given freely by the individuals, with little or no expectation of privacy
- whether the individuals explicitly consented to their information being disclosed in de-identified form for this secondary purpose and/or were properly notified at the time of collection of this data practice[11]

The result of the invasion of privacy assessment is a qualitative value; however, the amount of de-identification that is required to be applied to a data set is quantified numerically. To bridge this divide, once you have assessed the invasion of privacy value, you must translate the result into a numerical value, representing the amount of de-identification proportionate to that level of risk. This "re-identification risk threshold" represents, in general, the minimum amount of de-identification that must be applied to a data set in order for it to be considered de-identified, that is, for it to no longer contain personal information. Accordingly, it forms the baseline against which to compare your calculations concerning de-identification going forward.

When translating between the (qualitative) invasion of privacy value and the (quantitative) re-identification risk threshold, consider a key aspect of de-identification—namely, that de-identification does not produce data sets for which there is *zero probability* of re-identification. Rather, it results in data sets for which the probability of re-identification is *very low*, given the level of re-identification risk involved in the release. The amount of de-identification proportionate to the invasion of privacy value should be equal to a very low probability of re-identification given that level of risk.

The following table may be used as a guideline in determining what may be considered a very low value for the probability of re-identification for data sets with different invasion of privacy values.[12]

[11] See El Emam, *Guide to the De-identification of Personal Health Information*, 283–290. This section of El Emam's book also contains an assessment tool that may help in determining the level of risk to individuals posed by the release of a data set.

[12] See ibid., 228.

Invasion of Privacy	Re-identification Risk Threshold	Cell Size Equivalent
Low	0.1	10
Medium	0.075	15
High	0.05	20

When combined with the calculations involved in step 5, the values listed in the table are consistent with data release precedents across Canada and the United States.[13] The table also includes the cell size equivalent for each probability of re-identification for illustrative purposes only. Cell sizes apply to aggregate count or frequency tables, not individual-level structured data. Nonetheless, the concept can be used to illustrate the general effect of de-identification on such data sets. For example, a data set with a probability of re-identification of 0.1 means that each row in the data set will in general have the same values for quasi-identifiers as nine other rows, that is, have a "cell size" of 10.

STEP 4: MEASURE THE DATA RISK

Once you have determined an acceptable re-identification risk threshold, the next step is to measure the amount of re-identification risk in the data set itself. The data risk is used to determine the level of re-identification risk involved in the release.

Measuring the amount of re-identification risk in a data set is a two-step process. You must (1) calculate the probability of re-identification of each row, and (2) apply the appropriate risk measurement method based on the release model used.

4.1 CALCULATE THE PROBABILITY OF RE-IDENTIFICATION OF EACH ROW

Each row in a data set about individuals contains information about one individual. Accordingly, each row has a probability of re-identification. For a given row, the probability of re-identification is dependent on how many other rows in the data set have the same values for variables that are quasi-identifiers.

All the rows in a data set with the same values for variables that are quasi-identifiers form an "equivalence class." For example, in a data set with variables for gender, age and highest level of education, all the rows corresponding to 35-year-old men with post-secondary degrees would form an equivalence class. The size of an equivalence class is equal to the number of rows with the same values for quasi-identifiers.

For each row, the probability of re-identification is equal to 1 divided by the size of its equivalence class. For example, each row in an equivalence class of size 5 has a probability of re-identification of 0.2.

[13] See ibid., 279–282.

$$\text{Probability of re-identification for a given row} = \frac{1}{\text{Size of equivalence class}}$$

Rows with larger equivalence classes have lower probabilities of re-identification, since more rows and therefore more individuals in the data set have the same values for quasi-identifiers. Rows with smaller equivalence classes have higher probabilities of re-identification, since less rows (less individuals) have the same values for quasi-identifiers.

4.2 APPLY THE APPROPRIATE RISK MEASUREMENT METHOD

While the probability of re-identification of each row is equal to 1 divided by the size of its equivalence class, there are different ways to use these values to measure the amount of re-identification risk in the data set, depending on the release model used.

Public Data Releases: Maximum Risk

For public data releases, you should assume that someone will attempt a demonstration attack for publicity. These kinds of attacks will target the most vulnerable rows in the data set, which are those with the smallest equivalence classes and highest probability of re-identification. Because of this, you should use the maximum probability of re-identification across all rows to measure the amount of re-identification risk.

Non-Public Data Releases: Strict Average Risk

For non-public data releases, because access to the data set is limited to a select number of identified recipients, you should assume that no row is more vulnerable than others to a re-identification attack. Here, you should use the average probability of re-identification across all rows to measure the amount of re-identification risk in the data set. However, to protect against unique rows or equivalence classes with a high risk of re-identification, the average should be a “strict” average where no row may have a probability of re-identification that is greater than a specific value. A cut-off of 0.33 is often proposed, that is, the smallest size of equivalence class in the data set should be 3.[14] In practice, however, a maximum probability of re-identification of 0.5 may also be used, which in the case of strict average ensures that there are no unique rows and that the average risk is acceptably small.

Semi-Public Data Releases: Maximum Risk

Because semi-public data releases are available to anyone for download, you should assume that the most vulnerable rows will be more at risk of attack than others. Because of this, like public data releases, you should use the maximum probability of re-identification across all rows to measure the amount of re-identification risk.

[14] See El Emam and Malin, “Appendix B: Concepts and Methods for De-identifying Clinical Trial Data.”

STEP 5: MEASURE THE CONTEXT RISK

While the risk from the data set plays an important role in determining the level of re-identification risk involved in the release of a data set, it is not the only factor to consider. The re-identification risk is also a function of the kinds of re-identification attacks that are possible on the data set given the release model used. Further analyzing the re-identification risk in terms of possible attacks produces the context risk. Together with the data risk, this value is used to calculate the overall risk of re-identification involved in the release of a data set (in step 6).

The context risk is the probability of one or more re-identification attacks being launched against a data set. While re-identification attacks may be launched on any de-identified data set once it has been released, the adversaries and kinds of attacks differ depending on the release model used.

PUBLIC DATA RELEASES

The calculations used to measure the context risk for public data releases are straightforward. Because the data set is made available to anyone for download or use without any conditions, you should assume that someone will attempt a demonstration attack for publicity. The probability of an adversary launching a re-identification attack against the data set is therefore 1.

NON–PUBLIC DATA RELEASES

In contrast, the calculations for measuring the context risk for non-public data releases, in particular the methods and equations used to determine the probabilities of possible re-identification attacks, are more complex and may require specialized knowledge or skills to carry out. As noted in the introduction, if you are not confident in your abilities to carry out these calculations, you may wish to seek advice from technical staff or other experts in the field.

If technical or expert advice is not available, another option is to measure the context risk as though it were for a public data release using the (much simpler) method above. While this may result in a data set with lower utility, the amount of protection against re-identification attacks would be equal to a non-public data release, if not greater.

For non-public data releases, the probabilities of three different re-identification attacks or threats need to be determined:

1. deliberate insider attack
2. inadvertent recognition of an individual in the data set by an acquaintance
3. data breach

You should use the highest of these probabilities when measuring the context risk.

Attack 1: Deliberate Insider Attack

The probability of a recipient of a non-public data release attempting to re-identify one or more individuals in the data set is based on two factors:

1. the extent of the controls set out in the data sharing agreement regarding the privacy and security of the data
2. the motives and capacity of the recipient in regards to performing a re-identification attack

Both of these factors entail qualitative assessments, resulting in values typically in the range of "low," "medium" or "high."

Privacy and Security Controls

Depending on the privacy and security controls set out in the data sharing agreement for a non-public data release, the probability of a recipient attempting to launch a re-identification attack may vary. The higher the level of privacy and security controls, the lower the probability of a re-identification attack being launched. While a more complete list of controls is available,[15] some privacy and security controls that may be considered in a data sharing agreement include:

- recipient allows only "authorized" staff to access and use data on a "need-to-know" basis (only when required to perform their duties)
- a non-disclosure or confidentiality agreement (pledge of confidentiality) is in place for all staff, including external collaborators and subcontractors
- data will be disposed of after a specified retention period
- data will not be disclosed or shared with third parties without appropriate controls or prior approval
- privacy and security policies and procedures are in place, monitored and enforced
- mandatory and ongoing privacy, confidentiality and security training is conducted for all individuals and/or team members including those at external collaborating or subcontracting sites
- a breach of privacy protocol is in place, including immediate written notification to the data custodian
- virus-checking and/or anti-malware programs have been implemented
- a detailed monitoring system for audit trails has been instituted to document the person, time and nature of data access
- if electronic transmission of the data is required, an encrypted protocol is used

[15] See the list of privacy and security controls available in Appendix 1 of Khaled El Emam et al., "Evaluating the Risk of Re-identification of Patients from Hospital Prescription Records." *Canadian Journal of Hospital Pharmacy* 62, no. 4 (Jul-Aug 2009): 307–319, **http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2826964/**.

- computers and files that hold the disclosed information are housed in secure settings in rooms protected by such methods as combination lock doors or smart card door entry, with paper files stored in locked storage cabinets[16]

Motives and Capacity

Additional factors to consider when determining the probability of a recipient attempting to launch a re-identification attack are their motives and capacity. The more motivated and more capable the recipient is with respect to re-identifying one or more individuals in the data set, the higher the probability of a re-identification attack being launched. When assessing motives and capacity, consider:

- whether the recipient has worked with your institution in the past without incident
- whether possible reasons exist, financial or otherwise, for the recipient to attempt to re-identify one or more individuals
- whether the recipient has the technical expertise and/or financial resources to attempt any re-identification
- whether the recipient has access to other private databases or data sets that could be linked to the data to re-identify one or more individuals[17]

Probability of Re-identification Attack

Based on the level of privacy and security controls in the data sharing agreement and the motives and capacity of the recipient, the probability of a deliberate re-identification attack being launched by an insider may be estimated. The following table may be used as a guideline in determining what may be considered an acceptable estimate for the probability of a re-identification attack being launched against non-public data sets.[18]

Privacy and Security Controls	Motives and Capacity	Probability of Re-identification Attack
High	Low	0.05
	Medium	0.1
	High	0.2
Medium	Low	0.2
	Medium	0.3
	High	0.4
Low	Low	0.4
	Medium	0.5
	High	0.6

[16] See El Emam, *Guide to the De-identification of Personal Health Information*, 290–371. This section of El Emam's book also contains an assessment tool that may help in determining the level of privacy and security controls in a data sharing agreement.

[17] See ibid., 373–376. This section of El Emam's book also contains an assessment tool that may help in determining the level of motives and capacity of a recipient.

[18] See ibid., 208.

Attack 2: Inadvertent Recognition of an Individual by an Acquaintance

In addition to deliberately attempting a re-identification attack, the recipient of a non-public data release may also inadvertently re-identify one or more individuals. This could happen if, while analyzing the data, they recognize a friend, colleague, family member or acquaintance. The probability of such an "attack" occurring is equal to the probability of a random recipient knowing someone in the data set. To calculate this, the following equation may be used:

$$1 - (1 - p)^m$$

In this equation, *p* is the percentage of individuals in the population who have the condition or characteristic discussed in the data set and *m* is the number of people, on average, that an individual knows.[19] Take, for example, a data set about individuals who carpool to work. Based on values of *p* and *m*, the equation would give the probability that a random individual knows someone who carpools to work.

The value of *p* should be determined by recent population statistics. On the other hand, the value for *m* may vary depending on the kind of relationship with an individual required to have knowledge about them regarding the condition or characteristic discussed in the data set. For friends, you should in general use a value of *m* between an average of 150, that is, "Dunbar's number,"[20] and 190.[21]

Attack 3: Data Breach

The third attack to consider in the case of a non-public data release is that of a data breach on the part of the recipient. If a data breach occurs at the recipient's facilities, you should assume that an external adversary will attempt a re-identification attack. Therefore, the probability of such an attack occurring is equal to the probability of a breach occurring at the recipient's facilities. To calculate this value, you should use publicly available data on the prevalence of data breaches in the recipient's respective industry.

SEMI–PUBLIC DATA RELEASES

The possible re-identification attacks for semi-public data releases can be considered the same as those for non-public data releases. Accordingly, to measure the context risk for semi-public data releases, you should use the same method and equations as for non-public data releases, with one adjustment. With respect to "Attack 1: Deliberate Insider Threat," you should assume that the recipient has high motives and capacity and, at best, low privacy and security controls. This is because semi-public data releases are available to anyone for download and are limited in terms of the amount of protection they can provide.

[19] See ibid., 211.

[20] See "Dunbar's number," *Oxford Dictionaries,* **http://www.oxforddictionaries.com/definition/english/dunbar's-number**.

[21] See El Emam, *Guide to the De-identification of Personal Health Information*, 213.

When developing the terms-of-use agreement, you should include provisions that, at a minimum, prohibit recipients from:

- attempting to re-identify individuals in the data set
- linking to external data sets or information
- sharing the data set without permission

STEP 6: CALCULATE THE OVERALL RISK

Once the data risk and the context risk have been measured, the overall risk of re-identification can be calculated. The overall risk is equal to the data risk multiplied by the context risk.

Overall risk = data risk x context risk

The overall risk is equivalent to the probability of one or more rows being re-identified if an attack was launched. For example, if a data set has a data risk of 0.2 and a context risk of 0.5, the overall risk for the data set is 0.1.

STEP 7: DE-IDENTIFY THE DATA

For a data set to be considered de-identified, any identifiable information must be removed. The values of a data set may be transformed in various ways to remove any information that identifies an individual or for which there is a reasonable expectation that the information could be used, either alone or with other information, to identify an individual. Depending on the type and nature of the identifiers, different techniques may be applied. To remove any identifiable information, you should:

1. mask direct identifiers
2. modify the size of equivalence classes
3. ensure that the overall risk is less than or equal to the re-identification risk threshold

7.1 MASK DIRECT IDENTIFIERS

Variables classified as direct identifiers are not used for data analysis because, as noted above, they are not normally useful for research purposes. Because of this, the simplest, most privacy-protective way of dealing with them is to suppress their values in the data set by removing the column of the directly identifying variable.

However, depending on the nature of the research, there may be a need to contact the individuals involved and notify them of the results. In such cases, the directly identifying variables should be transformed using a different masking technique, such as:

- replacing the values with pseudonyms and maintaining the linking database in a secure location
- encrypting the values and storing the encryption key in a safe place

Because directly identifying variables can be used, either by themselves or in combination with other readily available sources of information, to identify individuals, the utmost care must be taken when performing such transformations. If a directly identifying variable is transformed improperly or in an insecure manner, an adversary may be able to re-identify a large number of individuals.

For example, a common technique for creating pseudonyms is to transform the value of a directly identifying variable into an irreversible code using a one-way hash function. However, this technique may be vulnerable to brute force attacks if the total number of possible values of the variable is small enough that the adversary can compute the hash values of all the possible values of the variable in a reasonable amount of time and use this to create a reverse lookup table of hashed and original values. To protect against such attacks, you should always add random data to the input of a one-way hash function and maintain this "salt" or "key" along with the linking database in a secure location.

7.2 MODIFY SIZE OF EQUIVALENCE CLASSES

For a data set to be considered de-identified, the overall risk of re-identification must be less than or equal to the re-identification risk threshold. If the overall risk is greater than the re-identification risk threshold, you must modify the size of equivalence classes in the data set in order to reduce the data risk.

Depending on the values of its quasi-identifiers, a data set may have equivalence classes of different sizes. De-identification involves transforming the values of quasi-identifiers in various ways to modify the size of equivalence classes in a data set. Two techniques to do this are generalization and suppression.

Generalization

Generalization is the process of removing precision from a value to produce a more general value. It may be applied in increasing amounts. For example, a full date may be generalized to month and year, which may in turn be generalized to year, which may in turn be generalized to five-year interval, 10-year interval, and so on.

When using generalization, you should apply it to all the rows of a variable. You should also ensure that the set of generalizations used within a variable are uniform and do not overlap. For example, a uniform set of five-year age intervals would be 10–14, 15–19, 20–24, 25–29, 30–34, and so on.

There is one exception to this. For continuous variables, you may introduce a cut-point at the top or bottom range of values to create a “catch all” category for outliers. For example, the age of individuals may be generalized to year, with a catch all category of “90+” for individuals who are 90 or older. This generalization technique is known as top- or bottom-coding, depending on where the cut-point is made.

Suppression

Suppression is the process of removing values from a data set. In contrast to generalization, which applies to all the rows of a quasi-identifier, suppression affects single rows only. Suppression of a value of a quasi-identifier may happen at different levels. For example, it may involve removing the entire row, the set of quasi-identifiers in the row or only the individual cell. While the less information removed from a data set the greater potential for a higher utility data set, when suppressing a value of a quasi-identifier, you may need to remove the entire row or a set of quasi-identifiers in the row to ensure that the equivalence classes are of the appropriate size.

7.3 ENSURE THAT THE OVERALL RISK IS LESS THAN OR EQUAL TO THE RE-IDENTIFICATION RISK THRESHOLD

If the size of any equivalence class in the data set has been modified, you must recalculate the overall risk of re-identification and compare it to the re-identification risk threshold. For a data set to be considered de-identified, the data risk must be sufficiently reduced so that the overall risk is less than or equal to the re-identification risk threshold.

STEP 8: ASSESS DATA UTILITY

There may be a trade-off between the amount of de-identification applied to a data set and the utility of the resulting information. The more the variables that qualify as quasi-identifiers are de-identified using techniques such as generalization and suppression, the higher the potential for a corresponding loss in the utility of the data set.

While generalization and suppression may be applied to a data set to ensure that the overall risk of re-identification is less than or equal to the re-identification risk threshold, these de-identification techniques may be applied in different ways and combinations to achieve this result. For example, one approach may rely more on generalization and reducing the precision of categories to increase the size of equivalence classes. Another approach may rely more on suppression and removing rows or cells of variables with equivalence classes that are too small. Depending on the properties of the data set, different applications and/or combinations of generalization and suppression may preserve more utility in the information while protecting the privacy of individuals.

As a general rule, suppression should be considered before generalization, unless more than five per cent of the rows in the data set already have some form of suppression.[22] Because suppression removes information from single rows, in contrast to generalization, which reduces the precision of all the rows in the data set, you may wish to consider suppression as a starting point for de-identification.

If the utility of the data set is low or could be improved—for example, more than five per cent of the rows have some form of suppression or further generalization could be avoided by suppressing certain rows or values—you may wish to repeat steps 7.2 and 7.3 above. Applying and/or combining the techniques of generalization and suppression in a new way could produce a higher utility data set while ensuring that the overall risk of re-identification remains less than or equal to the risk threshold.

STEP 9: DOCUMENT THE PROCESS

Each attempt at de-identifying a data set containing personal information should follow the same steps and evaluate the same set of issues. However, the variables and values, and the analysis to determine the amount and kinds of de-identification, will differ for each data release. To help guide you through the complexities and challenges involved in de-identifying personal information, you should consider producing a report documenting the process and its results. There are a number of benefits to this best practice, including:

- the ability to demonstrate due diligence and evidence of compliance, which may be important in the event of a privacy breach or complaint to the IPC
- confidence (of individuals, other institutions, partners and your own management) that best practices are being followed.
- increased transparency, awareness, understanding and trust in your organization’s information management practices

DE-IDENTIFICATION GOVERNANCE

Responsibility for releasing a de-identified data set does not end with the completion of the process for removing any identifiable information. Governance is an important aspect of releasing any de-identified data set. A robust de-identification governance process may include activities such as:

- protecting against attribute disclosure[23]
- ongoing and regular re-identification risk assessments

[22] See Khaled El Emam et al., “A Globally Optimal k-Anonymity Method for the De-Identification of Health Data,” *Journal of the American Medical Informatics Association* 16, no. 5 (Sep-Oct 2009): 670–682, **http://dx.doi.org/10.1197/jamia.M3144**.

[23] See El Emam and Malin, “Appendix B: Concepts and Methods for De-identifying Clinical Trial Data.”

- auditing data recipients to ensure that they are complying with the conditions of the data sharing agreement
- examining the disclosures of overlapping data sets to ensure that the re-identification risk is not increasing with new data releases, or that potential collusion among data recipients does not increase the re-identification risk
- maintaining transparency around the de-identification practices of the institution
- assigning responsibility and accountability for de-identification
- maintaining oversight of changes in relevant regulations and legislation as well as court cases
- developing a response process in case there has been a re-identification attack
- ensuring that individuals performing de-identification have adequate and up-to-date training[24]

While all of the above activities are important to consider when developing a de-identification governance process, the first two raise issues that are specific to de-identification.

PROTECTING AGAINST ATTRIBUTE DISCLOSURE

One of the reasons for releasing de-identified data sets is to provide others with an opportunity to study the values and properties of the raw data and draw inferences from them. This is the primary purpose of statistics and data analysis.

While de-identification techniques protect against the disclosure of individuals' identities, they do not protect against the disclosure of attributes relating to groups of individuals that may be stigmatizing to those individuals. Some inferences may be desirable insofar as they may enhance our understanding of a particular issue or topic. Others may subject groups of individuals to unjust or prejudicial treatment or would be considered offensive. For example, a data set showing whether children of parents with a particular religious affiliation are being vaccinated against certain viruses could result in stigmatization.[25]

The privacy protections set out in *FIPPA* and *MFIPPA* relate to the personal information of individuals only and do not include measures to address potential harms affecting groups of individuals. Nonetheless, as a best practice, you should consider whether any group attributes in a de-identified data set are stigmatizing before releasing the data set. An ethics review of the data set may be needed to achieve this.

[24] See Khaled El Emam, "The Twelve Characteristics of a De-identification Methodology," *Risky Business: Sharing Health Data While Protecting Privacy* (Trafford Publishing: 2013), 134–146 at 141.

[25] See El Emam, *Guide to the De-identification of Personal Health Information*, 9–10.

ONGOING AND REGULAR RE–IDENTIFICATION RISK ASSESSMENTS

Another important step in the process of de-identifying a data set is to classify variables, above all, quasi-identifiers. A challenge with classifying quasi-identifiers is in anticipating the possible sources of background knowledge that an adversary may have, especially since new sources of information may become available at any time.

The potential for individuals to be re-identified by combining new sources of information with otherwise de-identified data is an important privacy concern to consider. Unanticipated sources of information that were not available at the time of de-identification may become available and be used to re-identify individuals.

Once you have released a de-identified data set, you should consider monitoring whether any new sources of information have become available and whether such sources may be used to re-identify individuals in the data set. If so, you should re-assess the classification of variables. Depending on the re-assessment, you may need to mask or de-identify additional variables to ensure that the overall probability of re-identification is less than or equal to the re-identification risk threshold.

In addition, you may also wish to commission a staged re-identification attack on a data set to determine how difficult (or easy) it would be for an attacker to re-identify one or more individuals. This would provide an empirical measurement of the risk of re-identification. While more expensive than statistical evaluations, commissioned attacks should be performed on particularly high-risk data sets, or every few years on other data sets, to understand the attack landscape.[26]

CONCLUSION

De-identification is the process of removing information that identifies an individual or for which there is a reasonable expectation that the information could be used, either alone or with other information, to identify an individual.

De-identification can be a complex and technically challenging process. The risk-based approach developed in these guidelines outlines a step-by-step process for de-identifying data sets in accordance with *FIPPA* and *MFIPPA.*

When attempting to de-identify structured data or data sets, institutions may wish to seek advice from technical staff or other experts in the field, their freedom of information and privacy coordinator or legal counsel. Institutions may also wish to consider automated tools or de-identification software to facilitate the process.

De-identification results in data sets for which the probability of re-identification is very low, given the level of re-identification risk involved in the release. While de-identification techniques protect against the disclosure of individuals' identities, they do not protect against other risks, including the disclosure of stigmatizing group attributes. Institutions should consider instituting a robust de-identification governance process to address additional risks and concerns.

[26] See the "motivated intruder" test in the U.K. Information Commissioner's Office, *Anonymisation Code of Practice,* **https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf**.

APPENDIX A: RESOURCES

Emam, Khaled El. *Guide to the De-identification of Personal Health Information*. Boca Raton, FL: CRC Press, 2013.

Emam, Khaled El. “The Twelve Characteristics of a De-identification Methodology.” *Risky Business: Sharing Health Data While Protecting Privacy*. Trafford Publishing: 2013.

Emam, Khaled El and Luk Arbuckle. *Anonymizing Health Data.* Sebastopol, CA: O’Reilly, 2014.

Emam, Khaled El, Fida K Dankar, Régis Vaillancourt, Tyson Roffey and Mary Lysyk. “Evaluating the Risk of Re-identification of Patients from Hospital Prescription Records.” *Canadian Journal of Hospital Pharmacy* 62, no. 4 (Jul-Aug 2009): 307–319*,* **http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2826964/**.

Emam, Khaled El, Fida Kamal Dankar, Romeo Issa, Elizabeth Jonker, Daniel Amyot, Elise Cogo, Jean-Pierre Corriveau, Mark Walker, Sadrul Chowdhury, Regis Vaillancourt, Tyson Roffey and Jim Bottomley. “A Globally Optimal k-Anonymity Method for the De-Identification of Health Data.” *Journal of the American Medical Informatics Association* 16, no. 5 (Sep-Oct 2009): 670–682, **http://dx.doi.org/10.1197/jamia.M3144**.

Emam, Khaled El and Bradley Malin. “Appendix B: Concepts and Methods for De-identifying Clinical Trial Data.” *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk.* Institute of Medicine of the National Academies. Washington DC: The National Academies Press, 2015. **http://www.ncbi.nlm.nih.gov/books/NBK285994/**.

The Expert Panel on Timely Access to Health and Social Data for Health Research and Health System Innovation, *Accessing Health and Health-Related Data in Canada.* Ottawa: Council of Canadian Academies, 2015. **http://www.scienceadvice.ca/uploads/eng/assessments%20and%20publications%20and%20news%20releases/Health-data/HealthDataFullReportEn.pdf**.

Health Information Trust Alliance, *HITRUST De-identification Framework*, 2015. **https://hitrustalliance.net/de-identification/**.

National Institute of Standards and Technology, *NIST Internal Report 8053. De-identification of Personal Information*, 2015*.* **http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf**.

Office for Civil Rights (OCR), *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*, 2012. **http://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html**.

PhUSE De-identification Working Group, *De-identification Standard for CDISC SDTM 3.2*, 2015. **http://www.phuse.eu/Data_Transparency.aspx**.

U.K. Information Commissioner’s Office, *Anonymisation Code of Practice*, 2012. **https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf**.

“What is a quasi-identifier?” *Electronic Health Information Laboratory.* **http://www.ehealthinformation.ca/faq/quasi-identifier/**.

ABOUT THE INFORMATION AND PRIVACY COMMISSIONER OF ONTARIO

The role of the Information and Privacy Commissioner of Ontario is set out in three statutes: the *Freedom of Information and Protection of Privacy Act*, the *Municipal Freedom of Information and Protection of Privacy Act* and the *Personal Health Information Protection Act*. The Commissioner acts independently of government to uphold and promote open government and the protection of personal privacy.

Under the three Acts, the Commissioner:

- Resolves access to information appeals and complaints when government or health care practitioners and organizations refuse to grant requests for access or correction,
- Investigates complaints with respect to personal information held by government or health care practitioners and organizations,
- Conducts research into access and privacy issues,
- Comments on proposed government legislation and programs and
- Educates the public about Ontario's access and privacy laws.

Information and Privacy
Commissioner of Ontario

Commissaire à l'information et à la
protection de la vie privée de l'Ontario

Information and Privacy Commissioner of Ontario
2 Bloor Street East, Suite 1400
Toronto, Ontario
Canada M4W 1A8

Website: www.ipc.on.ca
Telephone: 416-326-3333
Email: info@ipc.on.ca

June 2016