

Tools for De-Identification of Personal Health Information

Prepared for the Pan Canadian Health Information
Privacy (HIP) Group

Authored by: Ross Fraser and Don Willison, September 2009

Executive Summary

This report identifies useful and available tools and techniques for the de-identification of personal information from interoperable electronic health records and health information-related data warehouses.

Section 1 contains a general introduction and defines some of the terms commonly used when discussing de-identification.

Section 2 describes the principles of de-identification, including two basic models of how data warehouse are operated; the distinction between record-level and aggregate data; the distinction between direct and indirect identifiers; a description of the types of secondary uses that data warehouses support (health research, health system planning, public health surveillance, and generation of de-identified data for system testing); an explanation of k-anonymity as a measure of de-identification; a discussion of the special problems inherent in free-text data; a discussion of the special problems posed by genetic information; and guidance to prevent unintended disclosures through lapses in security.

Section 3 describes the various approaches to de-identification, including a flow diagram of when to use each approach. Record-level data can be de-identified through data reduction (including removal of direct identifiers, reduction in the detail of the data, and sampling), data modification (random addition of noise to the data, randomization of data values, and data swapping), and data suppression. Each approach is briefly described and examples are given. Pseudonymisation is described; including the distinction between reversible and irreversible pseudonymisation, and the two basic ways in which pseudonymisation is carried out.

Aggregate data has its own approaches to de-identification, including restriction-based methods (cell suppression and changing the classification scheme for the data) and heuristics. These are described with examples.

Section 5 contains some best practices for de-identification. These include how direct identifiers should be handled in data warehouses; how date variables should be presented in released datasets; how location data such as postal codes should be handled in released datasets; special guidelines for diagnostic imaging data; how pseudonymous IDs should be handled in released datasets to prevent unintended data linkages to other datasets; and elements of contractual agreements on use and disclosure of datasets.

Section 5 contains a description of readily available tools for de-identification of both record-level data and aggregate data. Tools described for handling direct

identifiers in record-level data include Oracle Data Masking Pack, Camouflage, Informatica Data Privacy, and Data Masker. Tools for handling indirect identifiers in record-level data include PARAT, μ -Argus, and the Cornell Anonymization Toolkit. Additional tools for aggregate data include τ -ARGUS. Additional tools are also discussed for postal code conversion. Third-party evaluation of de-identification (or the lack thereof) is also briefly discussed.

Section 6 briefly discusses re-identification risks. The report concludes with two observations: that the tools described can significantly reduce the risk of re-identification but only when sensibly combined with administrative controls such as end-user agreements and good security practices; and that the de-identified data will only be of value to the end-users if the approach to de-identification supports the intended use.

Contents

Contents	iv
1 Introduction	1
1.1 Purpose	1
1.2 Scope	1
1.3 Terms and Abbreviations	2
2 Principles of De-Identification	3
2.1 Purpose-Built Data Warehouses	3
2.2 Record-Level Data vs. Aggregate Data	4
2.3 Direct vs. Indirect Identifiers	5
2.4 Types of Secondary Uses	6
2.5 K-Anonymity	7
2.6 Free-text Data	8
2.7 Genetic information	8
2.8 Intended vs. Unintended Disclosures	9
3 General Approaches to De-Identification	10
3.1 Disclosure of Record-Level Data	12
3.1.1 Data Reduction	13
3.1.2 Data Modification	14
3.1.3 Data Suppression	15
3.1.4 Pseudonymisation	15
3.2 Disclosure of Aggregate Data	16
3.2.1 Restriction-Based Methods	16
3.3 Heuristics	18
4 Best Practices for De-Identification	19

4.1	Storage of Direct Identifiers in Data Warehouses	19
4.2	Date Variables in Datasets	19
4.3	Location Data in Datasets	20
4.4	Diagnostic Imaging Data in Datasets	21
4.5	Pseudonymous IDs and Data Linking	21
4.6	Contractual Agreements on Use and Disclosure of Datasets	21
5	Automated De-Identification Tools	21
5.1	Requirements for Automated De-Identification Tools	22
5.2	Tools to Remove or Suppress Direct Identifiers in Record-Level Data	23
5.3	Tools to Mitigate Risk of Re-identification from Indirect Identifiers in Record-Level Data	26
5.4	Tools to Mitigate Risk of Re-identification from Aggregate Data	30
5.5	Miscellaneous Tools	30
5.6	Evaluation of De-Identification Tools	31
6	Residual Risk of Re-Identification	32
7	Conclusion	32
	References	33
	Acknowledgements	35

1 Introduction

1.1 Purpose

The purpose of this paper is:

1. to identify and summarize from the published literature useful and available tools and technologies for the de-identification of personal information from interoperable electronic health records and health information-related data warehouses. For completeness, this will include a brief description of how procedural approaches to disclosure control complement the technologies to limit disclosure.
2. to identify and summarize from the published literature assessments that have been done to date on these tools or technologies; and
3. to identify whether one or more organizations are well placed to assess such tools.

1.2 Scope

This report is undertaken in support of the work of the federal/provincial/territorial Health Information Privacy Group in developing common understandings and suggestions for pan-Canadian approaches to the de-identification of personal health information.

The following are within the scope of this report:

- a) a scan of the literature to identify and summarize any assessments that have been done on the tools or technologies
- b) identification of an organization or organizations well-placed to assess such tools.

As discussed later in this document, tools for de-identification require certain policy decisions be made to guide their effective use. Such policies are ultimately a matter for jurisdictions to decide based on their tolerance of re-identification risk and on the intended use of de-identified data. Such policy decisions are out-of-scope of this paper, as are the administrative procedures that implement them. However, best practices are discussed in section 4.

It is important to note that the tools reviewed in this paper are representative of those available at the time of writing. This list of tools is not exhaustive. Nor is the inclusion of tools in this paper an endorsement from the authors. Implementers

should carefully determine what their requirements are; select candidate tools that purport to meet these requirements; and then evaluate these tools carefully.

1.3 Terms and Abbreviations

A challenge in this field is the lack of standardized terminology. Below, we provide definitions for the terms we use in this report. Where available, we draw our definitions from the CIHR’s “Best Practices for Protecting Privacy in Health Research”. [2] Other sources are referenced at the end of the definition.

Term	Definition
Aggregate data	Data that have been averaged or grouped into ranges (e.g. 5-year or 10-year age groupings) across multiple records. Also referred to as <i>macro-data</i> . [2]
CIHR	Canadian Institutes of Health Research
Direct identifiers	Variables that provide an explicit link to a data subject. These include name, address, telephone number, and health insurance number. See also <i>indirect identifiers</i> . [2]
Generalization	A method of obtaining k-anonymity that involves replacing or re-coding a value with a less specific but semantically consistent value. [3]
Indirect identifiers	Variables that, in combination, could be used to identify an individual. Examples include date of birth, gender, postal code, national or ethnic origin, less common medical condition, unusual education or occupation. [2] See also <i>direct identifiers</i> and <i>quasi-identifier</i> .
Informativity	Refers to the robustness of a variable in terms of its ability to provide information for analysis. For example, date of birth is a highly informative variable, as it can be manipulated in many different ways in analyses – e.g. into age at the time of a particular hospital admission date. If date-of-birth were to be substituted with year of birth, that variable is somewhat less informative, and age category (e.g. age 50-60) is less informative again.
K-anonymity	A release of data is said to adhere to k-anonymity if each released record for a data subject contained in the data set cannot be distinguished from at least k-1 other individuals whose records also appears in the data set.

Term	Definition
	<i>Adapted from [3].</i>
Macro-data	See <i>aggregate data</i> .
Micro-data	See <i>record-level data</i> .
PCCF	Postal Code Conversion File (from Statistics Canada; see section 5.5)
Pseudonymous identifiers	Personal identifier that is different from the normally used personal identifier. This may be either derived from the normally used personal identifier in a reversible or irreversible way, or alternatively be totally unrelated. Usually restricted to mean an identifier that does not allow the derivation of the normal personal identifier.
Quasi-identifier	A set of <i>indirect identifiers</i> that in combination can uniquely identify individuals but that cannot do so in isolation. E.g., the combination of birth date and gender is sometimes considered a quasi-identifier.
Record-level data	Data at the level of an individual person. Record-level data need not directly identify the data subject but are more vulnerable to re-identification than are aggregate data. Also referred to as micro-data.
SDC	See <i>statistical disclosure controls</i> .
Statistical Disclosure Controls (SDC)	A collection of statistical approaches to the limiting of the risk of disclosure of individuals' identities.
Suppression	A method of obtaining k-anonymity by not releasing a value (e.g. in a table, suppressing several cells) or an entire record (in releases of record-level data. [3])

2 Principles of De-Identification

2.1 Purpose-Built Data Warehouses

Databases that support the interoperable electronic health record (iEHR) are structured for the primary purpose of effective and efficient delivery of health care to individuals. The data may reside in varying formats in separate repositories running on different technological platforms.

Before such data can be used for the purposes of health systems planning, public and population health, or clinical research, the data must be collated, cleaned and organized in a fashion that permits a population-level analysis. This requires a different data structure and a particular expertise which is typically not found in most institutions that are custodians of personal health information. Therefore, a general decision needs to be made – perhaps at the jurisdictional level – as to whether this ongoing task will be done by a trusted third-party or by the provincial or regional health authority.

2.2 Record-Level Data vs. Aggregate Data

The next decision that needs to be made is whether record-level data¹ (also referred to as micro-data) should be released at all by either the health authority or trusted third-party or whether infrastructures should be developed for in-house processing and analysis of such data, with only aggregate data being released (e.g. tabular summations or frequency data; also referred to as macro-data). The disclosure of data to outside researchers allows those researchers extended off-site access to de-identified data sets and, unsurprisingly, is the model preferred by most researchers involved in clinical research and epidemiology. This approach is employed (at least to some extent) for health data research in British Columbia, Quebec and Newfoundland and Labrador. The Canadian Institute for Health Information also releases record-level data to researchers under some limited and strictly controlled circumstances.² However, disclosure of data to outside researchers results in reduced control over uses of the data, including manipulation of the data – whether intentional or unintentional – in ways that may increase the risk of re-identification of individuals. In-house analysis is more privacy-protective but requires extensive in-house analytic expertise. The latter approach is employed by in the provinces of Ontario and Manitoba, where they have employed trusted third-parties—the Institute for Clinical Evaluative Sciences (ICES) and the Manitoba Centre for Health Policy (MCHP), respectively—to conduct these analyses. It is also the method employed for by Statistics Canada.

Record-level data are more prone to re-identification than are aggregate data (see the discussion in section 3). Therefore the decision over whether or not to release record-level data to outside researchers is a particularly important one. Ultimately, an informed policy decision must be made which is beyond the scope of this paper. The two approaches are discussed further in the paper “Data Data Everywhere” [4]

¹ Record-level data consists of a data set where each record relates to a single data subject. There may be one record per data subject or many, depending upon the structure and level of detail in the data set.

² All of the organisations mentioned above use legal and contractual means to restrict the uses that researchers can make of the data provided, as well as protections that must be in place to safeguard the datasets while in the hands of the researchers.

2.3 Direct vs. Indirect Identifiers

For health systems planning and population health research, the identity of individuals is immaterial to the analysis in the vast majority of cases. Consequently, direct personal identifiers (e.g. names, addresses, telephone numbers) should typically be removed before release of data for these purposes. An exception is in the area of public health where individual identifiable data may be necessary in the context of either contact tracing or disease outbreak management as well as for population level analyses. Even here, direct identifiers can be replaced with pseudonymous identifiers that allow re-identification of data subjects in a straightforward and controlled manner, should the need arise. (See the discussion in section 3.1.4).

When releasing data from which direct identifiers have been removed, there is always the potential for re-identification of individual-level records through a combination of variables that individually do not identify the data subject, but which, in combination, create a high risk of re-identification.. The most common such variables³ are:

- dates (e.g. birth, admission and discharge, procedures),
- geolocators (e.g. postal code, spatial data released on maps),
- gender, and
- diagnostic codes, especially when these refer to less common (though not necessarily rare) health conditions
- unusual education (e.g. PhD in statistical disclosure control procedures)
- unusual occupation (e.g. president of a major teaching hospital in Toronto).

Some of these variables share the property that their values are relatively uniformly distributed in a population (e.g., dates and gender). They combine to identify data subjects because each value reduces the number of individuals who could potentially be the data subject of a record. Conversely, diagnostic codes are distributed in a highly non-uniform manner. There is a handful of highly prevalent health conditions (e.g. hypertension, obesity) for which there is little concern over risk of indirect re-identification. By contrast, the prevalence of most health conditions is sufficiently low that, when combined with other quasi-identifying variables, the risk of re-identification increases dramatically. Postal codes are another variable that is not distributed uniformly, as urban areas may contain many more individuals in the area represented by the first four or five characters of a postal code than rural areas.⁴

³ Other variables such as race, ethnicity, religion, and income or other socio-economic indicators can also combine to reveal identity, though they are not as commonly encountered in Canadian EHR data as the variables in the list above.

⁴ The first three letters of the Canadian postal code are often used in healthcare data as a geolocator. But postal codes were designed for the efficient delivery of mail, not the protection of privacy. The first three characters of the postal code can refer to large urban areas or to small

To address this concern when disclosing data to other parties for purposes of analyses, no more information should be released than that necessary to accomplish the intended purpose of the release. This includes limiting the number of variables released and the level of detail in the variables to be released (e.g. date-of-birth vs. age vs. age category). The challenge is that methods employed to limit the level of detail in these variables also limits the capacity for conducting data analyses, so compromises must be made between achieving an “acceptable” level of risk of re-identification and having data that will have sufficient detail to accomplish the analytic task. This involves several judgment calls and an informed dialogue between the data custodian and the end user of the data. [ref to location where this will be discussed below]

2.4 Types of Secondary Uses

De-identification is typically applied for one of the following secondary uses:

1. **health research:** Most commonly, this includes the study of the impact of health services delivery and health policies on the health of individuals, and also the study of non-health-care system factors (e.g. education, income) on health. This nearly always requires record-level data for analysis purposes. Occasionally, it may also require follow-up with patients (if only to obtain consent) and, therefore, may require the capability to re-identify data subjects. See the discussion in section 3.1.4 on reversible pseudonymisation. Perhaps more importantly, health research may also require the ability to link records longitudinally over time. Here too, a consistently applied pseudonymous identifier (it need not be reversible) allows for the tracking of patients over an extended period of time.
2. **health system planning:** there is rarely any need for health system planners and analysts to access data containing direct identifiers. Where record-level data are needed (e.g., in planning involving points of care), consistently applied pseudonymous identifiers suffice for virtually all uses involved in health system planning (see the discussion of pseudonymous identifiers in section 3.1.4).
3. **public health surveillance:** like health research uses, public health surveillance requires record-level data. In some cases, it may be necessary to contact data subjects (e.g. for management of disease outbreaks) and hence requires the ability to re-identify data subjects where necessary. The use of pseudonymous IDs may still allow the removal of direct identifiers; limiting re-identification to a small subset of the data records on an “as needed” basis.

communities; populations referred to can vary from a few hundred to more than 40,000. A Statistics Canada report on mapping between postal codes and census areas can be found at <http://www.statcan.gc.ca/pub/92f0138m/92f0138m2007001-eng.htm>

- 4. generation of de-identified data for system testing:** while not typically considered a secondary use, the scrambling or randomization of direct and indirect identifiers (see sections 3.1.2 and 5.2) can generate realistic data for system testing without exposing PHI to vendors, implementers, system testers, and other third parties.

2.5 K-Anonymity

A data set provides k -anonymity for the data subjects represented if the information for each person contained in the data set cannot be distinguished from at least $k-1$ individuals whose information also appears in the data set. For example, a data set has 5-anonymity if, for every record in the data set that describe characteristics of a data subject, there are at least four other individuals also represented by records in the data set who share the same characteristics described by the record.

The following record-level data set exhibits 3-anonymity:

Table 1 Example of K-Anonymity where K=3

Pseudo ID	Age	Gender	ICD-10 Code
Patient 1	0 to 10 yrs	M	F106
Patient 2	20 to 35 yrs	F	F106
Patient 3	0 to 10 yrs	M	F106
Patient 4	51 to 65 yrs	F	F106
Patient 5	20 to 35 yrs	M	F106
Patient 6	51 to 65 yrs	F	F106
Patient 7	0 to 10 yrs	M	F106
Patient 8	20 to 35 yrs	F	F106
Patient 9	51 to 65 yrs	F	F106
Patient 10	20 to 35 yrs	F	F106
Patient 11	20 to 35 yrs	M	F106
Patient 12	20 to 35 yrs	M	F106
Patient 13	0 to 10 yrs	M	F106

The diagram illustrates 3-anonymity by grouping records that share the same characteristics. Three groups are highlighted with colored arrows pointing to the right:

- Orange arrows:** Point to Patient 1, Patient 3, and Patient 13, which all have the characteristics (0 to 10 yrs, M, F106).
- Green arrows:** Point to Patient 2, Patient 8, and Patient 10, which all have the characteristics (20 to 35 yrs, F, F106).
- Black arrows:** Point to Patient 4, Patient 6, and Patient 9, which all have the characteristics (51 to 65 yrs, F, F106).

2.6 Free-text Data

While, several tools are available that search free-text for obvious names (of patients, family members, and others) and then eliminate them without rendering the remaining text unreadable, these tools cannot guarantee anonymity. The continued presence of free-form text in record-level data is challenging from a privacy perspective as artificial-intelligence-based tools to analyse text and ensure that it does not identify the data subject are far from being commercially available. None of the tools described in this paper can completely eliminate the risk of de-identification of free-form text in record-level data.

There are tools available, however, to process text-based data, and then extract coded data from in a standardized nomenclature such as SNOMED. These tools can be used to replace the free-text data with a rigorously structured alternative: the free-form text is then discarded or suppressed, leaving only fully anonymised data in a structured format. Software tools exist to extract clinically significant data from free-form text and then replace the text with coded data⁵. Further information on the Canadian version of SNOMED can be found at <http://www.infoway.ca/lang-en/standards-collaborative/snomed-ct>.

2.7 Genetic information

While a data variable that denotes the presence or absence of a specific gene is, in theory, no different than any other variable that denotes the presence or absence of a factor contributing to morbidity (e.g., smoking or family history), a special concern is often associated with data denoting a genetic pre-disposition to a given disease, especially in terms of the harm it may do to data subjects who are denied insurance or employment opportunities. Special care must therefore be taken to limit the use of record-level data containing such variables (see section 3.1). Similar care must be exercised in regard to data on family histories.

Gene sequence data are much more challenging from a privacy perspective than data variables denoting the presence or absence of a specific disease-related gene. Even a few dozen gene markers may provide enough data to uniquely identify an individual from a genetic sample. The forensic use of such gene sequence data makes it at least as privacy invasive as a complete set of fingerprints. And unlike fingerprints, the traces of genetic material we constantly leave in our wake (skin flakes, dandruff and hair) cannot be wiped away like fingerprints or shielded with gloves. The tools described in this paper are not designed to mitigate the privacy

⁵ See for example, http://secure.cihi.ca/cihiweb/en/downloads/Kevin_Donnelly_-_SNOMED_CT_-_The_Global_Perspective.pdf

risks entailed by the disclosure of such data, especially if the data contains whole gene sequences.⁶

A particular challenge is in the area of research to establish the relationship between particular gene sequences and corresponding expressions of illness. Such research proceeds in a different manner than does research on population public health and even research on health policy. The latter are usually hypothesis driven. This allows the data custodian, through discussions with the researcher, to determine which variables to release for the research project and, of these, which variables may be released with less detail, so as to reduce risk of re-identification. By contrast, researchers seeking associations between gene sequences and illness have followed a data-mining model where associations are tested across all variables without prior hypotheses. A researcher whose analyses are driven by data mining will demand access to all data. This renders the data highly vulnerable to re-identification and statistical disclosure procedures cannot be readily applied in these circumstances.

Only strict application and enforcement of administrative procedures (such as user agreements that limit use and retention) and properly implemented security controls (e.g., encryption, access control, and physical security) are effective in maintaining the confidentiality of such genetic data.

2.8 Intended vs. Unintended Disclosures

The technical and statistical approaches discussed in this paper to control the disclosure of persons' identities only apply to intended disclosures of data. Appropriate security measures (physical, technical, and procedural) are critical to protect data against unintended disclosures of data from a variety of sources, including loss, theft and the actions of hackers. Several sources provide in-depth guidelines on the security of healthcare information, including ISO 27799 *Health Informatics – Information security management in health using ISO/IEC 27002*, published by the International Organization for Standardization and *For the Record*, published by the National Academy Press in 1997 [5]

Finally, administrative policies and procedures are needed to limit the conditions under which data are used. User agreements are a critical component of a privacy-protective management of disclosures for secondary purposes. These are out of scope of this paper.

⁶ There are two views regarding the protection of genetic information: the first treats genetic data in the same way as any other clinical information; the second considers genetic information to be qualitatively different from other healthcare information – this latter point of view has been called "genetic exceptionalism." In 2004, the European Directorate C (Science and Society), unit C3 (Ethics and Science) issued 25 recommendations on the ethical, legal, and social implications of genetic testing; including the recommendation that "genetic exceptionalism" should be avoided, internationally, in the context of the EU and at the level of its Member States, but that the public perception that genetic testing is different needs to be acknowledged and addressed. By contrast, the so-called "Montreux Declaration" at the 27th International Conference of Data Protection and Privacy Commissioners in September 2005 declared that "the fast increase in knowledge in the field of genetics may make human DNA the most sensitive personal data of all" and that "this acceleration in knowledge raises the importance of adequate legal protection and privacy".

3 General Approaches to De-Identification

Although not widely disseminated, the general statistical approaches to disclosure control have been well established for several decades and have been in use by large statistical institutions such as Statistics Canada and the U.S. Census Bureau. These approaches generally consist of ways of either reducing the informativity of the data or, if this does not sufficiently reduce risk of re-identification, suppression (i.e. withholding) of the data or record. They apply to the release of both record-level data and aggregate-level data presented in tabular or frequency format. [4,5]

The subsections that follow describe each type of tool. The following two diagrams illustrate when these various types of tools are best applied the associated decisions that must be made before applying them.

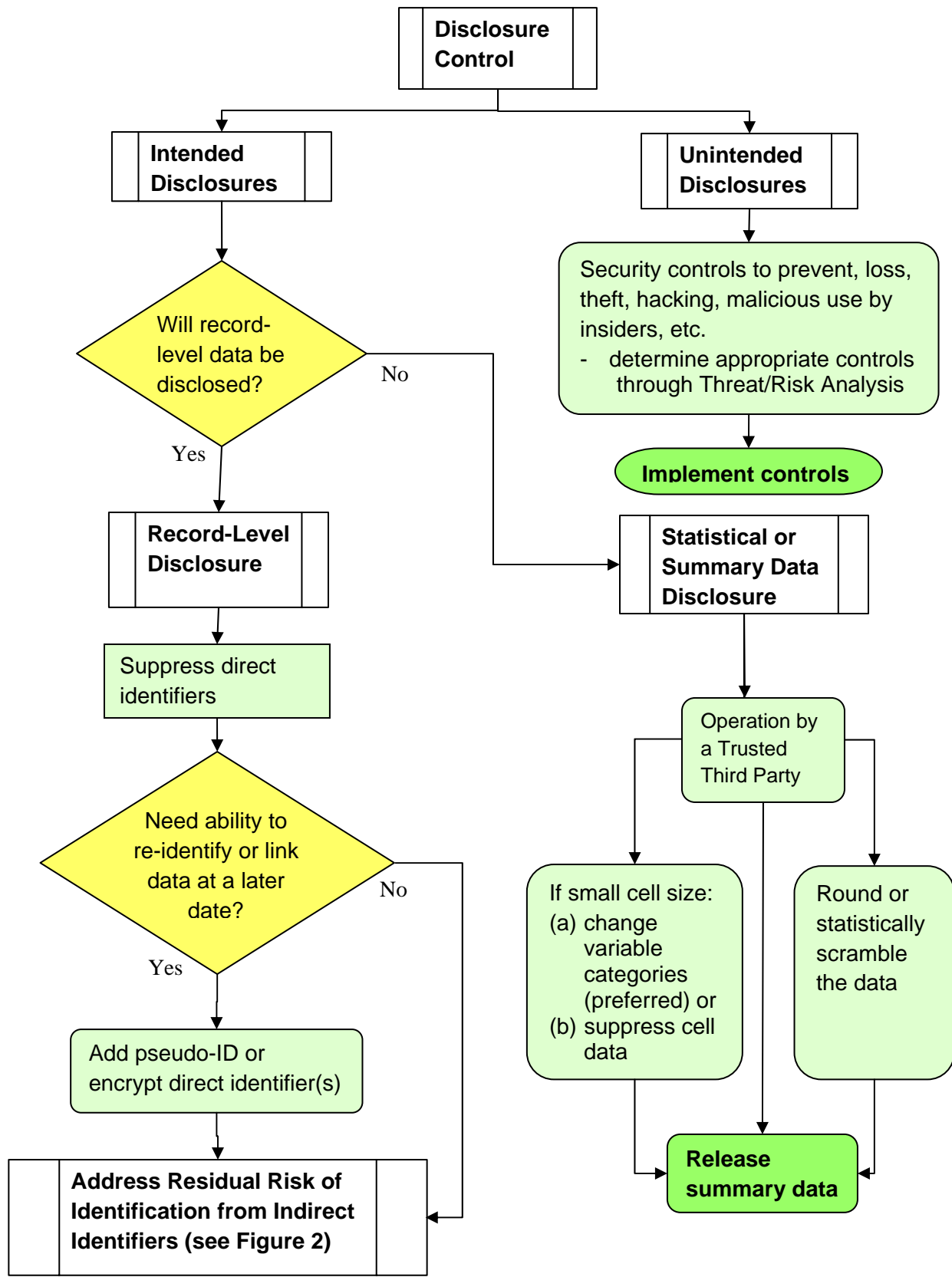


Figure 1 Applying Tools for Disclosure Control

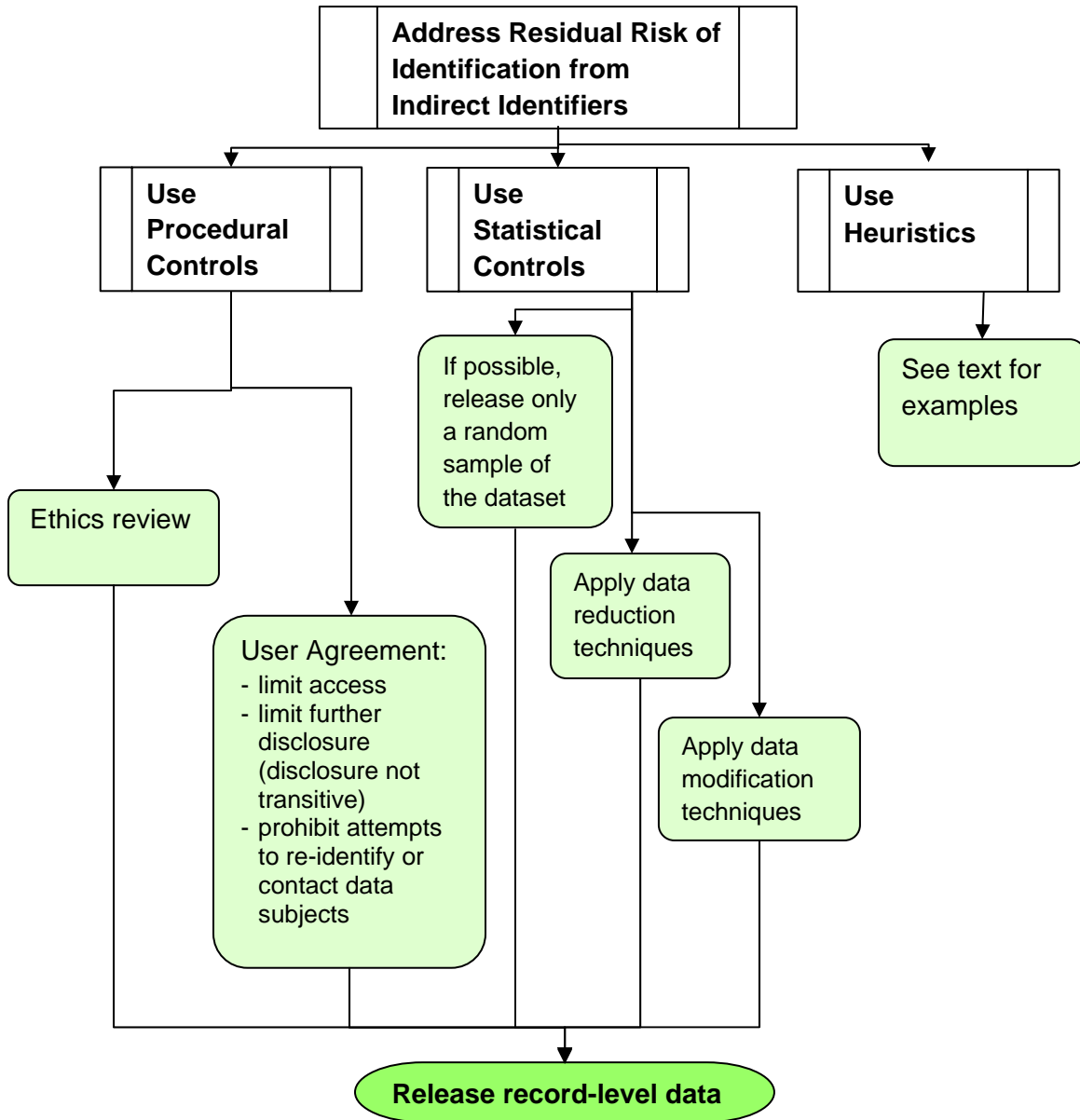


Figure 2 Applying Tools for Disclosure Control (Indirect Identifiers)

3.1 Disclosure of Record-Level Data

The three general approaches used when releasing record-level data (micro-data) are data reduction, data modification and data suppression.

3.1.1 Data Reduction

Data reduction techniques are more commonly used than data modification (see below). These generally consist of:

1. **Suppression of direct identifiers** – i.e. removal of directly identifying variables in the dataset. This is generally considered a necessary, but not sufficient, first step.
2. **Reduction in detail.** This is the most common approach used. It can be thought of as a form of rounding of the data or collapsing into larger categories. The goal is to reduce the number of records with a unique combination of potentially re-identifying variables – most commonly, dates, gender, geographic locators, and rare diagnoses – to some acceptable level of risk of re-identification. This is technically referred to as “k-anonymity” (see section 2.5). In most cases, if information on gender or diagnosis is required, these data elements cannot be reduced without severely compromising the analysis, but dates and geographic locators can. For example:
 - a) Date of birth can be rounded to year of birth. If year of birth does not provide sufficient protection against re-identification, age at a given time (say, the time of treatment) provide the same informativity required for analytic purposes. If not combined with other date variables (e.g. hospitalization date) disclosure of age would result in less potential for re-identification. If other date information were to be combined, then it may be necessary to move from age to age category (e.g. 10 year intervals) and year (instead of date) of hospitalization to reduce the risk of the combination of age plus admission date creating a unique record.
 - b) Full postal code is highly identifying. It is common to reduce this information to the first 3 digits of the postal code (forward sortation area). For some analyses, this may be too crude and so if greater detail is required regarding geographic location, greater reduction in detail must be applied to some other variable(s).
 - c) One iteratively adjusts the amount of reduction in detail for each variable of concern to levels that achieve a compromise between reducing the likelihood of identifiability and retaining the ability to conduct the analyses of interest. When combining rare health conditions with geographic location information, it may become very difficult to achieve a solution that allows for the analysis of interest without also creating high risk of re-identification – a challenge for which there is no easy solution other than analysis of data in-house and careful control over the tabular data released.
3. **Sampling.** If a dataset holds more records than required to accomplish the analysis, sampling is a particularly simple and effective approach to limiting identifiability. In the past, sampling was often used because it also addressed the challenge of the relatively high computational cost of data

analysis. In an era of high-capacity inexpensive computing power and large census-based datasets, sampling is often overlooked. However, it can be very effective in reducing the re-identification of individuals.

Here is how sampling works. Suppose an analysis was conducted using a random sample of one-fifth of the records from a database. Let's further assume that other data reduction techniques like anonymisation and reduction in detail were also used. If it happened that a unique record appeared to match a particular individual, because the data were drawn from a much larger dataset, there is still a high probability that the unique record belonged to someone else. On the other hand, if the unique record arose from an analysis of the complete dataset, the identification of a unique record results in a high likelihood of a positive identification of a particular individual. Sampling cannot, for example, protect against types of re-identification where a third party matches a dataset with another overlapping dataset. For this reason, sampling is often used in combination with other techniques, rather than on its own.

3.1.2 Data Modification

Data modification techniques involve more radical approaches to the data and have greater potential to reduce the informativity of the data. In some cases, though, they may be used as an alternative to data reduction techniques to permit analyses that would not be otherwise possible. Common examples of data modification techniques include:

1. **Random addition of "noise" to the data:**⁷ For example, suppose an analysis required high-fidelity information on age, such that the data request was for date-of-birth. One solution may be to randomly change the actual date of birth in the record within a certain pre-determined acceptable range (e.g. 6 months). Then the DOB in the record could be anything from one day to 6 months different than the actual date of birth of the data subject. As the actual date of birth of the individual in the released dataset is unknown, the ability to indirectly identify the individual is greatly reduced. Statistical packages such as SAS and SPSS can assist in this process.
2. **Randomization of data values:** This is most often used when generating test data for vendors and software developers. For example, data variables like first name and last name could be replaced by names randomly drawn from a large data set of real names typically used in Canada.⁸ Good randomization tools select names randomly with the same probability that they appear in the actual population to ensure that very uncommon names do not appear disproportionately in the de-identified data. Randomization can also randomly substitute field values such as health card numbers, while still obeying the structure and rules of real health card numbers (which typically contain a checksum digit).

⁷ The technical term for this is "perturbation" of the data.

⁸ There are many sources for such data, including mailing lists, directories of residences, electronic copies of phone books, and other lists of names and addresses.

3. **Data swapping:** Records of pairs of individuals of roughly the same characteristics of interest are identified. The values of particular variables are then swapped between the two records. As a result, a dataset is created with records that are no longer the original records but which, on aggregate analysis, will achieve the same results as would have been achieved using the original dataset. This may be acceptable for producing overall aggregate statistics but its acceptability for purposes like multivariable regression analysis is questionable. Unlike randomization of data values, data swapping always relies upon data values that appear somewhere in the original data set.

3.1.3 Data Suppression

If data reduction and modification techniques are unable to sufficiently de-identify the data, then the alternative is suppression of the records that are at high risk of re-identification. Depending on the extent of the suppression, it can introduce a high level of distortion in some types of analysis, as the suppression or loss of records is not completely at random.

3.1.4 Pseudonymisation

Whereas de-identification is the general term for any process of removing the association between a set of identifying data and the data subject, pseudonymisation is a special subcategory of de-identification. A pseudonym links de-identified data to the same person across multiple data records or information system without revealing the identity of the person. Pseudonymisation can be performed with or without a possibility of re-identifying the data subject and hence reference is made to reversible or irreversible pseudonymisation.

Pseudonymisation is recognized as an important method for protecting the privacy of personal health information. Applications include secondary use of clinical data, clinical trials and post marketing surveillance, pseudonymous care (e.g., of sexually transmitted diseases), public health monitoring and assessment, confidential patient-safety reporting (e.g. adverse drug effects), comparative quality indicator reporting, peer review, and equipment maintenance.

Reversible pseudonymisation is generally achieved in one of two ways:

- a) by encrypting identifiable information to produce the pseudonymous ID, or
- b) by deriving direct identifiers from the pseudonymous ID via a lookup-table.

Pseudonymisation is typically carried out in one of two ways:

1. pseudonymous IDs are maintained within or on behalf of an organization for a single purpose – in this situation, typically the identifiers assigned are used solely within the organization.

2. pseudonymous IDs are provided by a data warehouse custodian – in this situation, pseudo IDs can be provided to end-users that enable linking of patient health information (e.g., from datasets collected over time) while still protecting the identity of those patients.

Pseudonymisation can either map a given direct identifier to the same pseudonymous ID or else map a given direct identifier to different pseudonymous IDs in a way that is context dependent (e.g., by assigning different pseudo IDs to different researchers or research institutions) or location dependant (e.g., by assigning different pseudo IDs to data comes that comes from different data sources).

There is a technical specification on pseudonymisation from the International Organisation for Standardization (ISO) that provides guidance and best practices for pseudonymisation (ISO 25237 Health Informatics –Pseudonymisation). As well, there is additional guidance on pseudonymisation of radiology data available in *Pseudonymization of radiology data for research purposes*. Journal of Digital Imaging, 2007; 20(3):pages 284-295 (Numeir R, Lemay A, Lina J-M. authors).

3.2 Disclosure of Aggregate Data

It may be that a provincial agency or regional health authority chooses to not release record-level data but, instead, either conducts in-house analyses or enters into a contractual relationship with a trusted third-party to maintain the record-level data holdings and to conduct the analyses; releasing only aggregate or macro-level data. This holds certain advantages from the perspective of disclosure control but the release of macro-data is not without risk of re-identification – for example, in the case of tabular information with small cell counts. In such cases there are disclosure control techniques for the release of macro-data that parallel those for micro-data.

The general approaches are restriction-based methods and perturbation-based methods. They are described below.

3.2.1 Restriction-Based Methods

The two most common restriction-based approaches are cell suppression and changing the classification system:

1. ***Cell suppression*** – eliminating rows containing cells with small cell counts – is generally an inferior approach to masking tabular data. If done thoroughly, it substantially reduces informativity of the data to the end-user and, if not done thoroughly, it can often be relatively-easily reverse-engineered to fill in the missing cells. Therefore, this approach should

generally be discouraged in favour of other more effective restriction-based methods such as those described below.

2. **Changing the classification scheme** may be accomplished through either changing the cut-points for categories or through collapsing cells. Another approach is through a process called top-coding and bottom-coding. For example, it is common to have small numbers of observations at extremes of age. One solution is to collapse together all ages under a certain value (bottom-coding) or above a certain value (top-coding). Examples are provided below (see Table 2). In each case, the goal of eliminating small cell sizes is accomplished. The determination of what changes to cut-points are appropriate requires judgment and automated techniques should not be used as a substitute for informed judgment. However, automated tools can assist the user in making such judgements (see section 5). A more complete discussion of these techniques may be found in other sources. [5,6]

Table 2 Changing the Classification Scheme to Eliminate Small Cell Sizes by Changing Data Ranges

Changing the Classification Scheme by Changing the Cut-Points for Data Ranges					
Gender variable	Bottom-coding			Top-coding	
	Under 12	12 to 15	16 to 19	20+	Total
Males	23	20	18	19	80
Females	2	5	7	6	20
Total	25	25	25	25	100
Change cut points to eliminate small cell size:					
	Under 13	13 to 16	17 to 20	21+	Total
Males	26	20	19	15	80
Females	5	5	5	5	20
Total	31	25	24	20	100

Changing the Classification Scheme by Collapsing Cells To Combine Data Ranges:					
Gender variable	Bottom-coding			Top-coding	
	Under 12	12 to 15	16 to 19	20+	Total
Males	23	20	18	19	80
Females	2	5	7	6	20
Total	25	25	25	25	100
Collapse cells to eliminate small cell size:					
	Under 15		16 to 19	20+	Total
Males	43		20	15	80
Females	7		5	5	20
Total	50		25	20	100

3.3 Heuristics

In some cases, heuristics (rules-of-thumb) are applied to inform decisions about reducing disclosure risk. For example,

- a) the threshold rule (rule of 5) for small cell counts in tabular data where tables are not released with cells smaller than 5 – see the discussion in [8]; or
- b) setting a minimum population size (e.g. 100,000) for the release of a data set with geographic-specific data. This refers to any subunit of geographic analysis – by health district, by public health unit jurisdiction, by some subunit of postal code (whether it be 3, 4, or even 5 digits, depending on the number of people captured there). This heuristic is particularly crude because one has to look beyond broad population numbers and examine the target of the analysis within the dataset. So, for example, if one is examining a relatively rare health condition, a regional sample with 100,000 may still be too small. One has to know something about the underlying analyses intended before making informed decisions.

Heuristics have the advantage of being very easy to understand and follow. They are also useful when decisions must be made about identifiability *before* any data is collected (eg, when reviewing research protocols). But on their own, heuristics are

inadequate for all but the most simple of analyses, as they usually take into account a limited number of variables – often only one – whereas epidemiologic-type analyses usually take into account several variables that, together, may render the dataset more identifiable.

4 Best Practices for De-Identification

4.1 Storage of Direct Identifiers in Data Warehouses

The creation of a data warehouse provides an opportunity to secure direct identifiers and control their dissemination. Especially problematic are direct identifiers such as name, address, birth date, gender and health card number, as these could be used in identity theft as well as in a breach of privacy. Should direct identifiers be stored in data warehouses that maintain record-level data? If the possibility exists that some released records will need to be subsequently re-identified under controlled conditions (e.g., to contact the patient), then direct identifiers must be stored somewhere—they cannot be discarded entirely. It makes sense then that a lookup table be maintained that links the pseudo IDs assigned to records in the data warehouse with direct identifiers such as name, address and health card number. But this lookup table can be maintained separately from the data warehouse so that additional measures can be taken to tightly control access to the direct identifiers it contains. This lookup table is also needed if pseudonymous identifiers are being consistently applied to new data on patients as the data arrives in the data warehouse over time.

Using public identifiers as internal indexes within a data warehouse is not a security best practice. Internally generated record identifiers are always preferred because the clear separation of direct identifiers from other data tables in the data warehouse provides an additional level of security in the event of loss of data, data storage devices, or backup media. As above, the link between internal record identifiers and direct identifiers such as name or health card number can be done via a lookup table whose security can be tightly controlled.

The issue is discussed at length in Sanders & Protti [1].

4.2 Date Variables in Datasets

Other than direct identifiers such as name and address, date fields are among the most problematic variables in de-identified health datasets when it comes to risk of re-identification. Birth dates are often a matter of public record and can easily be obtained for many individuals. The friends and family of patients often know the date on which a procedure was performed (especially surgeries and other

procedures that provoke a visit to the hospital by loved ones). This is also true for dates of admission and dates of discharge. Members of the public often know the date when a trauma occurred (it may even be a news item in the public media). A patient may inform others of the date of a trip to the doctor's office (e.g.: an employer). In short, dates provide many opportunities for re-identification.

Whenever possible, dates should be eliminated in favour of durations or intervals. This can occur in one of three ways:

- substitution of an interval between one event and another, rather than reporting the dates of the two events: e.g., wait time in days between diagnosis and performance of a medical procedure instead of date of diagnosis and date of procedure, or age at a particular event (time from birth to event in years);
- anchoring, whereby a specific date is selected (say, the date of diagnosis) and all other dates are specified as increments/decrements from that anchor (e.g., visits would be specified as days before or after that anchor date) and
- substitution of a time period for a specific date: the most common example is substitution of birth year for birth date but many other substitutions can be used, including the combination of month and year, quarterly time periods (1st quarter for January to March, etc.), or even day of the week if it is relevant to the research at hand.

4.3 Location Data in Datasets

If regional codes are too specific, they should be aggregated. Where location codes are structured in a hierarchical way, the finer levels can be stripped. A typical example involves the handling of Canadian postal codes. The number of leading characters of the postal code that can safely be included in de-identified data (e.g., the first one to three characters) depends entirely upon the population of residents living within the geographic areas indicated by the codes. Tools are available to map postal codes to census areas and thence to census data on population. The Canadian Institute for Health Information (CIHI) uses one such tool (the Postal Code Conversion File (see section 5.5) in converting postal codes to meaningful and privacy protective geographic variables.

No strict guidelines exist on how large a population such location data codes should represent, as it is only one of potentially many indirect identifiers (gender, birth dates, etc.) that could interact to restrict the number of potential data subjects described by a record. For example, It has been shown that in data containing gender and birth date, there is no additional privacy protection benefit in further

restricting postal code data beyond the first three characters of the postal code.⁹ But inclusion of other indirect identifiers might invalidate this conclusion. The topic is extensively discussed in *Evaluating Predictors of Geographic Area Population Size Cut-offs to Manage Re-identification Risk*.⁹

4.4 Diagnostic Imaging Data in Datasets

Diagnostic imaging data may contain identifiable information within the data (e.g. a radiology image with patient identifiers embedded in the image). Identifiable data in the structured and coded DICOM header should be handled in accordance with DICOM Supplement 55, *Attribute Level Confidentiality*. Additional risk assessment shall be considered for identifiable characteristics of the image or notations that are part of the image.¹⁰

4.5 Pseudonymous IDs and Data Linking

Pseudo IDs, by definition, allow linking across data sets and longitudinally over time. Data custodians should therefore ensure that pseudo IDs are generated that are specific to a given research proposal or institute or researcher (as appropriate). This is easily accomplished technically and ensures that the capability to link data sets remains within the control of the data custodians.

4.6 Contractual Agreements on Use and Disclosure of Datasets

Acceptable use agreements are an essential contractual control that permits data custodians a measure of control over secondary uses of data provided to researchers (potentially including rights to audit and redress for misuse). Such agreements may be substantially more difficult to enforce against an individual than against an institution. As a result, data custodians should carefully consider whether agreements on use and disclosure are made with individuals who do not represent or otherwise obligate an institution or organization.

Another important consideration is whether the data custodian has the right to audit the data recipient for compliance with the contractual terms of the agreement. Data custodians should carefully consider the inclusion of a legal right to audit compliance with the terms and provisions of such agreements.

5 Automated De-Identification Tools

Heuristics are usually insufficient as a means of disclosure control, particularly for release of micro-data and non-automated mechanisms of manipulating multiple

⁹ Khaled El Emam, Ann Brown, Philip Abdelmalik: *Evaluating Predictors of Geographic Area Population Size Cut-offs to Manage Re-identification Risk*, J Am Med Inform Assoc. 2009;16:256–266. DOI 10.1197/jamia.M2902.

¹⁰ See ISO TS25237, Section 6.6.2.4.

variables to achieve k -anonymity are cumbersome. In response, a number of automated de-identification programs have been developed. These are grouped below into three categories:

- tools for efficiently handling direct identifiers in record-level data,
- tools for reducing the risk of re-identification from indirect identifiers in record-level data, and
- tools for handling aggregate data sets.

Below, we summarize information about currently available software packages that remove direct and/or indirect identifiers. These software tools use a wide assortment of techniques described in Section 4.

There are several approaches that are described in the academic literature that do not appear in our summary below—e.g., De-ID, Carafe, MeDS, Scrub, DBScrub, and Datafly. We have not included these for one or more of the following reasons:

- a) the tool was technologically superannuated by more recently available tools and techniques,
- b) the tool was no longer actively supported (i.e., available on current operating system platforms and making use of contemporary user interface technology), or
- c) the tool was not accompanied by appropriately detailed installation and user documentation.

Most of the software tools have not undergone extensive third-party evaluation to ensure that the product will deliver comparable results to those using unmodified data when doing complex multi-variable predictor analyses typically conducted in population health analyses. They also include techniques that are at high risk of altering the results of certain types of analyses if used without the application of careful judgment. None of these tools substitutes for good judgment when determining which indirect identifiers may be generalized to protect anonymity and which must remain at the fine-grained detail that existed in the original dataset.

5.1 Requirements for Automated De-Identification Tools

In the paper, *A Globally Optimal k -Anonymity Method for the Deidentification of Health Data*,¹¹ the authors state four requirements that de-identification tools for health data should possess:

¹¹ El Emam K, Dankar F. et al., *Globally Optimal k -Anonymity Method for the De-identification of Health Data*, *forthcoming*.

1. that indirect identifiers can be manipulated as a set of hierarchal values: many variables found in health datasets can have their values mapped to a hierarchy. For example, dates can be mapped to combinations of month and year, or to years, or to decades, etc. Postal codes can be mapped to the first three characters or to the first character. In this way the variables can be mapped from highly specific values to very general values. Tools for de-identification must be able to work with these naturally occurring hierarchies of values.
2. that discrete intervals in data ranges must be user-definable: some programs attempt to eliminate risks of re-identification by automatically mapping variables to highly irregular ranges. The authors give the example of age mapped to the ranges 0 to 9, 10 to 12, 13 to 25, and 26 to 60. Such irregular intervals may not be suitable for some kinds of research and the user needs to be able to customize the cut-points or try other intervals.
3. that when applying restriction-based methods, whatever is done to a variable to remap its values should be done consistently across all records in the data set and not just applied locally to specific records. For example, if patient age is grouped into the ranges "0 to 10", "11 to 20", and "over 20" for some of the records in the dataset, other records in the dataset should not be group into a different range such as "0 to 15", "15 to 25", and "over 25". Patient ages should be grouped into the same ranges for all records.
4. that the de-identification be globally optimal, achieving k-anonymity while minimizing information loss (i.e., preserving informativity). While it is difficult to compute an optimum balance between k-anonymity and information loss in all datasets, some programs do a better job than others.

The tools reviewed below generally meet the criteria above that are applicable to the type of tool under consideration.

5.2 Tools to Remove or Suppress Direct Identifiers in Record-Level Data

Oracle Data Masking Pack

Oracle provides an add-on tool called the Oracle Data Masking Pack that works with their Oracle 10i database. The software masks data using a variety of techniques, including encryption, data shuffling, and replacement.

- data substitution: data values can be randomly generated or replaced from a pre-defined set;
- data swapping (see the description in section 3.1.2);

- condition-based masking (e.g., one set of rules can be applied if the birthdate indicates a child or adolescent and another set of rules can be applied if the birthdate indicates an adult);
- suppression of direct identifiers, including automatically masking fields across all joined tables and masking across mirrored databases; and
- maintenance of referential integrity across database tables (i.e., if a data value is substituted for another in one table, the substitution will consistently be applied in other tables to that the tables still link together properly).

A scripting language is available to automate the masking process.

Further information can be found at:

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

Camouflage

Camouflage is a data masking tool developed by a Canadian company, Camouflage Software Inc., headquartered in St. John's, NL. Camouflage was developed with financial assistance from the National Research Council of Canada. 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 offers the following features:

- data substitution: data values can be randomly generated or selected from a pre-defined set,
- data swapping (see the description in section 3.1.2,
- maintenance of referential integrity across database tables (i.e., if a data value is substituted for another in one table, the substitution will consistently be applied in other tables to that the tables still link together properly)
- data obfuscation (a weak form of encryption) of direct identifiers to allow for reversible de-identification. Rules for masking are customizable.

Camouflage has partnered with IBM, Microsoft and Oracle to market Camouflage.

More information can be found at www.datamasking.com

Informatica Data Privacy

Informatica Data Privacy (formerly Applimation Informia Secure – Applimation was recently purchased by Informatica) 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 substitution, including generation of random values or selection from a pre-defined set (supplied);
- data swapping;
- data skewing;
- data encryption, allowing for reversible de-identification;
- maintenance of referential integrity across database tables (i.e., if a data value is substituted for another in one table, the substitution will consistently be applied in other tables to that the tables still link together properly);
- ability to mask data across different database platforms (e.g., composite data warehouses running in Oracle and Microsoft SQL) and
- extensive auditing features provided for compliance audits; these consist of audit logs and reports for all masking activities.

The company web site contains several white papers and technical reports. More information can be found at <http://www.applimation.com/leader-customers.asp>

Data Masker

Data Masker has been developed by a UK firm called Net 2000 and is used by many companies in the UK, as well as in the US and Canada. . Data Masker runs only on Windows platforms is available for Oracle (versions 7, 8, 9i, 10g and 11g), IBM DB2 UDB (versions 8.1 and greater), and Microsoft SQL Server (versions 7, 2000, 2005 and 2008). The company is preparing a Sybase version but as of September 2009, it is under development). Product features include:

- data substitution, including shuffling and replacement from user-defined substitution sets;
- data swapping; and
- data obfuscation (a weak form of encryption) via the Oracle DBMS Obfuscation Toolkit (Oracle databases only), allowing for a level of reversible de-identification.

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.

More information can be found at <http://www.datamasker.com/index.html>

IBM Optim Data Privacy Solution

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.

A related product called IBM Infosphere can analyse databases and look for indirect identifiers embedded in other fields of data (such as transaction numbers that are not meaningless unique numbers).

More information can be found at <http://www-01.ibm.com/software/data/data-management/optim/data-privacy-solution/>

5.3 Tools to Mitigate Risk of Re-identification from Indirect Identifiers in Record-Level Data

The tools below are designed to specifically address the risks of residual re-identification in record-level data sets from which the direct identifiers (e.g., name and street address) have been removed. The tools use sophisticated algorithms to accomplish an essentially simple task: apply the methods described in section

None of these tools substitutes for good judgment in determining which variables require greater detail and which require less. Moreover, most of these programs make use of data modification techniques that have not been well evaluated in terms of the validity of the results when used in conventional multi-variable epidemiologic analyses. Some of these tools may therefore be suitable for some uses (e.g., health care planning and evaluation) but not others (e.g., epidemiology). Discussion with researchers needs to take place to ensure that the tools are appropriate to the intended data use.

Optimal Lattice Anonymization / PARAT

The objective of Optimal Lattice Anonymization is to find a range of data values that minimizes information loss while still guaranteeing k-anonymity. The algorithm has been embedded in a commercial product called the **Privacy Analytics Risk Assessment Tool (PARAT)**. PARAT is a Windows based application and is

compatible with several databases, including Oracle, and Microsoft SQL Server. PARAT uses a multi-step process: the user selects the indirect identifiers to be released from the data set and then specifies a re-identification risk threshold. The program then performs a risk analysis on the indirect identifiers (i.e., after removal of the direct identifiers), based upon the presumed risk of re-identification from three sources of hypothetical attack: a prosecutor, a journalist, and a marketer. PARAT then applies several de-identification techniques to the data to reduce re-identification risk to an acceptable level.

The program is straight-forward to use and the Optimal Lattice Anonymisation algorithm is a sophisticated one. The illustration of re-identification risk by estimating the risk of re-identification from a prosecutor, a journalist, and a marketer may not be applicable to a health information custodian: custodians deal with the former through established legal processes and are unlikely to release any but the most aggregated data to a journalist. Appropriate acceptable use agreements with research institutions are typically in place to prevent disclosure or use by marketers. Nevertheless, suitably reinterpreted as an indication of re-identification risk from an adversary with a given level of sophistication, the program's indications of risk are useful heuristics in determining how far indirect identifiers need to be manipulated to reduce re-identification risk to an acceptable level. The terms are actually intended to characterize whole classes of attack. Risk of re-identification by a prosecutor risk is essentially k -anonymity whereas risk of re-identification by a marketer measures the risk of linking with another database and trying to re-identify all records. Likewise, journalist risk applies to all datasets that are samples. When re-interpreted in this way, the terms become meaningful to health information custodians. PARAT is one of the few tools that directly provides a measure of re-identification risk.

Protection can be applied through one of several methods. Global recoding is the process of recoding the categories of selected variables in the manner described above in section 3.2.1 (but with record-level data as opposed to aggregate data). Local suppression introduces missing values for some of the indirect variables of selected records.

Further information on PARAT can be found at <http://www.privacyanalytics.ca/technology/technology.html>

μ -Argus

μ -Argus (readers can search for m -argus) is made available by Statistics Netherlands, that country's national statistics bureau. The name is an acronym for "Anti-Re-identification General Utility System". The program runs under Windows 2000 and later versions. The program was developed by the *Computational Aspects of Statistical Confidentiality* project of the European Union. Recent

extensions of μ -ARGUS have been made possible during the European CENEX-SDC-project.

To use μ -Argus, the user first selects the indirect variables that could potentially re-identify data subjects. This, of course, relies upon the user's personal judgement, but once a variable has been declared (indirectly) identifying, it is then a fairly mechanical procedure to deal with the variable in μ -ARGUS.

The program first estimates the individual risk of re-identification for each record in the dataset. Technically, it estimates an upper bound for the probability of re-identification. Re-identification of a data subject can occur when this data subject is rare in the population with respect to a certain combination of values of indirect variables. In practice, this estimation can be difficult to accomplish: often the dataset is the only a sample of the population and rarity in the population (with respect to a certain combination of values of variables) can be hard to establish as there is generally no way to determine with certainty whether a person who is rare in the data set (with respect to a certain combination of values) is also rare in the population. As with other programs in this category, μ -ARGUS program therefore estimates rarity in the population based upon rarity in the dataset and calculates a risk of re-identification based on available combinations of variables.

The program also estimates the global risk of re-identification for the entire file in terms of expected number of re-identifications and the re-identification rate (whereas the first is a measure of disclosure that depends on the number of records in the file, the re-identification rate is independent of the number of records).

The user must then determine what is an acceptable level of risk. After the risk has been estimated, protection is applied, as with PARAT, through one of two options:

1. **global recoding:** recoding the categories of selected variables (global recoding) in the manner described above in section 3.2.1 (but with record-level data as opposed to aggregate data), or
2. **local suppression:** introducing missing values for some of the indirect variables of selected records.

Note that unlike global recoding (which changes the way a variable is encoded throughout all the records in the dataset), local suppression only affects the records with a high risk of re-identification.

μ -ARGUS identifies unsafe combinations of variables and iteratively allows the user to perform manual global recoding. Once the user has reduced the number of unsafe combinations, μ -ARGUS performs local suppression to eliminate all remaining unsafe combinations.

μ -Argus allows the user to experiment with different levels of acceptable risk to examine the effect each has on the resulting dataset (i.e., how much global recoding or local suppression must take place to reduce the risk to the user's preferred level).

The program can also deal with hierarchical data sets; e.g., household data which groups together individual members of a household. In this case, the program can calculate a re-identification risk for the entire household, as well as for individual household members.

The μ -Argus manual provides an excellent introduction to the methods described above in section 4, even for readers who will not be using μ -Argus. The full implementation and operational manual is available at <http://neon.vb.cbs.nl/casc/Software/MuManual4.2.pdf>

The software (version 4.2 from December of 2008) is available free of charge at http://neon.vb.cbs.nl/casc/Software/MU420_B1.zip

Incognito / Cornell Anonymization Toolkit

Incognito is an algorithm devised by Kristen LeFevre of the University of Wisconsin.¹² Incognito considers all possible subsets of the indirect identifiers. For example, in a data set with gender, age, and procedure code, the program would consider each of the three variables separately, then all combinations of gender and age, age and procedure, and gender and procedure; and then finally all combinations of all three variables together. In evaluating whether any of these combinations produce unique or rare combinations of data values, it uses certain optimizations to speed up its calculations to the point where use of the program on large data sets is still practical.

In 2009, the Incognito algorithm was implemented by a group at Cornell University and made available for general use as a program called the Cornell Anonymization Toolkit.

As with PARAT and μ -Argus, Incognito also implements global recoding.

The Cornell Anonymization Toolkit is available for download free of charge at <http://sourceforge.net/projects/anony-toolkit/>

The download does not include extensive documentation and users must puzzle their way through several program features. As the project is recent, this may be rectified in the near future.

¹² Kristen LeFevre, David DeWitt, and Raghu Ramakrishnan. *Incognito: Efficient Full-Domain K-Anonymity*. In ACM SIGMOD International Conference on Management of Data, June 2005

Further information can be found at

<http://portal.acm.org/citation.cfm?id=1559979&dl=GUIDE&coll=GUIDE&CFID=49698174&CFTOKEN=64870461>

5.4 Tools to Mitigate Risk of Re-identification from Aggregate Data

τ -ARGUS

τ -ARGUS is an extension of μ -ARGUS (see section 5.2) that is intended for tabular and frequency data – i.e. aggregate statistics. It applies similar statistical techniques to those incorporated into μ -Argus to minimize the risk of re-identification of individuals in the presentation of summary statistics. These include restriction based approaches (e.g. changing classification schemes, cell suppression) and techniques that introduce statistical “noise” into either the underlying data or to the summary statistics presented. .

τ -ARGUS requires a license to use one of two types of commercially available software packages for solving so-called LP problems: XPRESS-MP or CPLEX (only one is needed). The τ -ARGUS software is available free of charge at

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

Information on licensing XPRESS-MP can be found at

<http://neon.vb.cbs.nl/casc/Software/tauArgus2009prices.pdf>

Information on licensing CPLEX can be found at

<http://neon.vb.cbs.nl/casc/Software/CPlexForARGUS2005.pdf>

Further information on τ -ARGUS can be found at

<http://neon.vb.cbs.nl/casc/Software/TauManualV3.3.pdf>

5.5 Miscellaneous Tools

Postal Code Conversion

The Postal Code Conversion File (PCCF) is a digital file, available from Statistics Canada, that 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. PCCF was first created in 1098 by the Geography division of Statistics Canada and has been regularly updated ever since (the most recent update relies upon the 2006 census). A quality indicator for the confidence of this linkage is also available in the PCCF.

By converting from postal codes to standard geographic areas from the Census, custodians of databases are able to 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-identification.

Further information on the PCCF can be found at <http://www.statcan.gc.ca/bsolc/olc-cel/olc-cel?catno=92F0153G&CHROPG=1&lang=eng>

In-House Tools

CIHI is in the early stages of developing tools to support pseudonymisation. Statistics Canada is in the process of developing in-house tools, but these are not currently available for release to the public or other organizations.

The Régie de l'assurance maladie du Québec (RAMQ) has also developed tools in-house to assist in de-identification.

5.6 Evaluation of De-Identification Tools

At present, no organization or institution evaluates de-identification tools in a systematic manner. There are isolated comparisons in the academic literature where the authors compare their algorithm or approach to other selected approaches documented in the academic literature, but these cannot be considered an impartial evaluation of software tools for three reasons:

- while peer-reviewed, such evaluations nearly always compare previous approaches to those put forward by the authors of the evaluation, and hence are not truly impartial third-party evaluations; and
- the evaluations are often focussed on algorithms or conceptual models rather than on toolsets; the actual software tools are sometimes not publicly or commercially available or else are in a rudimentary form that would not meet many objective criteria for software maturity (e.g., robustness of development methodology, documentation, user interfaces, etc.); and
- the evaluations are not the product comparisons that are commonly available for popular categories of commercial software.

The authors are aware of proposals put forward to fund such evaluative activities but at the time of writing, no such funded activity has been approved in Canada or elsewhere.

6 Residual Risk of Re-Identification

The risk of re-identification can never be precisely zero. One has to accept some level of risk if useful data is to be used by secondary users. The important issues are how much risk is acceptable—a matter of policy—and how the risk can effectively be measured—a matter of technology. Some of the tools above, such as μ -ARGUS and PARAT, assist the user by providing risk metrics. These metrics go far in ensuring that the risk of re-identification stays within the boundaries of acceptable risk. But, ultimately, what constitutes an acceptable risk is a question that no software program can answer.

7 Conclusion

This report has examined several approaches to the de-identification of personal health data and has discussed a variety of tools that assist in the de-identification process. Used appropriately, these tools can significantly reduce the risk of re-identification in de-identified data. Appropriate use requires that these tools be combined with administrative controls and good security practices. While the focus of this paper has been on technical or statistical approaches to disclosure limitation, these technical controls are only useful when exercised in combination with other protections, including laws that provide the general framework for collection, use, and disclosure of the data and that specify sanctions for inappropriate collection, use, or disclosure. Also needed are best practices and codes of conduct to guide particular areas of secondary use. The CIHR *Best Practices for the Protection of Privacy in Health Research* [2] are a good example of such best practices.

There is always a degree of loss of control once information is released to another party. Therefore, any data custodian contemplating release of data to outside users should enter into specific agreements only with those institutions that can demonstrate that they have the necessary safeguards, policies and practices in place. Further, they should be prepared to audit the practices of any party to whom they have disclosed these data.

The tools and techniques discussed in this report are intended to protect privacy but they will only be useful if the de-identified data generated are appropriate for the type of research being undertaken. This requires a careful discussion with prospective researchers to ensure that the tools and techniques selected fit the uses to which the data will be put.

References

1. Sanders D, Protti D: **Data Warehouses in Healthcare: Fundamental Principles.** *Electronic Healthcare* 2008, **6**: 1-16. Available at: <http://www.longwoods.com/product.php?productid=19510&cat=524&page=1>
2. Canadian Institutes of Health Research Privacy Advisory Committee. **CIHR Best Practices for Protecting Privacy in Health Research** - September 2005. Available at: http://www.cihr-irsc.gc.ca/e/documents/pbp_sept2005_e.pdf. 2005. Ottawa, Public Works and Government Services Canada.
3. Sweeney L: **Achieving k-anonymity privacy protection using generalization and suppression.** *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* 2002, **10**: 571-588.
4. Black.C, McGrail K, Fooks C, Baranek P, Maslove L. **Data, Data, everywhere...: Improving access to population health and health services research data in Canada.** 2005. Ottawa, Canada, Canadian Policy Research Networks.
5. National Research Council: **For the Record. Protecting electronic health information.** Washington: National Academy Press; 1997.
6. Anonymous: **Report on statistical disclosure limitation methodology. Statistical policy working paper 22 (May 1994, revised 2005).** Washington, DC: Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget; 1994. Available at <http://www.fcs.gov/working-papers/spwp22.html>
7. Eurostat: **Manual of disclosure control methods.** Luxembourg: Office for Official Publications of the European Communities; 1996.
8. El Emam K: **Heuristics for De-identifying Health Data,** IEEE Privacy and Security, July/Agust, 2008.

El Emam K, Brown B, Abdelmalik P: **Evaluating Predictors of Geographic Area Population Size Cut-offs to Manage Re-identification Risk,** J Am Med Inform Assoc. 2009;16:256–266. DOI 10.1197/jamia.M2902.

El Emam K, Dankar F. et al., **Globally Optimal k-Anonymity Method for the De-identification of Health Data,** *forthcoming.*

Herting R, Barnes M.: **Large scale database scrubbing using object oriented software components.** Proc AMIA Annual Fall Symposium 1998: 508-512.

ISO/TR 22221 - **Health informatics -- Good principles and practices for a clinical data warehouse,**

LeFevre K, DeWitt D, and Ramakrishnan R. **Incognito: Efficient Full-Domain K-Anonymity.** In ACM SIGMOD International Conference on Management of Data, June 2005

Li J, Wang H, Jin H, Yong J: **Current Developments of k-Anonymous Data Releasing** , Proceedings of the National e-Health Privacy and Security Symposium 2006 (ehPASS'06) - ISBN: 1741071380

Malin B: **An Evaluation of the Current State of Genomic Data Privacy Protection Technology and a Roadmap for the Future.** Journal of the American Medical Informatics Association. *Forthcoming.*

Numeir R, Lemay A, Lina J-M.: **Pseudonymization of radiology data for research purposes.** Journal of Digital Imaging, 2007; 20(3): 284-295

Sweeney L.: **k-Anonymity: a Model for Protecting Privacy.** Source International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, archive Volume 10 , Issue 5 (October 2002), Pages: 557 - 570 ISSN:0218-4885

Sweeney L. **Replacing personally-identifying information in medical records, the Scrub System.** Proc AMIA Annual Fall Symposium, 1996: 333-337.

Acknowledgements

The authors wish to thank the following individuals for their time in answering our questions about material presented in this paper:

Mark Fuller
Director of Architecture
Canadian Institute for Health Information

Kerry LeFresne
Information Services Coordinator
Newfoundland and Labrador Centre for Health Information

Lucy McDonald
Chief Privacy Officer
Newfoundland and Labrador Centre for Health Information

Joan Roch
Chief Privacy Strategist
Canada Health Infoway