‘Best Practice’ Guidelines for Managing the Disclosure of De-Identified Health Information

Prepared by the:
Health System Use Technical Advisory Committee
Data De-Identification Working Group

October 2010

For more information on this document, please contact:

Canadian Institute for Health Information
495 Richmond Road, Suite 600
Ottawa, ON K2A 4H6
Phone: (613) 241-7860 (Communications)
Fax #: (613) 241-8120
Email: communications@cihi.ca
www.cihi.ca

Table of Contents

Table of Figures

1 Executive Summary

Health data in Canada, as in other countries, are used for a wide range of legally authorized purposes including the delivery of health programs and services, management of the health system and various clinical programs, public health monitoring and research. These uses require access to data in a variety of forms ranging from fully identifiable, record-level data to aggregated, summary-level data.

It is a basic principle to use health data in the least privacy intrusive way in accordance with the stated management, analytical or research objectives. There must be legal authority to use the data and all uses of identifiable data must comply with applicable privacy laws.

Often data need to be at the record-level but the identity of the individuals is not required to achieve management, analytical or research objectives. In these cases, the data can be 'de-identified' and these data are commonly referred to as 'de-identified data'.

Given the requirement to comply with the applicable privacy laws and the importance of being able to use health data for a wide range of purposes, it is essential that the processes to de-identify health data be effective and the related risks be managed.

The purpose of this paper is to identify current 'best practices' and to develop a guideline that outlines a process for data de-identification and management of risk, in the context of third party requests for disclosure, without consent, of record-level, health data. It is important to recognize that best practices need to be flexible and adaptable to various contexts, and may also evolve from time to time so as to be responsive to new, emerging technologies.

The primary audience for the guidelines includes health ministries, data custodians, and health data users for potential incorporation into their practices. A secondary audience includes interested parties such as health research funders.

The process is summarized as follows, specifically, the Data Provider (or Custodian):

1. Receives the disclosure request from a Data Requestor and reviews it collaboratively with the Data Requestor to ensure that it is complete, compliant and clearly states the analytical needs and planned data use. This is an important step since it:
    - Builds rapport between the parties
    - Clarifies the Data Requestor's objectives, analytical needs and data use, and
    - Provides an opportunity to clarify the expectations and obligations of each party in order to ensure proper data use, disclosure and management
    - Helps to assess re-identification risks, to establish the appropriate de-identification techniques and to determine necessary mitigating controls

2. Assesses the risk of re-identification based upon a thorough review of the disclosure request including the types of data requested, planned data use, Data Requestor's privacy and security policies, etc.
3. Establishes the appropriate de-identification techniques, iteratively applies each technique and re-assesses the re-identification risk until it is reduced to an acceptable level. (If the risk of re-identification cannot be reduced to an acceptable level, then the Data Provider can consider additional mitigating controls to manage the risk.)
4. Executes the required mitigating controls in a data sharing agreement[1] once the re-identification risk has been acceptably reduced. These controls work in conjunction with de-identification techniques to minimize the re-identification risk.
5. Discloses the data and monitors the Data Requestor's information usage as appropriate. This begins once there is a data sharing agreement in place.

There are also various decision points along the way where the Data Provider can decide to continue or exit the process and decline the disclosure request. The number and type of de-identification techniques can vary for each disclosure request and the formality and complexity of the overall process is commensurate with the re-identification risks associated with the disclosure.

The guidelines include a number of appendices that provide more details to support the 'best practice' process. These include:

- Sample disclosure request employing data de-identification techniques
- Checklists for reviewing disclosure requests
- Description of 'best practice' de-identification techniques
- Structured methodology for estimating re-identification risk levels
- Examples of alternatives to traditional disclosure
- List of applicable privacy statutes, regulations and policies by province
- Brief description of some commercially available, automated de-identification tools
- Glossary of terms
- List of reference documents

Health System Use (HSU) of data refers to the use of health information to improve health of Canadians and the health care system. It supports the delivery of care and patient outcomes.

[1] The term 'data sharing agreement' is used to denote an agreement between the Data Provider and Data Requestor that documents the expectations and obligations of each party vis-à-vis data use, disclosure and management. It can take various forms including a letter of authorization, a memorandum of understanding, a formal legal agreement, etc.

2 Introduction

In Canada there is a tradition of using health care data to understand and improve the health of Canadians and the Canadian health care system. This has been accomplished by leveraging the use of health data for variety of purposes, including the management of clinical programs and services, broader health system management purposes such evaluation and planning, monitoring the health of the public and research.

While health care data offer significant benefit, it is understood that the uses of health data need to respect individual privacy. Users of health data may require access to data in a variety of forms ranging from record-level data to summary-level data. However, even when record level data are required, the identity of the individuals is often not required to achieve the objectives. In these cases, the data can be 'de-identified'.

The purpose of this paper is to consolidate current 'best practices' and to develop a guideline that outlines a process for data de-identification. It outlines the following five-step process:

- Reviewing the HSU disclosure request
- Assessing re-identification risks
- Establishing and applying de-identification techniques
- Executing mitigating controls regarding the request
- Disclosing data and monitoring usage

This paper was developed by jurisdictional and industry experts from the following organizations:

- Manitoba Health
- Ontario Agency for Health Protection and Promotion
- Newfoundland and Labrador Centre for Health Information
- Institute for Clinical Evaluative Sciences
- Children's Hospital of Eastern Ontario Research Institute and University of Ottawa
- Canada Health Infoway
- Canadian Institute for Health Information

It was developed under the auspices of the Health System Use - Technical Advisory Committee, a collaborative effort of the federal/provincial/territorial ministries of health, the Canadian Institute for Health Information (CIHI) and Canada Health Infoway.

3 Scope and Underlying Principles

The issues involved in 'health system use' are complex. Thus, the scope is limited to the disclosure, without consent, of record-level health information that is identifiable or potentially re-identifiable for uses such as:

- Clinical program management, i.e., improving front-line health care programs and services
- Health system management, i.e. improving the effectiveness and efficiency of the health care system
- Monitoring public health, i.e., understanding the health of the public
- Research, i.e., identifying improvements to medical treatments and programs of care, or to better understand the health of the population, the factors influencing health, and the performance of the health care system

The scope is summarized as follows:

In scope	Out of scope
Record-level data	Aggregate-level data
Disclosures without consent	Disclosures for which consent is required or exists
EHR information including feeder systems and EMR data	Health information from other source systems such as bio-banks and genetic data
Health information that is identifiable or potentially re-identifiable	Anonymous or aggregated health information

One challenge is the breadth and complexity of the privacy statutes, regulations and policies across Canada governing the disclosure of health information. Although the discussion has been limited to high-level, 'best practice' guidelines, there is some value provided for all jurisdictions through references to applicable privacy statutes, regulations and policies.

The 'best practice' guidelines are based upon the following underlying principles:

- Disclosure of health information for 'health system use' is best made with the minimum amount of data and at the highest degree of anonymity while still meeting management, analytical or research objectives
- Organizations and individuals responsible for handling requests for disclosure of health information for 'health system use' need to be:
  - Well-informed and up-to-date on de-ID principles and methods
  - Capable of applying current de-ID tools and techniques
  - Compliant with statutory requirements related to de-ID
- An assessment of re-identification (re-ID) risk is completed iteratively in conjunction with the application of appropriate de-ID techniques. Residual re-ID risk is managed by

implementing mitigating controls to minimize unintended and unauthorized data use and disclosure

- It is important that the risk assessment process be consistent, repeatable and transparent
- Requests for disclosure of identifiable or potentially re-identifiable health information from individuals or organizations for 'health system use' should include a legal analysis to ensure disclosure is lawfully permitted
- The formality and complexity of the entire process for managing disclosure of de-identified health information is commensurate with the re-identification risks associated with the disclosure

4 'Best Practice' Process Model Overview

The following section provides an overview of a five-step, 'best practice' process for managing the disclosure of de-identified health information. It begins with the receipt, collaborative review and approval of the disclosure request; continues with the iterative assessment of re-ID risk and application of data de-ID techniques; follows on with implementation of mitigating controls; and concludes with the disclosure of the data and ongoing post-disclosure monitoring. The process employs risk assessment and data de-ID techniques that have broad applicability for all health information.

4.1 Process Model Assumptions

The proposed process model assumes that:

- The goal is to minimize the probability of an individual being re-identified and the expected number of re-identified data records
- Data are considered to be de-identified if the risk of re-ID is at an acceptable level
- Data are manipulated using prescribed techniques until an acceptable level of re-ID risk has been attained … if possible
- This creates an iterative loop of *applying de-ID techniques and re-assessing re-ID risk* until an acceptable level of re-ID risk has been attained
- If an acceptable level of re-ID risk cannot be attained, then the parties can negotiate the use of additional mitigating controls to manage the risk
- The process provides for a number of disclosure decision points either to continue or to notify the Data Requestor that the request is declined

4.2 Process Model Flow

- The five-step process is shown in Figure 1 and is summarized as follows:

Data Provider	**Exit Criteria**
1. Receives the disclosure request and reviews it collaboratively with the Data Requestor to ensure that it is complete, compliant and clearly states the analytical needs and planned data use	▪ If acceptable, go the next step ▪ If incomplete, non-compliant or unclear, work collaboratively to remedy the deficiencies ▪ If deficiencies cannot be remedied, inform the Data Requestor that the request is declined and provide the rationale
2. Assesses the risk of re-ID based upon a thorough review of the disclosure request	

Data Provider	**Exit Criteria**
3. Establishes the appropriate de-ID techniques, iteratively applies each technique and re-assesses the re-ID risks until an acceptable level of re-ID risk has been attained	▪ If an acceptable level of re-ID risk has been attained, go to the next step ▪ If an acceptable level of re-ID risk has NOT been attained, then: – Continue to de-ID and re-analyze until an acceptable level of re-ID risk has been reached – If an acceptable level of re-ID risk cannot be attained then *either* negotiate the use of <u>additional</u> mitigating controls to manage the risk *or* inform the Data Requestor that the request is declined and provide the rationale
4. Executes the required mitigating controls when the re-ID risk has been has been sufficiently reduced and/or manageable	▪ If a satisfactory data sharing agreement has been executed, proceed to the next step ▪ If a satisfactory data sharing agreement cannot be executed, inform the Data Requestor that the request is declined and provide the rationale
5. Discloses the data and continues to monitor the Data Requestor's information usage	

- Appendix A contains a brief, step-by-step example of a health data disclosure request to a provincial Ministry of Health from an academic researcher

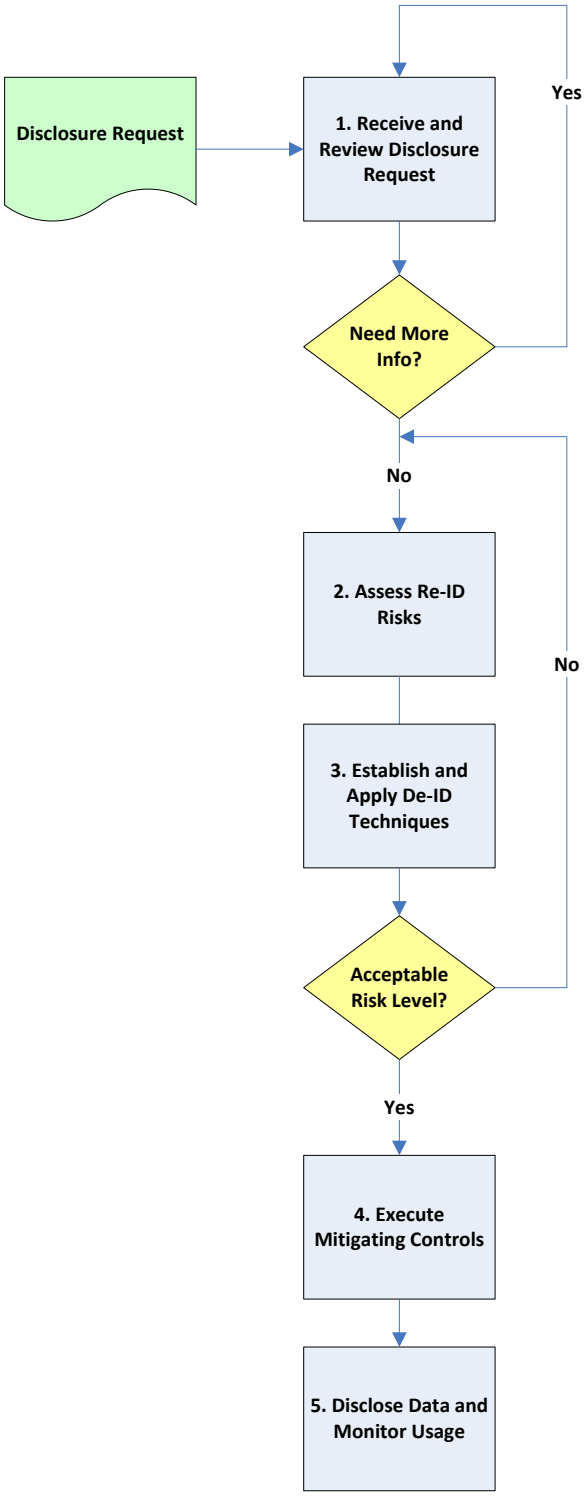

Figure 1 – Proposed De-ID Process Model

5 Receive and Review Disclosure Request

This section describes step 1 in the proposed de-ID process where the Data Provider receives the disclosure request and reviews it collaboratively with the Data Requestor to ensure that it is complete, compliant, and states the analytical needs and planned data use. Disclosure decision outcomes for this step are described at the end of the section.

5.1 Basic Principles

- Individuals or organizations need to formally submit 'health system use' data disclosure requests that are compliant with applicable privacy statutes, regulations and policies
- Reviewing the disclosure request is an important, collaborative process between the Data Provider and the Data Requestor. This collaboration:
  - Builds rapport between the parties
  - Clarifies the Data Requestor's objectives, analytical needs and data use, and
  - Provides an opportunity to clarify the expectations and obligations of each party in order to ensure proper data use, disclosure and management
  - Helps to assess re-ID risks, to establish the appropriate de-ID techniques and to determine necessary mitigating controls
- The formal disclosure request includes contact information for all involved including the applicant, project/initiative director and all co-applicants
- If applicable, disclosure requests for research purposes should include pertinent Research Ethics Board (REB) documentation for consideration.
- The formal disclosure request is signed and dated by the Data Requestor and others as appropriate, such as a supervisor for a student request
- The Data Provider and the Data Requestor work collaboratively to remedy an incomplete, non-compliant or unclear disclosure request

5.2 Suggested Disclosure Request Content

The disclosure request can be hard copy or web-based. Its contents can include:

- Contact information for each person named in the disclosure request including why their access is necessary, their role in the project/initiative, their related qualifications, and whether or not they will be accessing the record-level data
- Description and duration of project/initiative including objectives, methodology, participating organizations including their role, e.g., collaborations, data access, funding, etc
- Requested Information which can include:
  - Data variables requested, the rationale and frequency required for each
  - Description of any linkage of the requested data variables to other databases including the linkage processes

  - The Data Requestor's documented authority and approval to use the linking databases
  - Identification of the funding sources including reporting obligations to the funder
  - If the Data Requestor is asking for inclusion of identifiers, then provide rationale why the research cannot reasonably be accomplished with de-identified data
  - A description of the potential risks to individuals or groups, such as first nations, small communities, underserviced areas, disadvantaged or vulnerable populations or those with rare medical conditions
- Evidence of the Data Requestor's privacy and security policies such as:
  - A description of where the requested data will reside (address, room number) and how the data will be protected (administrative, technical and physical)
  - A copy of privacy policies, responsibilities, accountabilities and reporting including contact information of the Chief Privacy Officer (CPO), or equivalent
  - Date when last privacy and/or security audit of organization was conducted, including a summary of the outcome
  - A description of how privacy compliance will be monitored
  - The names of all those who will be authorized to access the requested data
  - If the Data Requestor is asking for inclusion of identifiers, the plan and timeline to fully de-identify or dispose of the data after the analysis is completed
- For academic research projects or initiatives that will be published in academic or peer-reviewed journals, provide pertinent REB documentation that can include:
  - Approved or pending applications
  - Approval letters and supporting documentation of requirements
  - Conditions imposed by the REB

5.3 Suggested Disclosure Request Review

Data Requested and Disclosure Request Content

- Review the data requested and disclosure request content as listed above
- Refer to Appendix B for a checklist of potential questions to be asked

Legal Authority and Compliance with Organizational Privacy Policies

- Does the Data Requestor have the authority to access the requested data without consent of the individual?
- Does the Data Provider have the legal authority to collect as well as disclose the requested data?
- Are there collection and disclosure limits imposed by the Data Provider's privacy statutes, regulations and policies?
- Are there limits to subsequent use and disclosure imposed by the Data Requestor's privacy statutes, regulations and policies?
- Has the Data Requestor notified the Privacy Commissioner (or other pertinent bodies as required in the jurisdiction) if they are planning to link the requested data to other databases?

- Refer to the applicable jurisdictional privacy statutes, regulations and policies provided in Appendix F

Other Criteria

- If applicable, is the protocol approved by the Research Ethics Board congruent with what has been described by the Data Requestor in their disclosure request?
- If the Data Requestor is doing the project/initiative on behalf of a jurisdiction, has the jurisdiction provided a letter of support?
- If other organizations need to be involved in the project/initiative, have they provided a letter of support?
- Is the Data Requester obligated to disclose the source data to journals etc. so the work can be verified?
- Are there current and pertinent (data sharing or research) agreements between the Data Provider and Data Requestor already in place?

5.4 Disclosure Decision

Based upon the analysis of this disclosure request the Data Provider can take one of the following courses of action:

1. If request is acceptable then proceed to the next step in the process
2. If the request is incomplete, non-compliant or unclear work collaboratively with the Data Requestor to remedy the deficiencies
3. If the request is unacceptable or the deficiencies cannot be remedied, inform the Data Requestor that the request is declined and provide the rationale for denying the request

6 Assess Re-ID Risks

This section describes step 2 in the proposed de-ID process where the Data Provider assesses the re-ID risks related to the data disclosure. Step 2 is tightly integrated, and completed iteratively with step 3 described in the next section.

6.1 Basic Principles

- The purpose of risk assessment is to determine how much de-ID to perform in order to reduce the risk of re-ID to an acceptable level
- It is important that disclosure requests undergo an assessment of re-ID risks both at the outset and as required over time
- Data Providers need to clearly define what constitutes a quasi-identifier (a data variable that can be used to probabilistically identify an individual) in order to select the variables to which they need to apply de-ID techniques
- It is important that an assessment of re-ID risks include variables that can infer quasi-identifiers, e.g., diagnostic codes can sometimes be used to infer gender
- Data Providers could consider establishing flexible guidelines for acceptable levels of re-ID risk that can address a range of disclosure scenarios

6.2 Background

- Identifying data variables can be classified as one of the following. For more examples refer to Appendix H:
  - Directly identifying variables can be used to uniquely identify an individual either by themselves or in combination with other readily available information. Examples can include name, phone number or email address
  - Indirectly identifying variables (quasi-identifiers) can be used to probabilistically identify an individual either by themselves or in combination with other available information. Examples can include sex, date of birth or postal code
- However these distinctions can vary depending upon the context, i.e., a variable can be directly identifying in one instance and a quasi-identifier in another. For example, a postal code can be a directly identifying variable or quasi-identifier depending upon the location. In general if a specific variable is of analytic importance, then reduce the identifiability of other variables to preserve the usefulness of the specific variable
- Another category of variables, sensitive variables, is not directly manipulated during de-ID. They contain sensitive health information about the individual. Examples can include sexual orientation or diagnosis codes. If a dataset has sensitive variables then it will require more de-ID
- The objective is creation of de-identified data that minimizes the probability of an individual being re-identified and the expected number of re-identified data records

- Data are considered de-identified if the risk of re-ID is acceptable, which can depend upon a number of factors including the:
  - Data Provider's tolerance of re-ID risk
  - Sensitivity of the data and the potential harm resulting from unintended or unauthorized data use and subsequent disclosure
  - Data Requestor's past practices, intentions, capacity to re-identify, internal privacy and security practices and access to additional data sources
- A good re-ID risk assessment approach considers these dimensions of risk that attempt to discern when a particular disclosure is risky. If the risk of re-ID is too great, then reduce the risk by performing more de-ID
- The four levels of decreasing identifiability are provided in the table below. This list also indicates a deceasing probability of re-ID

State	Description
1. Identifiable data	The data have directly identifying variables or sufficient quasi-identifiers that can be used to identify the individual.
2. Potentially De-identified data	Manipulations have been performed on the identifying variables but attempts to disguise the quasi-identifiers may be insufficient. The data may not be fully de-identified, may be partially exposed and may represent a re-ID risk.
3. De-identified data	An objective assessment of re-ID risk has been done and it is concluded that all directly identifying variables have been adequately manipulated and quasi-identifiers adequately disguised to ensure an acceptable level of re-ID risk.
4. Aggregate data	These are summary data such as tables or counts, where there are no identifying variables or quasi-identifiers.

6.3 Approaches to Managing Risk

- The purpose of the risk assessment is to decide how much de-ID to perform. The more de-ID that is performed the lower the risk of re-ID and the lesser the requirement for other mitigation controls. Then again, de-ID can distort the data and reduce data quality. The challenge becomes ensuring adequate de-ID while still safeguarding the data against re-ID. There are two general approaches for managing re-ID risk:

Heuristics

- Heuristics are essentially 'rules of thumb' and ideally are evidence-based. They are suitable before data are collected or when it's impossible to access the data. In general there are two types of heuristics, i.e., those based on:

- Uniqueness and rareness in the population, e.g., those with a rare medical condition
- Record linkage using public registries, e.g., geo-codes for people living in a geographic area with a small population

Analytics

- These methods analyze the data itself in order to measure re-ID risk and to decide how best to de-identify the data. The main principles of data de-ID are that the estimation is risk-based and considers the usefulness of the data. Steps in this process can be summarized as follows:
  - Determine the quasi-identifiers (privacy legislation often defines some of these)
  - Set an acceptable level of risk
  - Evaluate the risk of re-ID
  - Iteratively apply de-ID techniques and re-evaluate risk
  - Stop when the acceptable level of risk has been attained, if possible

6.4 Evaluating Re-ID Risk

- The evaluation of the level of risk of re-ID is a complex and multifaceted and can involve qualitative and/or quantitative approaches:

Qualitative

- A qualitative risk assessment is subjective (scored as low, medium or high) and depends primarily on the context under which the data are to be disclosed
- Much of the information on which to base the risk assessment comes from a review of the Disclosure Request
- Factors to consider can include the following. Refer to Appendix B for a more complete list:
  - Has the Data Requestor previously worked with the Data Provider?
  - Is there an existing data sharing agreement between the parties?
  - Are the requested data highly detailed or sensitive?
  - Where will the Data Requestor store the requested data?
  - Does the Data Requestor have adequate administrative, technical, and physical security controls to protect the requested data?
  - To what additional databases does the Data Requestor have access?
  - What is the impact if there was an unintended or unauthorized use and subsequent disclosure of the requested data?
  - Has the Data Requestor tried to limit the number and types of data variables requested?

Quantitative

- A quantitative risk assessment can involve evaluating the probability of uniquely identifying an individual as a measure of the theoretical risk of re-ID

- For example, one may want to reduce the number of data records with a unique combination of potentially re-identifying variables, i.e., increase k-anonymity
- The risk of re-ID can be measured as the probability that someone will find the correct identity of a single individual:
  - Assume that someone is looking for a specific 45-year female
  - If there are five 45-year old females in the dataset, the probability of re-ID is 1 in 5 or 20%
  - Assume that all ages are rounded (reduction in detail) from years to decades and that there are 25 females in their 40's
  - There are now twenty-five 40-something females in the dataset and the probability of re-ID is 1 in 25 or 4%

- Appendix D outlines a structured methodology that can be used for estimating the risk levels and establishing how much de-ID is required

---

- The qualitative and/or quantitative evaluation of re-ID risk is done iteratively with the application of appropriate de-ID techniques (described in the next section) until the risk of re-ID has been sufficiently reduced
- If the application of de-ID techniques alone does not adequately reduce the re-ID risk then the parties may consider including additional mitigating controls in the data sharing agreement to manage the risk

7 Establish and Apply De-ID Techniques

This section describes step 3 in the proposed de-ID process where the Data Provider establishes and applies the techniques appropriate to the risks and the planned data use. This step is done iteratively with step 2 described earlier. In other words, steps 2 and 3 are iterated jointly until the risk of re-ID has been sufficiently reduced. Disclosure decision outcomes for this step are described at the end of the section.

7.1 Basic Principles

- Disclosures are best made with the minimum amount of data and at the highest degree of anonymity while still meeting the management, analytical or research objectives
- Organizations and individuals responsible for handling disclosure requests need to be well-informed and up-to-date on de-ID principles and methods, capable of applying current de-ID tools and techniques and compliant with statutory requirements related to de-ID
- There are a number of common techniques that can be applied to data in order to reduce re-ID risk (refer to Appendix C):
  - *Reduction in Detail* is most often used for quasi-identifiers
  - *Substitution* and *Pseudonymization* are most often used for direct identifiers
  - *Suppression* can be used for both direct identifiers and quasi-identifiers
  - *Random Addition of 'Noise'* can also be used for both but is not preferred since it distorts the data
- No single technique can independently meet the diverse data de-ID needs related to 'health system use'
- The appropriate number and types of techniques vary for each disclosure request
- De-ID can be supplemented by other mitigating controls in the data sharing agreement to manage the risk

7.2 Applying De-ID Techniques

- In order to determine what de-ID technique(s) will be most effective, it is important that the parties first thoroughly discuss the analytical needs
- De-ID generally begins with *Reduction in Detail* followed by *Suppression*. These are the most accepted techniques in practice, the least expensive to apply, the easiest to understand and easiest to predict re-ID risk
- The other techniques are more expensive to apply, more difficult to understand and more difficult to predict re-ID risk

Manipulating Direct Identifiers

- Manipulating the direct identifiers in the dataset can be accomplished by means of *Suppression, Substitution* or *Pseudonymization*, as described in Appendix C

- Examples of direct identifiers can include name, phone number, email address, health insurance card number, credit card number and social insurance number
- Automated de-ID tools that manipulate direct identifiers are discussed in Appendix G, and can include:
  - Oracle Data Masking Pack
  - Camouflage
  - Informatica Data Privacy
  - Data Masker
  - IBM Optim Data Privacy Solution

Determining and Disguising Quasi-Identifiers

- Disguising the quasi-identifiers in the dataset can be accomplished by means of *Reduction in Detail* and *Suppression*, as described in Appendix C
- However, first It is necessary to ascertain the quasi-identifiers and for each:
  - Establish why it is required in the analysis
  - Determine the de-ID precision hierarchy, e.g., dates can be expressed as day/month/year, month/year, quarter/year, year, decade, etc
  - Apply the best de-ID technique that would still meet the data needs
- Examples of quasi-identifiers can include sex, date of birth or age, geo-codes, first language, ethnic origin, aboriginal identity, total years of schooling, marital status, criminal history, total income, visible minority status, profession, health event dates, health-related codes, country of birth, birth weight, and birth plurality
- Automated de-ID tools that disguise quasi-identifiers are discussed in a Appendix G, and can include:
  - PARAT – Privacy Analytics Risk Assessment Tool
  - µ-Argus – Anti-Re-ID General Utility System

7.3 De-ID Examples

- The figure below provides examples for applying the techniques with geo-code, numeric and alpha variable data types

	Geo-Code	Numeric	Alpha
Reduction in Detail	• Reduce postal codes to first 3 characters, i.e., Forward Sortation Area	• Round birth dates to year • Express dates relative to a milestone date	

	Geo-Code	Numeric	Alpha
Suppression	As a rule of thumb, suppress geo-codes when they contain five observations or less	As a rule of thumb, suppress numbers when they contain five observations or less	As a rule of thumb, suppress alpha variables when they contain five observations or less
Substitution	• If postal code is manipulated then ensure telephone area code is consistent	• If health card number is manipulated then ensure the new number can pass a checksum validation check	• Select new names in same proportion as in general public • If surname is manipulated then ensure the new name has the same number of characters and ethnicity
Pseudonymization	• Can be applied to most geo-data	• Can be applied to most numeric data	• Can be applied to most alpha data

Figure 2 – De-ID Examples by Variable Data Type

7.4 Disclosure Decision

1. If the evaluated risk of re-ID has been reduced to an acceptable level, then proceed to the next step in the process
2. If the evaluated risk of re-ID has NOT been reduced to an acceptable level, then:
   a. Continue to de-identify the data and re-analyze the re-ID risk until an acceptable level of risk has been achieved
   b. If it is concluded that the evaluated risk of re-ID cannot be reduced to an acceptable level, then *either* negotiate the use of additional mitigating controls to manage the risk *or* inform the Data Requestor that the request is declined and provide the rationale

8 Execute Mitigating Controls

This section describes step 4 in the proposed de-ID process where the Data Provider executes the mitigating controls when the re-ID risk has been sufficiently reduced and/or manageable. Disclosure decision outcomes for this step are described at the end of the section.

8.1 Basic Principles

- Mitigating controls work in conjunction with de-ID techniques to minimize the re-ID risk
- They are included in the data sharing agreement between the Data Provider and Data Requestor that documents the expectations and obligations of each party vis-à-vis data use, disclosure and management
- The form and complexity of the data sharing agreement is commensurate with the associated re-ID risks and can take various forms including letter of authorization, memorandum of understanding, formal legal agreement, etc.

8.2 Data Sharing Agreement

Execute a data sharing agreement between the Data Provider and Data Requester regarding data confidentiality and security that can include one or more of the following:

- Data Provider's right to audit the Data Requestor for compliance to the data sharing agreement
- For the Data Requestor:
  - Limits on the use and disclosure of the data without the Data Provider's prior written approval
  - Penalties for any unintended or unauthorized use and disclosure of the data
  - Provision of liability insurance and agreement to indemnify the Data Provider from damages related to any unintended or unauthorized use and disclosure of the data
  - Agreement to safeguard and protect the data from unauthorized access
  - Limits to linking the data to other databases without the Data Provider's prior written approval
  - Commitment to NOT attempt to re-ID the data
  - Commitment to NOT publicly release small cells (with less than an agreed number of observations)
  - Commitment to properly dispose of the data and to attest to the destruction
  - Agreement to provide the analytical code, if practicable
  - Agreement to have its staff to swear an oath with the Privacy Commissioner to safeguard data privacy, if applicable

- For researchers, an agreement to:
  - Submit a detailed research plan to the Data Provider
  - Limit source data and other disclosures to academic journals
  - Obtain Research Ethics Board approval as per institutional policies
  - Allow the Data Provider to review the manuscript before it is published to review data use for privacy purposes

8.3 Disclosure Decision

The Data Provider can take one of the following courses of action based upon the outcome of the creation and execution of the data sharing agreement:

1. If a satisfactory data sharing agreement has been executed, proceed to the next step
2. If a satisfactory data sharing agreement cannot executed, inform the Data Requestor that the request is declined and provide the rationale

9 Disclose Data and Monitor Usage

This section describes step 5 in the proposed de-ID process where the Data Provider discloses the data and continues to monitor the Data Requestor's information usage. Appendix E outlines several, alternate approaches to traditional disclosure.

9.1 Basic Principles

- In the data sharing agreement, the Data Provider specifies the right to audit the Data Requestor for compliance with the contractual terms of the data sharing agreement

9.2 Disclosure Process

- Disclose the data once there is an adequate data sharing agreement in place between the Data Provider and the Data Requestor
- Note that in cases where an audit was required to demonstrate compliance with good security and privacy practices, a security audit certificate may be needed before disclosure

9.3 Monitoring Process

On an ongoing basis the Data Provider can:

- Audit the Data Requestor to ensure compliance with the data sharing agreement
- Receive a copy of the Data Requestor's analytical code if practicable
- Ask Data Requestor to submit a form at the end of the project to state that the project is completed and confirm data retention or destruction
- In the case of a research project review the manuscript before it is published

10 Appendix A – Sample Disclosure Request Employing De-ID

This section provides a brief, step-by-step example of a health data disclosure request to a provincial Ministry of Health from an academic researcher.

10.1 Receive and Review Disclosure Request

A provincial data custodian (Ministry of Health) receives a data disclosure request from an academic researcher (Dr. Anon). Dr. Anon has completed the requisite disclosure request form found online on the Ministry's website. Along with Dr. Anon's contact information and list of research collaborators, the following information was submitted:

- Title of research study
- Funding agency
- Study protocol (including hypotheses and complete methodology)
- List of databases and data variables required as follows:

Database	Years of Data	Data Variables Requested
Health Registry	2000 – 2009	• Personal Health Identification Number (PHIN) • Date of Birth • Date of Death or coverage cancellation date
Physician Claims	2000 – 2009	• PHIN • Diagnostic code (ICD9) • Service Date
Prescription Claims	2000 – 2009	• PHIN • Drug Information Number (DIN) – to identify test strips • Drug-dispensed Date

Dr. Anon's application for data also indicates that the data will be used strictly for this specific academic project with the intention of publishing the results in an academic journal.

The project has already received approval from the University's REB. Dr. Anon states that the data will be stored on a personal computer in a locked office on campus.

10.2 Assess Re-ID Risks

Upon review of the disclosure request, the Ministry identifies several directly identifying variables are being requested including the real PHIN and the full date of birth/death.

Dr. Anon is contacted and agrees that for the purpose of these analyses, a study ID can replace the PHIN, and the date of birth/death can be replaced with the individual's age group (within a 5 year grouping).

The Ministry is also concerned that the physician claim data (diagnostic code and service date) and the prescription claim data (DIN and drug-dispensed date) could be used to re-identify one or more individuals. Dr. Anon points out that these data are critical to his analysis and prefers that they not be de-identified. The Ministry agrees to disclose these potentially identifiable data but plans to address the risk of re-ID through the data sharing agreement with Dr. Anon.

The data protection and storage process described by Dr. Anon is also an issue. The Ministry's policies stipulate that greater data security is required. A copy of the University's policies on information storage and protection is requested and reviewed. The Ministry requires that the data be stored only on a secure network drive of the University mainframe to which only Dr. Anon (or a designate) has password-protected access.

10.3 Establish and Apply De-ID Techniques

Staff from the Ministry extracts the data requested by Dr. Anon and applies the following de-ID techniques based upon the analytical needs and planned data use:

- Substitution – Replace the PHIN with a study ID while ensuring that multiple records for the same individual are replaced consistently to ensure referential integrity
- Reduction in Detail – Round the value for date of birth/death to just the year birth/death and then further into the appropriate 5-year category
- Other dates are not altered since they are critical to the analysis

10.4 Execute Mitigating Controls

The Ministry enters into a formal, legal data sharing agreement with Dr. Anon respecting the data that are ultimately disclosed. The data sharing agreement has been approved by the legal counsel of both organizations, constitutes the conditions under which the Ministry discloses data, and describes any access and use restrictions. Specifically, Dr. Anon agrees to:

- Provide the manuscript to the Ministry for review before submission to journal
- Store the data only on a secure network drive maintained by the University
- NOT attempt to re-ID the data
- NOT publicly release small cells (with less than 5 observations)
- Destroy all copies of the data after a set period of time and to allow the Ministry of Health to audit for the data's destruction

10.5 Disclose Data and Monitor Usage

The approved data are encrypted and password-protected, saved onto DVD, and sent via bonded courier to Dr. Anon's office. The password is sent to Dr. Anon via email.

Upon completion of the analyses, as per the conditions outlined in the data sharing agreement:

- Dr. Anon prepares a manuscript for submission to an academic journal and, prior to its submission to the journal, submits it to the Ministry for review of potential breaches of confidentiality (e.g., inclusion of small cell-sizes), and for appropriate use of the data in accordance with the original, approved protocol
- In addition, the Ministry reviews the description of the data sources for appropriate representation of the Ministry and its data

No issues are identified and Dr. Anon's manuscript is submitted and ultimately published in a top-ranking academic journal.

After the set period of time, Dr. Anon arranges for the University to destroy all copies of the data and confirms the data's destruction via e-mail to the Ministry.

11 Appendix B – Disclosure Request Check Lists

11.1 Data Requested

The following questions are suggested during a discussion of Data Requestor's analytical needs. It is also important to inform the Data Requestor of any data quality and limitation issues that become evident during the discussion:

- What is the entire body of data that is being used in this project/initiative?
- Has the Data Requestor tried to limit the number and types of data variables?
- Has the Data Requestor justified the use of each requested data variable?
- To what additional data does the Data Requestor have access (thus increasing the probability of re-ID)?
- What data variables will be used to link with data from other databases?
- Who will perform the linkages?

11.2 Request Content

Review the disclosure request as defined previously in section 5.2 as follows:

- Have all relevant participants been identified and included in the request?
- Has the Data Requestor adequately justified the need for all the data requested?
- Will the Data Requestor accept a random sample (subset) of the data rather than the entire dataset?
- Has the methodology for the use of the requested data been clearly articulated?
- Are the privacy and security policies of the Data Requestor's organization adequate?
- Has the Data Requestor demonstrated sufficient administrative, technical, and physical security controls to protect the requested data?
- What is the Data Requestor's track record of safeguarding data?
- Does the Data Requestor's organization have adequate privacy accountability structures, compliance reporting and privacy audit practices?

11.3 Project-Specific Privacy Impact Assessment

As an alternative to a formal disclosure request, one Data Provider currently asks Data Requestors to complete a form that poses the same questions one would ask in a project-specific Privacy Impact Assessment (PIA) including:

- Project purpose
- Databases being used or available for use
- Potential data linkages including rationale
- Participants and roles
- Public benefit of project
- Estimate of potential harm of unintended/unauthorized data use and disclosure
- Alternative data sources available
- Project and data retention plan and timeframe

- Project financial information including funding sources and obligations
- Approvals required and obtained
- Signatures of approval and confirmation

12 Appendix C – De-ID Techniques

12.1 Reduction in Detail

- This is the most common technique used and involves reduction in the detail of the data through rounding or collapsing the data values into larger categories
- The objective is to reduce the number of data records with a unique combination of quasi-identifier values
- Some of the common data values reduced in detail are dates and geo-locators

  Dates
  - Birth dates can be rounded to year of birth. Ages are less identifying than birthdates but can still pose high re-ID risk. De-ID is more thorough if age groups or categories are used, but the data then become less informative for analysis
  - Clinical event dates can be expressed relatively to milestone date, e.g., days from diagnosis
  - Be cautious with dates that can infer other dates, e.g., autopsy date (date of death), mother's discharge date (baby's birth date)
  - **Geo-Codes**
  - Postal codes are highly identifying. Provide no greater detail than required to accomplish the designated purpose. It may be necessary to de-identify other variables to a greater extent to mitigate the increased risk of re-ID with more detailed postal code information
- It is often necessary to iteratively reduce the detail for certain quasi-identifiers until one achieves an acceptable compromise between sufficiently reducing the likelihood of identifiability and retaining the usefulness of the data

12.2 Suppression

- Suppression can be done at the level of a variable, a record or a cell.

  Variable Suppression
  - This technique involves the removal or withholding of a data variable's values
  - All other variables in the record, i.e., those that are not quasi-identifiers, remain untouched
  - It may not always be plausible to suppress some variables because that will reduce the utility of the data

  Record Suppression
  - If variable suppression and reduction in detail techniques do not adequately de-identify the data then the alternative is the removal and withholding of the data records that create a high re-ID risk
  - However extensive suppression can introduce a high level of distortion in some types of analysis since the loss of records is not completely random and may reduce the usefulness

Cell Suppression

- A special case of suppression concerns 'outlier variables' such as rare diagnoses, uncommon medical procedures, some occupations or distinct deformities that can uniquely identify an individual
- Specific quasi-identifier values are suppressed such that the amount of suppression is minimal but still maintains an acceptable re-ID risk
- Entire variables and entire records are not suppressed
- This is the preferred suppression method because it reduces the amount of distortion to the data

12.3 Random Addition of 'Noise'

- This technique adds random 'noise' to the values of a variable in order to disguise its true value. It is also called data perturbation or scrambling
- It often works best with numeric or structured variables that can be randomly altered within a given range. For example:
  - Add or subtract a random number of days to a birth date within a defined range to disguise the date but preserve the age
  - Add or subtract a random number of inches to a height within a defined range to disguise the height but preserve a height range
  - Alter a postal code to a randomly selected nearby postal code to disguise the code but preserve the general location, e.g., consider shifting geo-codes by 0.5 km or more
- Data Requestors dislike this approach because they cannot trust the data anymore. For this reason *Reduction in Detail* that does not randomly alter a variable is preferable

12.4 Substitution

- This technique removes the association between the individual and the associated identifying data by replacing original data values with values that have been:
  - Randomly drawn from large databases (randomization of data values), or
  - Exchanged with values in other records in the dataset (data swapping)
- When using substitution it is helpful to replace the original data values with realistic values that look and behave like the original ones. For example, replace real names and addresses with false (but real) names and addresses
- A good name substitution tool selects a fake name:
  - With the same probability that it appears in the actual population to ensure uncommon names do not appear disproportionally, or
  - Ensuring that the replacement name has the same number of characters or that it is of the same ethnicity as the original name
- A good health insurance, social insurance or credit card substitution tool selects a fake number:
  - That will pass a validation check such as a 'modulus 10' checksum, or

  - Ensures the replacement number is the same card type or from the same financial institution
  - Note: To facilitate linkages across databases, the number generated for different records corresponding to the same individual must be consistent.
- If postal code is being substituted then manipulate the telephone number area code consistently
- A further consideration is ensuring that multiple records for the same individual are substituted consistently to ensure referential integrity
- The main drawback with substitution is that it is difficult to assess the difficulty of reversing the replacements, i.e., the of re-ID risk. The Data Provider must decide whether substitutions ensure that the risk re-identifying the data is sufficiently low.

12.5 Pseudonymization

- This technique removes the association between the individual and the associated identifying data by replacing an individual's identifying data variables with one or more pseudonyms (also called 'coding' or 'alias assignment')
- For example, a record containing an individual's real name and date-of-birth could have these identifying variables replaced with the unique pseudonym '098737'
- It is important that pseudonyms not be some deterministic code, e.g., consisting of initials and date of birth but rather independently generated.
- To facilitate linkages across data records and databases, the pseudonym generated for the same individual must be consistent
- As a result, pseudonymous data records can be associated since they allow associations between sets of characteristics but not with the individual
- Pseudonymization can be performed with or without the possibility of re-identifying the individual (called reversible or irreversible respectively)
  - Reversible pseudonymization is discussed below
  - In irreversible pseudonymization, the pseudonymized data do not contain information that allows the reestablishment of the link between the individual and the pseudonymized data
- An irreversible pseudonym can be random or unique depending upon whether it is different or identical each time it is generated for the same individual:
  - When it is risky to provide a Data Requestor with access and potential linkage of different coded datasets, a random pseudonym is generated for each individual every time the dataset is disclosed
  - When the ability to link different coded databases is desired, the same, unique pseudonym is generated for each individual every time the dataset is disclosed

12.6 Reversible Pseudonymization

- In reversible pseudonymization, the pseudonyms can be linked with the individual by applying procedures typically restricted to authorized users under prescribed circumstances and protections

- Reversible pseudonymization works well for research projects because it allows data cleansing while retaining the ability to reference original identifiers
- With reversible pseudonymization, a transformation table and related set of re-ID 'keys' are generated and maintained to allow the pseudonyms to be mapped back to the original data values
- Reversible coding schemes can often involve single or double coding
  - Single coding means that identifying data values are removed and each record is assigned a pseudonym. The original values are kept in a separate transformation table with the pseudonym to allow linking back to the original data. As shown in the figure below, an individual's real name and birth date can be replaced with a unique pseudonym '098737'

Figure 3 – Single Coded Pseudonymization

  - Double coding adds a linking database that connects each pseudonym to a second pseudonym. The transformation table links the second pseudonym to the original data value
  - The linking database can be maintained in a secure location by a trusted third party to provide double protection against re-ID. As shown in the figure below, an individual's real name and birth date can be replaced with a unique pseudonym '098737' that is double coded to the value '145635'

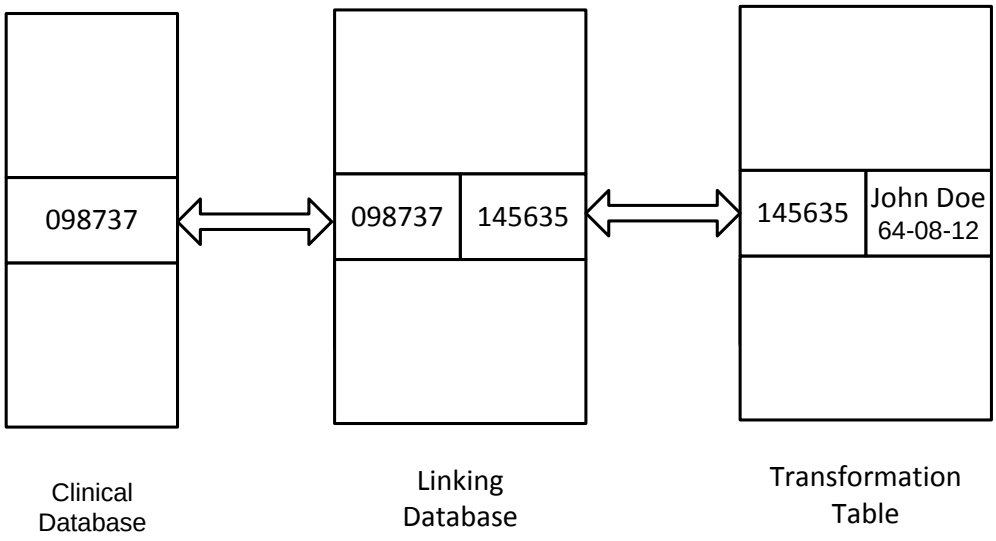

Figure 4 – Double Coded Pseudonymization

- Consider whether or not a 'key' or linking database is maintained by the Data Provider to allow validation related to concerns about data accuracy
- Consideration could be given to using a trusted third party for the generation and management of the decoding 'keys' including the:
  - The reliable and secure binding of unique pseudonyms to individuals
  - Protection of the pseudonyms from unauthorized re-ID
  - Provision of authorized re-ID of the original identifier(s) in accordance with agreed re-ID policy parameters
- The criteria for re-ID can be defined, automated, and securely managed by the trusted third party

12.7 Handling Freeform Text

- Freeform text cannot be assured anonymity with existing de-ID approaches
- All freeform text is subject to a risk analysis and a mitigation strategy for re-ID risks.
- Re-ID risks of freeform text can be mitigated through:
  - Implementing policies that prohibit freeform text from containing any identifying data variables, e.g., individual numbers and names
  - Verifying that freeform content is unlikely to contain any identifying data variables
  - Revising, rewriting or converting the freeform data into coded form
- Several tools are available to search free-text for direct and indirect identifiers and eliminate them without rendering the remaining text unreadable, but none can catch all instances of these identifiers all the time

12.8 Handling Small Cell Sizes

- Considering aggregate-level data (table structures) is beyond the scope of this document. However, the following general guidelines have been included for completeness:
  - Maintain a minimum number of individuals per cell, e.g., at least six individuals in a cell
  - Regardless of the number of individuals in a cell, if a small number contribute a large percentage of a cell's value then consider the data to be sensitive. As a rule of thumb, no individual can represent more than 70 percent of a cell's total value
  - Consider the size of the population size from which the data are drawn, e.g., if the data are from a small, rural population rather than a larger urban population the risk of re-ID may be higher
  - Display data using percentages rather than actual counts
  - Report either totals or averages (without counts) to display cost or account data
  - Display distributions in combined groups of 10 or 20 percent of the total
  - Combine sensitive data that are complimentary or in neighbouring cells
- Change the frequency of reporting, e.g., quarterly instead of monthly to achieve minimum cell size

13 Appendix D – Structured Methodology for Estimating Re-ID Risk Levels

The following five-step process model as proposed in the paper entitled *De-ID Risk Assessment Model* (16) can help quantify the risk threshold, i.e. an acceptable upper limit for the probability of re-identification. The model is not yet considered mainstream. It considers the Data Requestor's intentions for the requested data, capacity to re-ID, the mitigating controls at the Data Requestor's site and how much harm will accrue if there is an unintended or unauthorized use and subsequent disclosure.

13.1 Assess 'Intention and Capacity to Re-ID'

Assess the Data Requestor's intentions for the requested data, capacity to re-ID the data if it were given to them in de-identified form. This is based on information gained through the collaborative disclosure request review process and the collaborative process. Factors to consider include:

Intention

- Has the Data Requestor previously worked with the Data Provider?
- Can the Data Requestor potentially gain financially from re-identifying the data?
- Is there a non-financial reason for the Data Requestor to try to re-ID the data?

Capacity

- Has the Data Requestor the technical expertise to attempt to re-ID the data?
- Has the Data Requestor the financial resources to attempt to re-ID the data?
- Has the Data Requestor access to other databases that can be linked with the data to re-ID individuals?

Score the Data Requestor's 'intention and capacity' as low, medium or high based upon the information provided in the disclosure request and the responses to the above items.

13.2 Assess mitigating controls

Assess the mitigating controls in place at the Data Requestor's site. These controls can be assessed from the information describing the privacy and security practices in place at the Data Requestor's site, which was provided in the disclosure request. The better the privacy and security practices in place, the higher the mitigating controls. Score the mitigating controls as low, medium or high.

13.3 Estimate 'probability of a re-ID attempt'

Using the results from steps 1 and 2, estimate the 'probability of a re-ID attempt'. This assesses the likelihood or probability that someone will attempt to re-ID the data, defined as one of Remote, Occasional, Probable or Frequent. For example, if

the probability is Frequent then there is a very high chance that someone will attempt to re-identify the data.

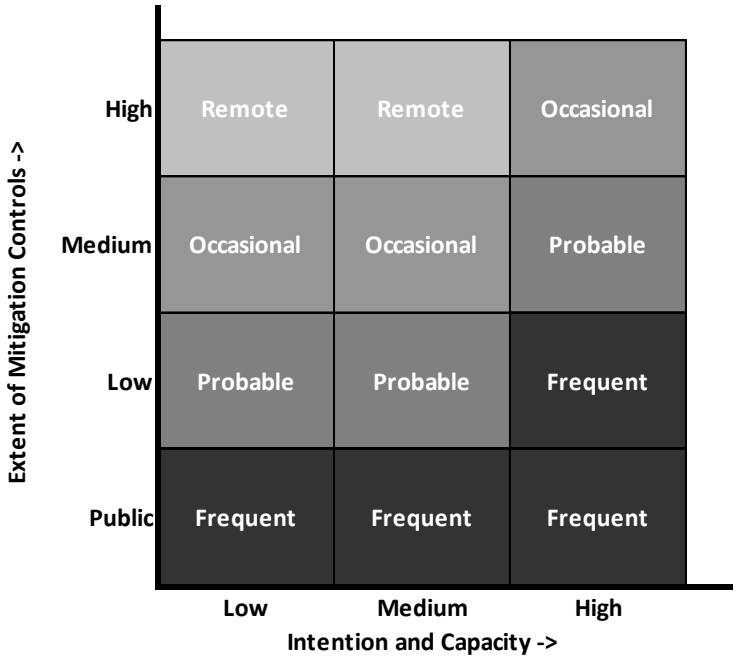

Figure 5 – Probability of a Re-ID Attempt

13.4 Evaluate potential for 'invasion-of-privacy'

By measuring the potential for 'invasion-of-privacy', the Data Provider can decide how much de-ID must be done. Assume that an 'invasion-of-privacy' can occur under three conditions:

- The Data Provider inappropriately discloses the data to the Data Requestor or there is an inappropriate use of the data
- The Data Requestor inappropriately processes the data
- There is an unintended or unauthorized use and disclosure at the Data Requestor's site

Factors to consider include:

- Are the data highly detailed or sensitive?
- Do the data come from a highly sensitive context?
- Will a considerable impact occur if there was an unintended or unauthorized use and subsequent disclosure?
- If there was an unintended or unauthorized use and subsequent disclosure, will it result in direct and quantifiable damages and injury to the individuals?
- If the Data Requestor is located in a different jurisdiction, is there is a possibility that the data sharing agreement will be difficult to enforce?

- Does the Data Requestor have little to lose if there is an unintended or unauthorized data use and disclosure?

Score the 'invasion-of-privacy' as low, medium, high depending on the responses to the above items.

13.5 Estimate how much de-identification is required

This matrix below combines the results from steps 3 and 4.

- The *x-axis* is the likelihood that someone will attempt to re-ID the data, i.e., Remote, Occasional, Probable or Frequent
- The *y-axis* is the potential for 'invasion-of-privacy', i.e., low, medium or high

The cell values in the matrix suggest a risk threshold. For example, a value of 5% means that the re-ID risk is high and extensive de-identification is required, i.e., the probability of re-identifying the data must be kept below 5%. It could be that the Data Requestor has poor privacy and security and practices, the resulting vulnerability is high and extensive de-identification is needed. Conversely if the suggested value is 33% then the overall risk is low and the data can be 'more lightly de-identified'. The risk threshold values in the matrix can be seen as suggestions only and may be modified by the Data Provider to reflect their perceptions of risk.

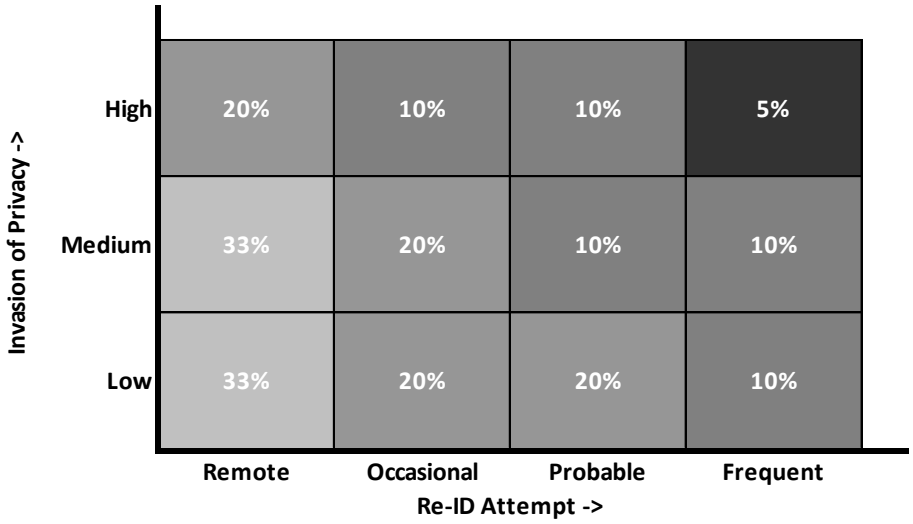

Figure 6 – Risk Threshold to Use

14 Appendix E – Alternatives to Traditional Disclosure

If a disclosure request is turned down, the Data Provider and Data Requestor can explore other disclosure alternatives including:

14.1 Controlled Access on Data Provider's Site

- The Data Requestor is granted controlled, secure access to the requested data on the Data Provider's premises.
- The Data Requestor has access to the data only while on the Data Provider's premises and leaves only with final analytical results
- This approach can work when the Data Requestor:
  - Has limited or no requirement to link to other databases
  - Has inadequate privacy and security practices
  - Has not previously worked or collaborated with the Data Provider
  - Has the internal technical expertise to attempt to re-ID the data
  - Has access to databases that can be linked with the data to re-ID individuals
  - Is located in another jurisdiction and there is a possibility that the data sharing agreement might be difficult to enforce
  - Could affect many people if there was an unintended or unauthorized use and subsequent disclosure
  - Could cause direct and quantifiable damages and injury to the individuals if there was an unintended or unauthorized use and subsequent disclosure
- The approach can also work when the acceptable level required to manage the risk is too low and extensive de-ID will result in data of limited use

14.2 Data Access from a Secure Satellite Facility

- The Data Provider:
  - Establishes a fully secure, satellite data-access facility at an academic research institution (university)
  - Screens and pre-qualifies fully-appointed, health researchers at the institution (university) as approved scientists
  - Grants each approved scientist with secure access from the satellite facility to de-identified data for research purposes over secure and encrypted data communication facilities
- The benefits can include:
  - The Data Provider can monitor and log all data access activity
  - The Data Requestor gets faster access to research data to a wide variety of tools
  - The process is simplified since scientists are pre-screened and pre-approved
  - The process fosters collaboration among approved scientists
  - The process facilitates an increased opportunity for knowledge transfer

15 Appendix F – Privacy Statutes, Regulations and Policies

15.1 Province of British Columbia

Freedom of Information and Protection of Privacy Act, 1996 (FOIPP)
http://www.bclaws.calRecon/document/freeside/--%20F%20-/Freedom%20of%20lnformation%20and%20Protection%20of%20Privacy%20Act%20%20RSBC%201996%20%20c.%20165/00_Act/96165_00.htm

E-Health (Personal Health lnformation Access and Protection of Privacy) Act [SBC 2008] c. 38 [NB: new act not all enabling regulations have been adopted yet]
http://www.bclaws.ca/Recon/document/freeside/-- c -/ehealth_personal health information access and%2Oprotection of privacy_act_sbc 2008_c. 38/00_08038_01.xml

172/2009: Disclosure Directive Regulation
http://www.bclaws.ca/Recon/doment/freeside/-- c --/ehealth_personal health information access and protection of privacy_act_sbc 2008_c. 3 8/05_regulations/l10_172_2009 .xml

15.2 Province of Alberta

Health Information Act, 1996 (HIA)
http://www.qp.alberta.ca/574.cfm?page=H05.cfm&leg_type=Acts&isbncln=9780779746682

Electronic Health Record Information Exchange Protocol (IEP)
http://www.albertanetcare.ca/11.htm

15.3 Province of Saskatchewan

The Health lnformation Protection Act, effective 2003 (HIPA)
http://www.qp.gov.sk.ca/documents/English/Statutes/Statutes/HO-021.pdf

15.4 Province of Manitoba

Personal Health Information Act, May 2010 (PHlA)
http://www.gov.mb.ca/health/phia/index.html

Reg. 64/2003
http://web2.gov.mb.ca/laws/regs/2003/064.pdf]

15.5 Province of Ontario

Personal Health Information Protection Act, 2004
http://www.e-laws.gov.on.ca/html/statutes/English/elaws_statutes_04p03_e.htm

Ontario Regulation General 329/04

15.6 Province de Québec

An Act respecting access to documents held by public bodies and the protection of personal information (Loi sur l'accès) R.S.Q. A2.1
http://www2.publicationsduquebec.gouv.qc.ca/dynamicSearch/telecharge.php?type=2&file=/A_2_1/A2_1_A.html

Formulaire de demande d'autorisation de recevoir des renseignements nominatifs à des fins de recherche, d'étude ou de statistique.
http://www.cai.gouv.qc.caJindex.html

Health and Social Services Act (LSSS) R.S.Q. S-4.2
http://www2.publicalionsduquebec.gouv.qc.ca/dynamicSearch/telecharge.php?type=2&file=/S_4_2/S4_2_A.html

Health Insurance Act (LAM) R.S.Q. A-29
http://www2.publicationsduquebec.gouv.qc.ca/dynamicSearch/telecharge.php?type=2&file=/A_29/A29_A.html

15.7 Province of Nova Scotia

Freedom of Information and Protection of Privacy Act, 1999,
http://www.gov.ns.ca/legislature/legc/statutes/freedom.htm

Regulations
http://www.gov.ns.ca/just/regulations/regs/foiregs.htm

Personal Health Information Act (Bill 64)
www.gov.ns.ca/legislature/legc/bills/61s_1st/1st_read/b064.htm

15.8 Province of Newfoundland and Labrador

Personal Health Information Act, SNL2008 ch. P-7.0 1
http://www.assembly.nl.ca/legislation/sr/statutes/p07-01.htm
NB: Only selected provisions are in force – other are yet to be proclaimed.

Health Research Ethics Authority Act, SNL2006 ch H-1.2
http://www.assembly.nl.ca/legislation/sr/statutes/h01-2.htm

NB: Not in force yet

15.9 Jurisdictions without Specific Health Privacy Legislation

The following jurisdictions do not have specific health privacy legislation but find direction from the statutes noted.

New Brunswick	PIPEDA – Private Sector Protection of Personal Information Act
Prince Edward Island	PIPEDA – Private Sector Freedom of Information and Protection of Privacy Act – Public Sector
Yukon	PIPEDA – Private Sector Access to Information and Protection of Privacy Act – Public Sector
Northwest Territories	PIPEDA – Private Sector Access to Information and Protection of Privacy Act – Public Sector
Nunavut	Access to Information and Protection of Privacy Act – Public Sector

16 Appendix G – Automated De-ID Tools

The following appendix provides an informational list of commercially available, automated de-ID tools that generally meet the criteria for automated de-ID. This overview is an updated version of the tools identified in the 2009 paper entitled *Tools for De-Identification of Personal Health Information (5)*. It does not constitute and endorsement of any one product in particular.

The list is divided into tools that mask[2] direct identifiers and tools that mitigate re-id risk from indirect identifiers, both at the record level. It is not intended to be an exhaustive list but rather provide a brief overview and link to the applicable product website.

The number in parentheses following the name of each tool indicates a source of further information as listed the Reference Documents.

16.1 Requirements for Automated De-ID

The paper entitled *A Globally Optimal k-Anonymity Method for the De-identification of Health Data (8)* identifies four requirements for a de-ID algorithm to ensure that it is practical for use. These represent the minimum set of requirements:

1. **Quasi-identifiers can be represented as hierarchies**: A de-ID algorithm must be able to deal with the hierarchical nature of variables. One common way to satisfy the k-anonymity criterion is to reduce the detail of quasi-identifiers, i.e., to reduce the precision of the variables as they move up the hierarchy. For example, a less precise representation of a six-character postal code 'K1H 8L1' is the first three characters 'K1H'. Likewise, a birth date can be represented less precisely as the year of birth. Numeric variables can be represented hierarchically, e.g., discrete ages can be converted to intervals such as [0–9], [10–20] etc.
2. **Discrete intervals are user-definable:** Since the reduction in detail of quasi-identifiers requires the user to make judgment calls, a de-ID algorithm must allow users to define the interval sizes that are appropriate for their analysis. If a de-ID algorithm automatically defines intervals then it may produce categories that are not meaningful or useful for analyses. For example, in an attempt to create equal numbers of record in each age category, an automated program may partition age

[2] The Infoway *White Paper on Information Governance of the Interoperable Electronic Health Record* uses the term 'masking' synonymously with the term 'locking,' where 'locking is the 'ability of a patient to expressly withhold or withdraw consent to the disclosure of a portion of his or her personal health information for healthcare purposes, except during a medical emergency.' The white paper acknowledges that, 'the term 'masking' has also been used occasionally as a synonym for anonymization (a process which is sometimes engineered to be irreversible) or as an informal way of referring to the process of encryption'. The automated de-ID tools discussed here use the term 'masking' in this latter context.

into intervals such as [0–9], [10–12], [13–25], (26–60], etc. If a user cannot define and control interval sizes then it could make data analysis overly complex and reduce data quality or usefulness.

3. **De-ID techniques need to be applied globally rather than locally:** De-ID techniques need to be applied consistently to all quasi-identifiers across all records in the dataset. For example, one record has a 17-year-old's age reduced to an interval of [11–19] while another record has a 17-year-old's age reduced to an interval of [16–22]. This inconsistency can make the data very difficult or even impossible to analyze.
4. **The solution satisfies k-anonymity and minimizes information loss:** While it is difficult to calculate an optimal balance, a de-ID algorithm must be able to achieve a balance between satisfying the k-anonymity criterion and minimizing information loss. Some programs do a better job than others of balancing k-anonymity and information loss.

16.2 Mask Direct Identifiers at the Record Level

The tools listed below manipulate direct identifiers in record-level data. They generally employ data substitution including randomization and data swapping. They can be summarized as follows:

Masking Technique	Oracle	Camouflage	Informatica	Data Masker	Optim
Suppression			√	√	
Random 'Noise'		√	√	√	
Substitution	√	√	√	√	√

Oracle Data Masking Pack (30)

Oracle provides a tool called the Oracle Data Masking Pack (ODMP) that works with the Oracle 11g database. The software reduces re-ID risk by irreversibly substituting original data with fictitious data.

ODMP provides a centralized library of 'out-of-the-box' masking formats for common types of indirect identifiers such as credit card and phone numbers. Users can extend this library with their own masking formats to meet their specific data de-ID requirements.

ODMP supports a variety of masking technique including:

- **Conditional masking:** Applying different masking rules depending upon certain conditions, e.g., apply one set of rules if the birth date indicates a child or adolescent and another set of rules if the birth date indicates an adult
- **Compound masking:** Ensuring that a set of related variables are masked as a group to ensure consistency, e.g. city, province/territory and postal codes values need to be consistent after masking

- **Deterministic masking:** Ensuring referential integrity, i.e., if a data value is substituted for another in one record/database, the substitution is consistently applied across other records/databases

ODMP can also support masking of data in other databases, such as IBM DB2 and Microsoft SQL Server. Further information is available at:

http://www.oracle.com/technology/products/oem/pdf/ds_datamasking.pdf

Camouflage (31)

Camouflage is a data-masking tool developed by a Canadian company, Camouflage Software Inc based in St. John's, Newfoundland and Labrador. Camouflage is a standalone tool available for desktops (Windows, UNIX, and Linux) and also in configurations that run on servers (Windows and Linux). It supports a variety of database platforms including Oracle, IBM DB2, Microsoft SQL Server, Sybase, and MySQL:

Camouflage provides the following features:

- **Random addition of 'noise':** Modifies numeric variables by incrementing or decrementing value, or increasing or decreasing by a percentage
- **Data substitution:** Including both data swapping and randomization of data values (generated or selected from a pre-defined set)
- **Maintenance of referential integrity across records/databases:** If a data variable is substituted for another, the substitution is consistently applied in other records/databases to ensure they link together properly

Camouflage has partnered with IBM, Microsoft and Oracle to market the tools. Further information is available at:

www.datamasking.com

Informatica Data Privacy (32)

Informatica Data Privacy (formerly Applimation Informia Secure) is a toolkit that works with a wide variety of database platforms (Oracle, DB2, SQL Server, Sybase, and Teradata) and runs on a variety of platforms (Windows, UNIX/Linux, and z/OS). The tool has the following data protection features:

- **Data suppression:** Replaces data variables with null values
- **Random addition of 'noise':** Includes some data skewing
- **Data substitution:** Including support of both randomization of data values (generated or selected from a pre-defined set) and data swapping
- **Maintenance of referential integrity across records/databases:** If a data variable is substituted for another, the substitution is consistently applied in other records/databases to ensure they link together properly
- **Mask data across different database platforms:** Can be used for composite data warehouses running in Oracle and Microsoft SQL) and

- **Extensive auditing features:** Consisting of audit logs and reports for all masking activities

The company web site contains several white papers and technical reports. Further information is available at:

http://www.informatica.com/products_services/data_privacy/Pages/data-privacy-features.aspx

Data Masker (33)

Data Masker was developed by a UK firm called Net 2000 and is used by many companies in the UK, US and Canada.

Data Masker runs only on Windows platforms is available for various versions of Oracle, SQL Server and DB2 UDB. A Sybase version is under development as of June 2010. The product has the following data protection features:

- **Data suppression:** Replaces data variables with null values
- **Random addition of 'noise':** Modifies numeric variables number by a random percentage of its real value
- **Data substitution:** Includes support of both randomization of data values (from a user-defined substitution set) and data swapping
- **Data encryption**: Leaving the data in place and visible to those with the appropriate key thus allowing for reversible de-ID

Data Masker is optimized for large databases. A fully functional copy of Data Masker can be obtained from the company web site for evaluation purposes without charge. The program is significantly less expensive than some of the other tools in this section. Further information is available at:

http://www.datamasker.com/index.html

IBM Optim Data Privacy Solution (34)

In 2007, IBM acquired a software company called Princeton Softech that developed enterprise data management software. IBM has rebranded the product as Optim and sells it as a suite of products and services for managing privacy. Product features include:

- **Data substitution:** Including various manipulations of substrings, arithmetic expressions, as well as random or sequential number generation, date aging and concatenations; and
- **Pre-defined data transformations:** For common identifiers such as the Canadian social insurance number

IBM's Infosphere data warehousing and archiving product can analyze databases and look for embedded indirect identifiers, e.g., transaction numbers that are not meaningless unique numbers). Further information is available at:

http://www-01.ibm.com/software/data/data-management/optim-solutions/data-privacy.html

16.3 Mitigate Re-ID Risk from Indirect Identifiers at the Record Level

The tools below are designed to address the risks of residual re-ID resulting from the presence of quasi-identifiers in record-level data from which the direct identifiers have already been removed.

PARAT – Privacy Analytics Risk Assessment Tool (35, 36)

Privacy Analytics has commercialized the technology developed by the Electronic Health Information Laboratory (EHIL) at the Children's Hospital of Eastern Ontario Research Institute and the University of Ottawa.

The objective of the PARAT software is to find a range of data values that minimize information loss while still guaranteeing k-anonymity. PARAT is a Windows based application and is compatible with several databases, including Oracle, and Microsoft SQL Server. PARAT uses a four-step process:

- The User selects the quasi-identifiers to be released from the data set
- The User specifies the acceptable re-ID risk threshold
- PARAT performs a risk analysis on the indirect identifiers based upon the presumed risk of re-ID from three hypothetical sources of attack: a prosecutor, a journalist, and a marketer
- PARAT applies several de-ID techniques to reduce re-ID risk to an acceptable level

PARAT uses several de-ID techniques including suppression and reduction in detail. PARAT is straightforward to use although its algorithms are sophisticated. Further information is available at:

http://www.privacyanalytics.ca/technology/technology.html

µ-Argus – Anti-Re-ID General Utility System (37)

µ-Argus is made available by Statistics Netherlands, that country's national statistics bureau. The program runs under Windows and was developed by the Computational Aspects of Statistical Confidentiality project of the European Union.

The µ-Argus software employs a five-step process

- The User determines the quasi-identifiers that could potentially re-identify data subjects
- µ-Argus first estimates the individual risk of re-ID for each record in the dataset, i.e., an upper bound for the probability of re-ID
- µ-Argus also estimates the global risk of re-ID for the entire file in terms of expected number of re-IDs and the re-ID rate

- The User then determines an acceptable level of risk
- After the risk has been estimated, μ-Argus applies several de-ID techniques including suppression and reduction to reduce re-ID risk to an acceptable level

μ-Argus allows the User to experiment with different levels of acceptable risk to examine the effect each has on the resulting dataset. Further information is available at:

http://neon.vb.cbs.nl/casc/Software/MuManual4.2.pdf

16.4 Other Automated Tools

τ-ARGUS – Anti-Re-ID General Utility System (37)

τ-ARGUS is an extension of μ-ARGUS. It is intended for aggregate-level (tabular and frequency) data.

τ-ARGUS applies similar statistical techniques to those incorporated into μ-Argus to minimize the risk of re-ID of individuals in the aggregate-level data. These include changing classification schemes, cell suppression and random addition of 'noise' into either the underlying or summary data. Further information is available at:

http://neon.vb.cbs.nl/casc/Software/tauInno3_3_B2.zip

Canadian Postal Code Conversion (38)

The Postal Code Conversion File (PCCF) was first created in 1998 by the Geography division of Statistics Canada and has been regularly updated ever since (the most recent update relies upon the 2006 census).

The PCCF allows Canada Post Corporation six-character postal codes to be mapped to Statistics Canada's standard geographic areas for which census data and other statistics are produced. Through the link between postal codes and standard geographic areas, the PCCF permits the integration of data from various sources.

By converting from postal codes to standard geographic areas from the Census, Data Providers can determine the population size of each area and ensure that the geographic data does not contain too few individuals. It may also provide a finer-grained geographic breakdown than the first three characters of the postal code without increasing risk of re-ID. Further information is available at:

http://www.statcan.gc.ca/bsolc/olc-cel/olc-cel?catno=92F0153G&CHROPG=1&lang=eng

17 Appendix H – Glossary of Terms

Aggregate-level data	Data that have been collected at the record-level then tabulated and reported as a sum or frequency to ensure that there are no directly identifying variables or quasi-identifiers
Cell Suppression	A special case of suppression where entire variables and entire records are not suppressed but rather specific quasi-identifier values are suppressed such that the amount of suppression is minimal but still maintains an acceptable level of re-ID risk
Clinical program management	The use of data to improve front-line health care programs and services, e.g., reduce hospital-acquired infections, improve the delivery of surgical programs, improve programs for chronic diseases like diabetes, understand why discharged patients need to be re-admitted, understand how many patients within a physician's practice have diabetes and their related complications to develop targeted programs of care
Data linkage	The connecting of two or more data records of health information or de-identified data to form a composite record for a specific individual
Data provider (also called 'data custodian')	An organization that collects and discloses health information including ministries of health, regional health authorities and similar bodies, hospitals, other health care facilities and professional colleges
De-ID processes	Processes that manipulate health information so that the identity of the individual cannot be determined by a reasonably foreseeable methods. They can include: • Reduction in Detail • Suppression • Random Addition of 'Noise' • Substitution • Pseudonymization
De-identified data	Health information that has been manipulated using appropriate de-ID processes. The directly identifying variables have been adequately manipulated and quasi-identifiers adequately disguised to ensure that the re-ID risk is acceptable.
Directly identifying variables	Data variables that can be used to uniquely identify an individual either by themselves or in combination with other readily available information. Examples include name, phone number, email address, health insurance card number, credit card number and social insurance number. See also indirectly identifying variables.

Disclose	To release or make available health information or de-identified data other than to the original Data Provider or the individual to whom the data pertain
Health information	A broad term including but not limited to financial information about health and health care, health information, de-identified data and aggregate data
Health system use (also called secondary use)	The use of health information for clinical program management, health system management, monitoring the health of the public, and research, all of which lead to improved patient care and health outcomes. This includes clinical program management, health system management, monitoring public health and research (q.v.)
Identifiable data	Data that have directly identifying variables or sufficient quasi-identifiers that can be used to identify the individual
Indirectly identifying variables (also called quasi-identifiers)	Data variables that can be used to probabilistically identify an individual either by themselves or in combination with other available information. Examples include sex, date of birth or age, geo-codes, first language, ethnic origin, aboriginal identity, total years of schooling, marital status, criminal history, total income, visible minority status, profession, health event dates, health-related codes, country of birth, birth weight, and birth plurality. See also directly identifying variables
K-anonymity	A criterion to ensure that there are at least k records in a dataset that have the same quasi-identifier values. For example, if the quasi-identifiers are age and gender, then it will ensure that there are at least k records with 45-year old females
Potentially De-identified data	Data where manipulations have been performed on the identifying variables but attempts to disguise the quasi-identifiers may be insufficient. The data may not be fully de-identified, may be partially exposed and may represent a re-ID risk
Monitoring public health	The use of data to understand the health of the public, e.g., identify a potential outbreak such as H1N1, understand rates of cancer and how they differ across age groups and regions of the country, monitor the coverage of newborn vaccine programs
Pseudonymization (also called 'coding' or 'alias assignment')	A technique that replaces identifiers with unique pseudonyms. A random pseudonym will be different if it is generated multiple times for the same individual. A unique pseudonym will be the same if it is generated multiple times for the same individual … perhaps to provide linkages across databases

Random 'noise' (also called data perturbation or scrambling)	A technique that adds random 'noise' to the values of a variable in order to disguise its true value
Record Suppression	A technique that involves removing and withholding data records that create high re-ID risk, though this may also reduce the usefulness of the data
Record-level data	Data in which each record is related to a single individual. Contrast to aggregate level data
Reduction in Detail	A technique that reduces the data detail by rounding or collapsing its into larger categories. The objective is to reduce the number of data records with a unique combination of potentially re-identifying variables, i.e., increase k-anonymity. For example, round an individual's age into a pre-defined range of ages, though this may also reduce the usefulness of the data
Research	The use of data to identify improvements to medical treatments and programs of care, and to better understand the health of the population, the factors influencing health, and the performance of the health care system, e.g., understand the impact of medical treatments (e.g., chemotherapy) on illnesses (e.g., breast cancer) and the link to future health problems, track the progression of patients with chronic diseases to determine the effectiveness of different programs of care, model and forecast health trends and utilization of health services, understand how factors like lifestyle and behaviour impact the overall health of the population
Reversible Pseudonymization (also called 'coding' or 'alias assignment')	With this technique, coding is reversible since it allows individuals to be re-identified if necessary. Reversible coding schemes can often involve single or double coding. • *Single coding* means that identifiers are removed and each record is assigned a new code (a pseudonym). Identifiers are kept in a different identity database linked via the pseudonym to the original data • *Double coding* means that the pseudonyms in the original data and in the identity database are different. The link between them is kept in a separate linking database maintained in a secure location by a trusted third party
Sensitive variables	Variables not really useful for re-ID purposes but containing sensitive health information about the individual. Examples include sexual orientation, diagnosis codes, history of depression

Substitution	A technique that replaces the actual values in a dataset with values, which look and behave like the original values. For example, replacing the real names and addresses with false (but real) names and addresses
Variable Suppression	A technique that involves the removal or withholding of a data variable's values. This is often a necessary step but may reduce the usefulness of the data

18 Appendix I – Reference Documents

Reference numbers refer to the numbers embedded in file names in the reference document repository:

1 Khaled El Emam, and Anita Fineberg, 'An Overview of Techniques for De-identifying Personal Health Information', 14 Aug 2009

2 Khaled El Emam, 'The five levels of identifiability', 31 Dec 2009

3 Fida Kamal Dankar and Khaled El Emam, 'A Method for Evaluating Marketer Re-ID Risk', Pais'10, Lausanne, Switzerland, 22 March 2010

4 Ross Fraser and Don Willison, 'Tools for De-ID of Personal Health Information', Presentation prepared for the Pan Canadian HIP Group, 21 Sep 2009

5 Ross Fraser and Don Willison, 'Tools for De-ID of Personal Health Information', Draft report prepared for the Pan Canadian HIP Group, 21 Sep 2009

6 Don Willison, 'Use of Data from the Electronic Health Record for Health Research – current governance challenges and potential approaches', Mar 2009

7 Price Waterhouse Coopers, 'Transforming healthcare through health system use of health data', 2009

8 Khaled El Emam, PHD, Fida Kamal Dankar, PHD et al, 'A Globally Optimal k-Anonymity Method for the De-Identification of Health Data', Journal of the American Medical Informatics Association Vol. 16 No. 5 September–October 2009

9 ISO/TS 25237, 'Health Informatics: Pseudonymization,' 1 Dec 2008

10 Canadian Institute for Health Information (CIHI), Privacy Policy on the Collection, Use, Disclosure and Retention of Personal Health Information and De-Identified Data (Bilingual), Jun 2009

11 Canadian Institute for Health Information, Privacy and Security Framework, Jan 2010

12 Personal Health Information Protection Act, 2004, Ontario Regulation 329/04 General

13 Personal Health Information Protection Act, 2004, S.O. 2004, Chapter 3, Schedule A

14 Khaled El Emam, 'Heuristics for De-identifying Health Data', Aug 2008

15 Khaled El Emam and Fida Kamal Dankar, 'Protecting Privacy Using k-Anonymity', Oct 2008

16 Khaled El Emam, 'De-ID Risk Assessment Model', 2009

17 Pan-Canadian HIP Group, Privacy-protective trans-jurisdictional disclosures of information from the interoperable electronic health record, Some pan-Canadian common understandings for discussion, Version 8, 25 May 2010

18 Lorie Wijntjes, Jessica Pollner, Complying with HIPAA's Privacy Rule - Tabular Data, Info Management Direct, July 2007

19 The Government of Manitoba, 'An Agreement Respecting a Research Project Data Set Comprised of Information Provided by Manitoba Health', 13 Mar 2009

20 Manitoba Health, 'Request for Access to Health Information Held by the Government of Manitoba', Version 3, 15 Jan 2008

21 Institute for Clinical Evaluative Sciences, 'Risk Register Scoring Template', 6 Nov 2009

22 Institute for Clinical Evaluative Sciences, Confidentiality Agreement for Non-Closure, 2010

23 Institute for Clinical Evaluative Sciences, 'Project-Specific Privacy Impact Assessment Form', Sep 2008

24 Institute for Clinical Evaluative Sciences, Data Sharing Agreement, Dec 2009

25 Canadian Institute of Health Research, National Sciences and Engineering Research Council of Canada, Social Science and Humanities Research Council of Canada; 'Ethical Conduct for Research Involving Humans', 2005

26 Canada Health Infoway, Interoperable Electronic Health Record Solutions (EHRS) Blueprint Executive Overview, Version2, April 2006

27 Newfoundland and Labrador Centre for Health Information, Application to Request Record-Level or Identifying Information – General, Version 1, 22 Jan 2010

28 Newfoundland and Labrador Centre for Health Information, Application to Request Record-Level or Identifying Information – Research, Version 1, 22 Jan 2010

29 Khaled El Emam, Fida K Dankar, Régis Vaillancourt, Tyson Roffey, and Mary Lysyk, 'Evaluating the Risk of Re-ID of Patients from Hospital Prescription Records', Canadian Journal of Hospital Pharmacists, Vol. 62, No. 4, July–August 2009

30 Oracle Corporation, 'Oracle Enterprise Manager 10g Data Masking Pack', 2007

31 Camouflage Software, 'Data Masking Best Practices - Four Phases of Evaluating and Implementing a Data Masking Solution', White Paper, March 2010

32 Informatica, 'Data Privacy Best Practices for Data Protection in Nonproduction Environments', White Paper, June 2009

33 Net 2000 Ltd (Data Masker), 'Data Sanitization Techniques', White Paper, 2005

34 IBM Corporation, 'Closing the data privacy gap: Protecting sensitive data in non-production environments', 2009

35 Privacy Analytics, 'PARAT – The Tool for Anonymizing Health Data'

36 Privacy Analytics, 'De-Identification Reduce Privacy Risks When Sharing Personally Identifiable Information', 2009

37 Statistics Netherlands, 'μ-ARGUS User's Manual', version 4.2, December 2008

38 Statistics Netherlands, 'τ-ARGUS User's Manual', version 3.3, December 2008

39 Statistics Canada, Postal Code Conversion File Reference Guide, December 2006