Pan-Canadian De-Identification Guidelines for Personal Health Information

Khaled El Emam, CHEO Research Institute
Elizabeth Jonker, CHEO Research Institute
Scott Sams, London School of Economics
Emilio Neri, TrialStat Corporation
Angelica Neisa, CHEO Research Institute
Tianshan Gao, CHEO Research Institute
Sadrul Chowdhury, University of Ottawa

(April 2007)

This report was produced for the Office of the Privacy Commissioner of Canada.

Document Information

Document Title:	Pan-Canadian De-Identification Guidelines for Personal Health Information
Original Document Date:	27th April 2007
Current Version Date:	14th May 2007
Document Version:	11
Copyright:	CHEO Research Institute, 401 Smyth Road, Ottawa, ON K1H 8L1, Canada
Contact:	Khaled El Emam (kelemam@ehealthinformation.ca)
More Information:	http://www.ehealthinformation.ca/

Other Relevant Reports

K. El Emam and S. Sams: "Anonymization Case Study 1: Randomizing Names and Addresses", April 2007.

M. Lysyk, K. El Emam, C. Lucock, M. Power, D. Willison : "Privacy Guidelines Workshop", December 2006

K. El Emam and G. Atherley (eds.): "Second Annual Electronic Health Information and Privacy Conference", November 2006.

K. El Emam: "Overview of Factors Affecting the Risk of Re-Identification in Canada", May 2006.

K. El Emam: "Data Anonymization Practices in Clinical Research: A Descriptive Study", May 2006.

These reports are available from
http://www.ehealthinformation.ca/

Table of Contents

Executive Summary

Personal health information is increasingly being collected, used, and disclosed for research, policy, and commercial purposes. However, the rise in security breaches and the pressure to share personal information in general is a concern for the Canadian public. One way to minimize the impact of security breaches that result in the loss of personal information and to facilitate legitimate sharing of personal information is to anonymize it.

In this report we present the results of a series of studies to inform data anonymization practices. The focus of the studies was to determine how anonymized data can be re-identified (i.e., reverse anonymization). By understanding how re-identification can be performed and what the risks of successful re-identification are, we can develop more effective techniques for anonymization. The focus was on re-identification through record linkage.

The list of studies we performed and a summary of their findings is as follows:

- We examined the availability of public information that can be used for re-identification attacks. A number of public sources were identified such as the Private Property Security Registry, Land Registry, telephone directory, and Canada Post address reverse lookup directory. These public sources pertain to sub-populations. It was not possible to construct a full population database suitable or a re-identification attack using record linkage (e.g., all of Canada or all of Ontario).
- When there is insufficient public information to launch a re-identification attack on a database, it may be possible to infer some of that information. We found that it was possible to predict gender and year of birth from the first names and graduation year respectively. It was not possible to accurately predict urban postal codes from other postal codes in the record (for example, a record that has the work postal code of an individual but is missing their home postal code, and we wish to predict their home postal code), but it may be possible to do so relatively accurately for rural postal codes.
- Once we have public sources of information, augmented with additional information from inference attacks, what is the probability of someone being actually able to launch a successful re-identification attack on a Canadian data set? It was found that the success rate can be quite high under specific circumstances.
- The risk quantification that we did indicates that the re-identification risks are not trivial; however, people tend to be willing to trade their privacy for some personal benefit. We examine what type of personal data Canadian job seekers are willing to expose on the public web. Are they willing to expose the type of information that is needed for a re-

identification attack? The answer is yes. Job seekers post sufficient personal information about themselves on the public web for some simple re-identification attacks.

- We then examine the kind of data that Canadians leave on their computer disk drives when they non-destructively de-commission them. The study collected 60 second hand disk drives across the country and extracted their data remnants. We found that there is little personal health information about the drive owners, but there is more personal health information on the drives that belongs to others. Furthermore, there was plenty of personal information (e.g., financial records) about the drive owners and others.

Based on our findings, we have formulated a concrete data anonymization process, with some automated tool support. Following this data anonymization process will allow custodians to manage re-identification risks in their data releases and protect the privacy of Canadians.

1 Introduction

There is growing adoption of Electronic Medical Records (EMRs) [1-5]. This is raising concern among patients, and the public in general, about unauthorized disclosure and use of their Personal Health Information (PHI) [6-10]. Rates of medical identity theft have been increasing, and the risks are exacerbated with the use of EMRs [11]. Furthermore, research and clinical applications are merging [12] and researchers are increasingly turning to EMRs as a source of clinically relevant patient data [13]. These privacy concerns therefore extend to data used for research. One way to address such privacy concerns is to anonymize PHI [14].

The objective of the studies described in this report is to produce practical guidelines for anonymizing PHI in Canada. The approaches that would work in a Canadian context are not necessarily similar to what would work in other jurisdictions because of differences in privacy laws, values of the population, and the availability of information about the population. We therefore embarked on a program of research to inform the development of such Canadian guidelines that can be of immediate practical utility.

In addition to the stipulations in federal and provincial privacy legislation, there are two other primary drivers for anonymizing PHI: (1) the rise in security breaches resulting in the inappropriate disclosure of personal information, and (2) the rise in the collection, access, and disclosure of personal information by businesses, governments, and researchers. Both of these drivers are described below.

1.1 Security breaches

A random scan of media reports on any single day will find multiple stories of personal data being lost by or stolen from corporations and governments (see http://ehip.blogs.com/ehip/ for an on-going tally). While many of these incidents are reported in the US, an examination of reports from the Ontario Privacy Commissioner indicates that there are many Canadian incidents as well where privacy breaches result in PHI being inappropriately disclosed (see Table 1).

When data management tasks requiring PHI, such as transcription and coding, are outsourced overseas, there are similar risks of breaches but with optics that are far worse. An infamous example is described as follows: “A women in Pakistan doing cut-rate medical transcription for the University of California at San Francisco medical centre threatened to post patients’ confidential files on the Internet unless she was paid more money” [15]. To prove her point, she attached several patient records to her email to UCSF. As it turns out, she was at the end of a long chain of sub-contracting: The records she was handling had been subcontracted four times.

The Pakistani woman withdrew her threat after being paid about $500 from a subcontractor indirectly involved in the scandal [15].

Report #	Summary
HI-040001-1 [16]	Computers containing progress notes (including personal health information of the patients) were stolen from a rural hospital in Ontario.
HI-050007-1 [17]	A computer reported stolen from a laboratory in Ontario included personal health information as well as the health card number of patients.
HI-050011-1 [18]	Loss of a personal digital assistant that was not even password protected by a manager at a community care access center containing PHI.
HI-050015-1 [19]	Loss of a computer from a company that provides nursing services.
HI-050022-1 [20]	A case manager at a community care access centre had a laptop computer stolen from her car which contained the profiles of 2 patients including their PHI (with diagnoses), health card number and physician information. The laptop case also contained a list of the centre's patients which included their names, addresses, admission dates, tracking numbers, and the services provided to each.
HI-050021-1 [21]	A clinic, which experienced a break-in, found that a laptop computer was missing which contained 2 years of scheduling information. This information consisted of patients' names and telephone numbers.
HI-050031-1 [22]	Three laptop computers were reported stolen from the office of a hospital's physiotherapists. One of the computers was thought to contain the PHI of 5 patients. All of the laptops were password protected.
HI-050019-1 [23]	A briefcase was stolen from the car of a public health nurse which contained the files of 2 patients.
HI-050016-1 [24]	A former employee reported the loss of a memory stick containing the evaluation data of a hospital program. This data included the hospital identification number, gender, age, marital status, education, diagnoses, dates of participation in the program, duration of illness and test results of the program's participants.
HI-050004-1 [25]	A courier's van containing patients' test reports was broken into and the reports were reported missing. These reports contained the names, birth dates, health insurance numbers and test results of various patients.
HI-050039-1 [26]	A laptop computer was stolen from a therapist's home which contained PHI of patients of the rehabilitation centre where he worked.
HO-001 [27]	Patients' health records from an x-ray and ultrasound clinic were found on the street in downtown Toronto. A reporter found the documents and printed a story about them. 10 records were turned over to the IPC office which contained the PHI of patients of the clinic.

Report #	Summary
HO-002 [28]	A woman complained that her PHI was inappropriately accessed and illegally used/disclosed during her treatment at the Ottawa Hospital. She believed that her medical record was accessed by a nurse who was not involved with her treatment, and who was also the girlfriend of her estranged husband. She believed that the information was then disclosed by the nurse to her husband.
HO-003 [29]	A medical and rehabilitation clinic left boxes containing PHI when they vacated their office space. The boxes were later found by the landlord.
HI-050044-1 [30]	A laptop computer, which contained 51 assessment reports, was stolen from the car of a psycho-educational consultant working for the school board.
HI- 050047-1 [31]	The theft of 2 laptop computers was reported after a break-in at a rehabilitation clinic. These computers contained the names, birth dates, addresses, PHI, and insurance info of 3000-4000 patients.

Table 1: Summary of some health care breaches documented by the Ontario Privacy Commissioner's Office. They vary in severity and the number of records containing PHI that are affected by an incident. The majority of reported breaches get resolved informally at the outset, and therefore there is no public report about the incident.

Security breaches that result in inappropriate disclosure of personal information can have a broad effect on society by raising the public's concern about who has access to their data and how it will be used [32]. Taking healthcare as an example, concern about privacy has caused some members of the public to take steps that may be detrimental to their well being [33], such as not being totally honest with their health care provider [10]. A survey in the US found that as many as 15% of adults have changed their behavior to protect their privacy [7]. Those behavior changes include: going to another doctor, paying out-of-pocket when insured to avoid disclosure, not seeking care to avoid disclosure to an employer, giving inaccurate or incomplete information on medical history, and asking a doctor not to write down the health problem or record a less serious or embarrassing condition. More than a quarter of teens indicated that they would not seek out health care if they had concerns about their information confidentiality [34]. In a survey of physicians in the US, nearly 87% reported that a patient had asked that information be kept out of the record, and nearly 78% of physicians said that they had withheld information from a patient's record due to privacy concerns [35]. Similar behaviors have been reported in Canada. A survey estimated that 12% of Canadians have withheld information from a health care provider because of concerns over who the information might be shared with, or how it might be used [36], and an estimated 735 thousand Canadians decided not to see a health care provider for the same reasons [37]. Such behavior changes can reduce the accuracy of health data [38-41].

Due to inaccurate data, patient safety may be jeopardized: clinicians may make treatment errors [42] or make errors ordering medications [43]. Furthermore, researchers may underestimate disease prevalence [44], and health system managers may underestimate compliance with

standards of care such as vaccination guidelines [45]. Health care organizations may be fined if they report inaccurate data to government agencies [46].

In addition to the impact such breaches have on individuals and providers, there is evidence that commercial corporations suffer a non-trivial loss in their share price after the announcement of a security breach, with greater losses when the breach involves unauthorized access to confidential data [47, 48]. Furthermore, the losses are correlated with the number of personal records affected [47]. In effect, markets punish corporations for the increased business, reputation, and liability risks when confidential data is inappropriately disclosed. Where legislation demands notification to customers when a breach occurs [49], there are also costs associated with the notification itself [50].

Moreover, individuals lose trust in organizations that collect data from them when there are security breaches involving personal data [47, 51], which means decreased loyalty and higher churn among the customer base. One survey found that 58% of respondents who had self-reported that they received a notification of a data security breach involving their personal data said that the breach event has diminished their trust and confidence in the company [52]. Nineteen percent of the same respondents indicated that they have already discontinued or plan to discontinue their relationship with the company because of the breach, and a further 40% said that they might discontinue their relationship [52]. Therefore, there are potentially severe financial consequences to corporations that lose or expose the personal data of their clients and users.

1.2 Broad data collection, access, and disclosure

Even if no security breaches have occurred, there are pressures to make personal information more generally available. For example, in January 2004 Canada was a signatory to the OECD Declaration on Access to Research Data from Public Funding [53]. This is intended to ensure that data generated through public funds are publicly accessible for researchers as much as possible [54]. To the extent that this is implemented, more of Canadians' personal data will be made available to researchers in Canada and internationally. In the context of research, the sharing of raw research data is believed to have many benefits [55, 56]. Researchers in the future may *have to* disclose their data. The Canadian Medical Association Journal has recently contemplated requiring researchers to make the full data set from their published studies available publicly on-line [57]. The Canadian Institutes of Health Research (CIHR) has drafted a policy that would consider researchers' track records in providing access to their outputs, including data, when considering applications for funding [58].

To the extent that the above contemplated disclosures of PHI are done with full informed consent of the patients, then arguably there is little to be concerned about. However, there is some pressure not to require consent from patients to collect and use their PHI, at least for observational research purposes. It has been shown that requiring consent does introduce biases

in recruitment because those individuals who do not consent, or are difficult/impossible to request express consent from tend to be different on important characteristics than those who consent and are actually recruited; and in some cases the express consent requirements also increase the cost and duration of doing the research [59-70]. Excessive restrictions on researchers' access to identifiable health information is considered detrimental to society at large because many beneficial studies could not be done [71, 72].

1.3 The case for anonymization

There are clearly risks to data custodians, as well as real consequences to the well-being of the Canadian public from continued and future actual and perceived invasions of their privacy.

One approach to minimize the risk of disclosing PHI due to a security breach is not to collect, retain, and transmit identifiable information. Effectively this means anonymizing data sets. For example, one recommendation from a recent order after a security breach was not to allow *identifiable information* to leave the health care facility [73].

There is more literature discussing the use of anonymized data sets for academic research purposes. Hence, we will use that as our main example in making the case for anonymization.

To safeguard patient privacy in research, often one of the requirements for waiving express consent is that the data be anonymized at the earliest opportunity [74]. This is important because there is evidence that individuals can be re-identified using common variables (such as zip code, date of birth, and gender) by linking to publicly available information [75, 76]. In addition, identifiability is a key consideration for Research Ethics Boards (REBs) in deciding whether or not consent is required [77].

Canadians, in general, are more willing to have their data used for research purposes if it is anonymized. One qualitative study found that Canadian patients wanted assurances that their electronic health information (from EMRs) would be anonymized before it is used for research [78]. When asked if their anonymized PHI can be used by researchers on subsequent studies without re-consenting, 80% of respondents to another survey in Newfoundland and Labrador agreed [79]. A survey in Alberta asked if respondents would agree to their anonymized PHI being obtained by researchers without consent, and 69% agreed [80]. A qualitative study found that the public would feel more comfortable if their information is anonymized before being used for research [81]. A review of public opinion surveys indicates that the majority of the public were generally unconcerned when researchers had access to their PHI if no identifying information was disclosed [82]. Similar patterns on what Canadians perceive as acceptable use of their PHI were observed for disclosures to public health organizations and the government [80].

Therefore, an effective way to safeguard privacy when data are collected, accessed, and disclosed (either deliberately as part of a legitimate business transaction, or inappropriately due

to, say, a security breach) is to anonymize the data. Furthermore, the Canadian public would be more comfortable having their data disclosed if it was anonymized.

1.4 Anonymization guidelines

To effectively anonymize data, it is necessary to have a good understanding of how data can be re-identified (i.e., reverse anonymization). Once re-identification techniques and risks are understood, then it will be easier to develop ways to anonymize data that will minimize the probability of successful re-identification. Therefore, our concern in this report will initially be on re-identification techniques and risks, which will then lead to the formulation of guidelines for effectively anonymizing data.

Research on re-identification risk also addresses another practical issue: the scope of health privacy legislation in Canada. Specifically, these acts refer to identifying information and identifiable individuals, without precisely defining what that means. For example, the Alberta Health Information Act states that “individually identifying” information, when used to describe health information, means that the identity of the individual who is the subject of the information can be readily ascertained from the information; and “non-identifying” information, when used to describe health information, means that the identity of the individual who is the subject of the information cannot be readily ascertained from the information. The Ontario Personal Health Information Protection Act defines “identifying information” as information that identifies an individual or for which it is reasonably foreseeable in the circumstances that it could be utilized, either alone or with other information, to identify an individual. Saskatchewan’s Health Information Protection Act defines “de-identified personal health information” as personal health information from which any information that may reasonably be expected to identify an individual has been removed. These definitions leave considerable room for interpretation, making consistency more of a challenge.

1.5 About this report

In our research we have conducted five sets of studies to examine the issue of re-identification in Canada:

- We examined the availability of public information that can be used for re-identification. This is reported on in Chapter 3.
- When there is insufficient public information to launch a re-identification attack on a database, it may be possible to infer some of that missing information. We examined different types of inference attacks for some common types of variables in Chapter 4.
- Once we have public sources of information, augmented with additional information from inference attacks, what is the probability of someone being actually able to launch a

successful re-identification attack on a Canadian data set? This question is addressed in Chapter 5.

- Our risk quantification indicates that the re-identification risks are not trivial; however, people tend to be willing to trade their privacy for some personal benefit. In Chapter 6, we examine what type of personal data Canadian job seekers are willing to expose on the public web. Are they willing to expose the type of information that is needed for a re-identification attack?
- Chapter 7 examines the kind of data that Canadians leave on their computer disk drives when they non-destructively de-commission them. The study collected 60 second hand disk drives across the country and extracted their data remnants.

We conclude the report with a set of anonymization recommendations and concrete guidelines based on our results. We also outline a research agenda to investigate some questions that were raised by our findings.

2 Background

In this chapter we provide some basic background on the problem of anonymization and scope the aspect of the general problem that we will be addressing.

2.1 Data release scenario

The scenario we assume in this report is that of a data holding agency, A, releasing data to someone outside A. Let that external individual or entity be denoted by E. We will also assume that the data pertains to individuals[I].

The most common ways in which data can be released is either in raw data format (also known as *microdata*), or in tabular format. Microdata consist of a record for each individual. Tabular data can consist of frequencies (counts or estimated counts) or magnitudes (for example, means, totals, and ranges). Our focus here will be on the disclosure of microdata. We assume that each individual appears only once in the microdata (i.e., there are no duplicate records for the same individual).

The original microdata file is denoted as s'. The file, s', is considered to be a sample from a larger finite population, U, where the individuals in s' would be a subset of the individuals in U. The variables in s' would be the same as or a subset of the variables in U [II].

Some transformations are made to s' in order to anonymize it. The anonymized microdata file, s, is then released. The anonymization should control the risk of personal information being inappropriate disclosed in s.

There are two types of disclosure that could be considered: identity disclosure or attribute disclosure [83]. With identity disclosure a real individual is linked to a record in s. Even if nothing new is learned about the identified individual, the fact that an individual can be linked to a record in s would be considered a successful compromise of the microdata file. With attribute disclosure something new may be learned about a real individual even if they are not linked to a record in s. For example, if unionized plumbers in Ottawa all earn the same salary and s contains some

[I] This particular assumption helps with the explanations and makes the examples more concrete, although the basic principles discussed here would be equally applicable if the data pertained to other units, such as households, communities, or businesses.

[II] An actual population data set may or may not exist in reality, but the concept of U is important for understanding some of the re-identification concepts that will be presented in this report. For example, if s is data directly from a sample survey, then a data set from a population survey (i.e., a census) will not necessarily exist.

records for unionized plumbers, then we can infer the salary of all unionized plumbers in Ottawa even though we did not re-identify any individual record in s.

Our focus in this report is on identity disclosure.

A must ensure that the risk of re-identification of the individuals in the released data s is exceedingly small. We will assume that A will not release s unless the risk of re-identification for that data set is lower than some threshold τ.

2.2 The attacker

We use the term "attacker" in this report to denote an individual or entity who is trying to re-identify the individuals in s. Other terms that are used in the statistical disclosure control literature include "data user", "intruder", "data spy", "data snoop" and "snooper". While many of these terms do have a negative connotation, this is not intended to imply that the attacker is doing something inappropriate or illegal – the re-identification attempt may be part of a legitimate business function.

For example [84], an attacker may hold a commercial database used for marketing and wished to use the data in s to update the information that they have on individuals. Or, an attacker may be a journalist or computer hacker who wishes to mischievously re-identify individuals in s to discredit A or some of A's practices.

The attacker may not be the entity E itself but may be facilitated by E. For example, the data may be shared with a researcher who is very careful and follows good practices for protecting the identity of individuals in s. As part of the research project s/he makes the data available to a number of graduate students and co-investigators. It may be one of these other individuals who plays the role of an attacker.

The nature of the attacker may change over time as well. For example, that same researcher may be careful for the duration of the project and not share the data, but then when the project is over sells the computer with all the data on it through a classified ad in the local newspaper. The purchaser finds the data and decides to do something with it. In this case the purchaser would be the attacker.

An important assumption that we make about the attacker is that s/he is rational. This means that the attacker will make optimal decisions based on his/her perceived probability of successfully re-identifying individuals in s and on the perceived value of the information to be gained from such re-identification.

Furthermore, we assume that an attacker who is attempting to re-identify individuals in s is doing so with intent. Spontaneous disclosure where a particular record is accidentally recognized (e.g.,

a data analyst building a model using s accidentally recognizes that a particular record belongs to her neighbor) would not be considered an attacker with intent.

2.3 The re-identification threshold

The threshold τ is an important parameter that has an impact on the amount of resources that A needs to put into anonymization. The lower the value of τ the greater the resources that needs to be spent on anonymizing the data.

The value of the threshold will have a direct impact on the ability of an attacker to re-identify individuals in the data set. If τ is high then it will be easier for an attacker to re-identify records in s. If it is low then it will be much harder for the attacker to do so.

One has to be careful in the choice of τ because even a low value may be risky. For example, if we assume that an attacker knows that 5% of people with a particular diagnosis will respond to a direct mail campaign. Also, assume that the attacker gets a database s of people with that same diagnosis and is able to re-identify only 10% of the records. If the attacker launches the direct mail campaign the expected response rate would be 0.5%, which is actually a respectable response rate for a direct mail campaign. This example emphasizes the need to have an understanding of attackers in order to anonymize effectively.

The choice of τ should be driven by three considerations: (a) the probability of an attacker attempting to re-identify individuals in s, (b) the consequences of successful re-identification, and (c) the impact of anonymization on the utility of the data.

If there is a high probability that an attacker exists who will attempt to re-identify the data set, then one should ensure that the threshold is low. This will make it more difficult for an attacker to re-identify individuals in s. A number of factors would influence an attack attempt [85]:

- The ability to get access to s is important for someone to attempt re-identification. For example, if s is a public use microdata file that can be downloaded off a web site then getting access to s is trivial. Whereas if there are more stringent screening requirements for access to the data, then there will be fewer opportunities for an attack. Other ways that an attacker may get s is through collusion with an authorized user of the data, and computer hacking or theft.
- The means of the attacker, including the financial means available to the attacker, the technical skills available to the attacker, and computing resources that can be used by the attacker. For example, some public registries have fees for their use and such fees may act as a financial deterrent if the attacker has limited funds.

- A rational attacker is more likely to try to re-identify s if s/he believes that the probability of success is high. As noted above the attacker's threshold for success may not be that high, depending on their motivation.
- The availability of other means to the attacker for achieving his/her objectives would influence the effort that one would make to re-identify individuals in s. For example, if the attacker is trying to add more variables to a marketing database and there are other simpler means of getting that data, then it is unlikely that s/he will attempt an attack with a low probability of success.
- Another factor that drives an attacker is the value of the identified data. If the data in s is perceived to have value to an attacker, then there is a higher likelihood of an attack. Value could mean that the data is useful for augmenting an existing database, or that the data contains sensitive health information.
- Controls on those with access to the data are important. For example, in cases where the data is being made available for public use (i.e., minimal controls), one can make the worse case assumptions about the probability of an attack. On the other hand, if the data is being made available to a researcher who often works with A, is bound by a non-disclosure or data sharing agreement which allows A to perform short-notice audits, and the researcher is known to have good records management practices in place, then there is a lower likelihood that the researcher would launch an attack.

This makes it important that the agency, A, have some understanding of the possible attackers and their motives, means, and opportunities.

If a successful re-identification will cause irreparable harm to the individuals in s because of the sensitivity of the data, for example, then the threshold should be small. Otherwise it may be reasonable to maintain a higher threshold.

The lower the value of the threshold the more the data is likely to be perturbed during anonymization. Such perturbation reduces the utility of the data. Therefore, it is important to consider how the data, s, will be used and whether the threshold is set too low. Setting the threshold too low will reduce its utility for the end-user community.

Deciding on the threshold encompasses making trade-offs. For example, the inclusion of a particular highly sensitive variable in s may make it more vulnerable to a re-identification attack. But that by itself does not mean that the variable ought to be removed. The other factors noted above must be taken into account when making that determination.

2.4 Scenarios of attack/disclosure

An attacker may launch different attacks on s depending on his/her motives. The chances of success will depend on the scenario of attack. Common scenarios of attack are:

- The attacker looks for unique individuals in s with interesting characteristics and attempts to re-identify them. For example, the attacker may notice a record with *profession=mayor* and *criminal record=true*. The attacker may then search for direct or indirect geographical information in the record to determine which town or city that individual lives in, and re-identify them. This type of attack is likely if the attacker wishes to discredit the data custodian A.
- The attacker may have a list of one or more known individuals, but is not sure if these individuals are in s. Because the records in s may contain additional information about these individuals, s/he wishes to re-identify them in s, if they exist there. For example, a spouse in a divorce case may wish to determine financial or health information by re-identifying their adversary or adversaries in a publicly released database, or a parent may look for their child's record knowing that they participated in a sexual behaviour survey.
- The attacker wishes to re-identify all individuals in s by linking records in s with records in another database which has much more information about these individuals. For example, the attacker may be a marketing company that matches an anonymized database of lawyers with diabetes with another database of all lawyers. Once the subset of lawyers with diabetes have been re-identified, they would be targeted with a marketing campaign for a new but expensive treatment.

There are different re-identification techniques that can be used to implement the scenarios. The one we will focus on in this report is called *record linkage*. In principle, record linkage techniques can be used in the above three cases, but we will assume that the attakcer is following scenario three as that is potentially the most damaging.

2.5 Terminology and definitions

Inconsistent terminology in the disclosure control literature is a recognized problem [86]. It is therefore important to be precise by defining terms.

2.5.1 Types of variables

It is useful to categorize the variables in s because the classification would have an impact on the best way to anonymize them. A common classification scheme is as follows [87]:

Identifying variables. These are variables that can directly identify individuals, either individually or in combination. Examples of identifying variables include name, email address, telephone number, home address, social insurance number, and medical card

number. Since these variables are obvious identifiers, if they are included in s then the data set is not anonymized. In some cases more than one identifying variable is needed to identify an individual uniquely. For example, the name "John Smith" appears 298 times in a search of the White Pages in Ontario. However, when the name is combined with a telephone number, the individual can be easily identified.

Quasi-identifiers. These are variables that do not directly identify an individual, but can be used for indirect re-identification. There is no universal definition of what are quasi-identifiers. There are some quasi-identifiers that have been studied more extensively than others, however, such as gender, date of birth, and postal/zip code. Quasi-identifiers may differ across data sets. For example, gender will not be a meaningful quasi-identifier if all of the individuals in s are female. We assume that these quasi-identifiers are categorical in nature or have a finite set of discrete values.

Non-identifying variables. Such variables may contain personal information on individuals, but are not useful for re-identification. For example, an indicator variable on whether an individual has pollen allergies would most likely be a non-identifying variable. In such a case the incidence of this allergy is so high in the population that it would not be a good discriminator among individuals.

It should be noted that the boundaries between the classes are somewhat fuzzy, and a variable's classification may change based on the context. However, it is still beneficial to have a way to address groups of variables in this way as each variable class is treated differently in practice.

2.5.2 Sample and population uniques

We assume that $|s| = n$ and $|U| = N$, which give the number of records in each microdata file. The categorical variable formed by cross-classifying all quasi-identifiers is denoted by X. Each of these values corresponds to a possible combination of values of the quasi-identifiers. Let X_i denote the value of X for record i in the microdata.

Let Z_{ik} be a match indicator which represents whether the i^{th} record has the same value as the k^{th} record ($i \neq k$):

$$Z_{ik} = 0 \quad if \; |X_i - X_k| = 0 \qquad \text{Eqn 1}$$

and $Z_{ik} = 1$ otherwise.

For any individual denoted by k, where $k \in U$, if:

$$\sum_{i \neq k, i \in U} Z_{ik} = N - 1 \qquad \text{Eqn 2}$$

then that individual is *population unique*. Similarly, for any individual denoted by k, where $k \in s$, if:

$$\sum_{i \neq k, i \in s} Z_{ik} = N - 1 \qquad \text{Eqn 3}$$

then that individual is *sample unique*. An individual who is population unique will by definition be sample unique if s/he appears in the sample microdata. However, an individual who is sample unique will not necessarily be population unique. Also, an individual who is not sample unique will, by definition not be population unique.

2.5.3 Coding and anonymization

A recent terminology review noted that the identifiability of a data set can be categorized as: coded or anonymized [88]. These are further described below.

Coding can be reversible or irreversible. For example, if all identifying variables were hashed then that would be irreversible. However, if they are encrypted, then that would be reversible if one has the encryption key. Common reversible coding schemes that are used are single or double coding. Single coded data means that identifiers are removed from the data set and each record is assigned a random code. Identifiers are kept in a different data set with the code to allow linking back to the original data. This is illustrated in Figure 1. The identity database would normally be kept separate from the clinical database with different access control permissions (e.g., only highly trusted individuals would have access).

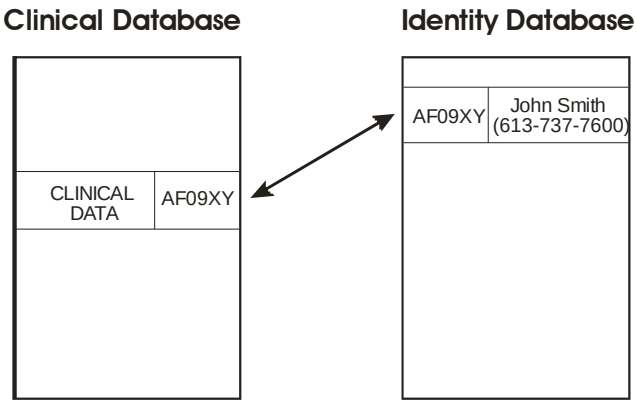


Figure 1: A clinical example showing single coding.

Double coded data means that the codes associated with the original data and the identifier data set are different, and the information linking them is kept in a separate linking database. The linking database is maintained in a secure location with a trusted third party, for example. This is illustrated in Figure 2.

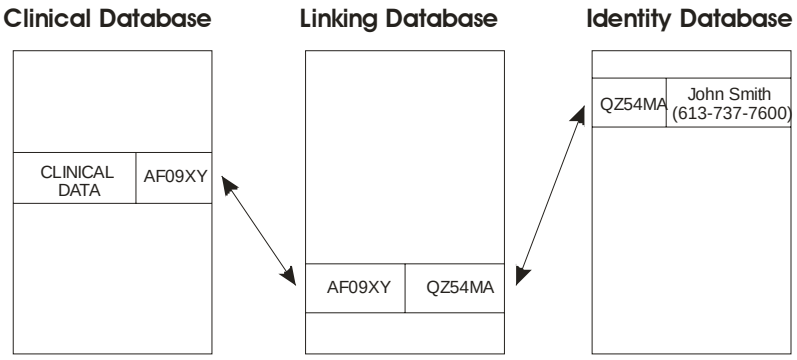


Figure 2: A clinical example showing double coding through a linking database.

To anonymize a data set, both the identifying variables and the quasi-identifiers need to be anonymized. Identifying variables can be anonymized by removing them, or through randomization. Randomization is described in more detail below.

The remainder of this report focuses on the anonymization of quasi-identifiers. The objective of that kind of anonymization is to ensure that the risk of re-identification through quasi-identifiers is below some threshold τ.

2.6 Anonymizing identifying variables through randomization

To illustrate randomization, we use as an example a data set s' contains the following identifying information: full names, addresses of individuals, and their credit card numbers. Further, we assume that the data set also contains sensitive information pertaining to these individuals. There are a number of scenarios where the disclosure of such a data set could be seen as an invasion of privacy. Example scenarios are:

> **Software Testing.** A software testing team needs to run comprehensive tests through a health insurance data processing application, and they need real data to make sure that the tests are as realistic as possible. The data that is needed includes names and addresses, as well as diagnosis and financial information about a company's clients. We also assume that the test team is separate from the main business units of the organization. For example, the testing team may be at another site, the testing function

has been contracted out to another company, or both of the above with the addition that the testing function was outsourced to a company in India. It would be risky to give the test team real customer data from the production environment.

Providing Researchers Data. A researcher wants to perform analysis on a data set that is being held by a health care facility. The data contains very sensitive medical information about the facility's patients. The facility cannot provide the real data to the researcher but is willing to take the researchers' SAS program, run it on their data and send him/her the results back. The only problem is that the researcher cannot write a SAS program that s/he knows will work on the facility's data set without first knowing what exactly the data looks like.

There are three common techniques that are used for de-identifying this kind of information:

1. Use one of the coding schemes mentioned above.
2. Remove all of the identifying information from the data.
3. Anonymize the identifying information through randomization.

The first two options will not actually meet the needs of our two scenarios. In the first scenario the test team needs realistic data which includes actual names and addresses, otherwise they would not be able to test the data processing application properly. For example, the application may not work properly with names having special characters (e.g., French names with accents). The only way testing will discover that bug is if a large number of realistic customer names are tried and some of these are names with special characters. So removing the names, hashing them, or encrypting them will not do. In the second scenario the researcher needs to get realistic data to write his/her SAS program. If the names are removed s/he may write a program that works with data sets without names and addresses, but then the program may not actually work with the real data.

Anonymization through randomization is one approach to anonymize the data. The basic idea is that an anonymizer will replace the real names and addresses with bogus names and addresses. The bogus names and addresses are taken from a large database of real Canadian names and addresses. The anonymizer ensures that all of that information looks real (for example, if it replaces a male name with a randomly selected name from its database, it will also be a male name).

The data set snippet shown (from the PrivacyAnalytics tool described in Appendix A) in Figure 3 is an example of an original data set. Here we can see the names and genders of the individuals. Figure 4 shows the data set snippet after the names have been randomized. The imputed first names are gender correct, except where the gender is unknown. The surnames are always randomized.

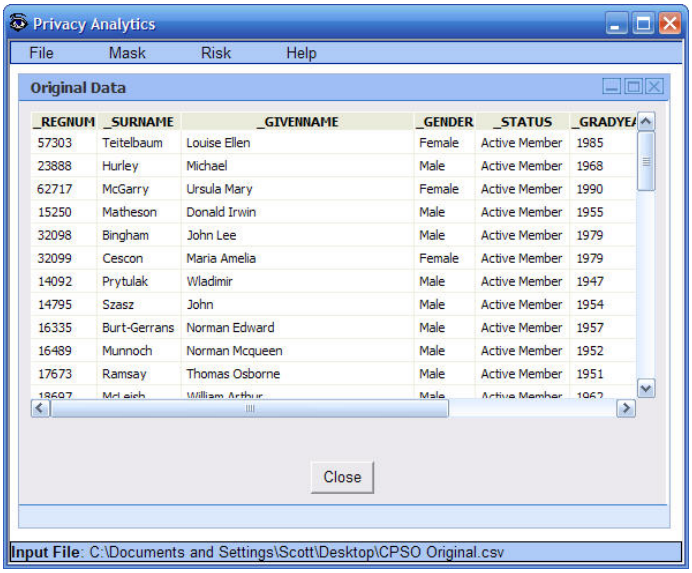


Figure 3: Screen shot of PrivacyAnalytics showing the original data set.

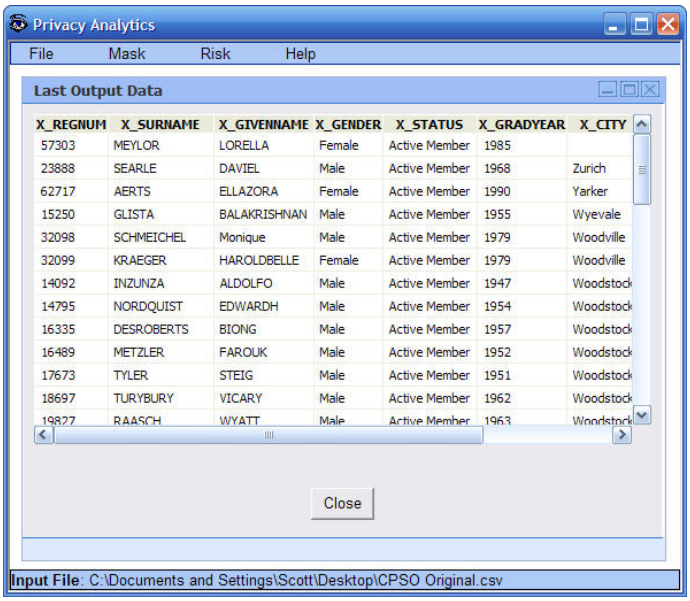


Figure 4: Screen shot of PrivacyAnalytics showing the data set after name randomization.

A similar approach is used to randomize addresses. In this case, actual addresses are replaced by other randomly selected addresses in the same Forward Sortation Area (FSA), the same city, or the same province. This would include the street addresses, postal codes, and telephone number information. Ensuring consistency across address variables is important because some end-users of s may require such consistency in their analyses. A detailed description of this anonymization case study is provided in [89].

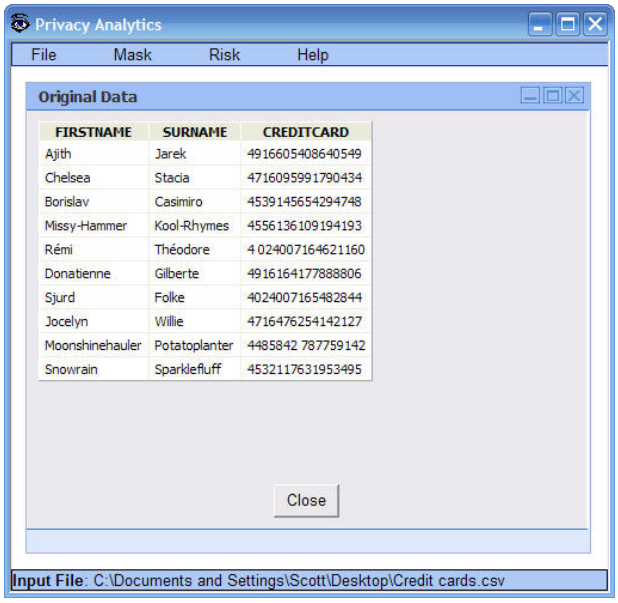


Figure 5: An example data set containing names and credit card numbers.

The data set in Figure 5 contains data on consumer credit cards (name and number). Credit card numbers can be randomized to generate valid numbers (that will pass the checksum) as shown in Figure 6. Through various settings this type of randomization can be controlled (e.g., ensure that the randomized credit card number is for the same issuer). Similar techniques can be applied for health insurance numbers as well as social insurance numbers.

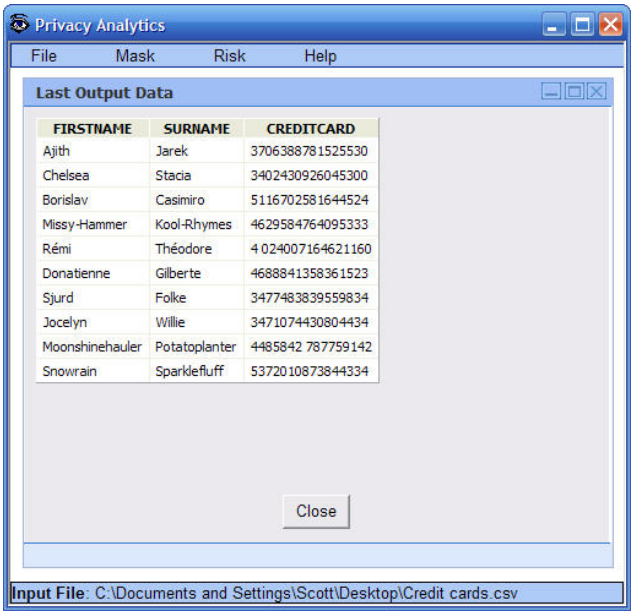


Figure 6: The credit card data set has had the credit card numbers randomized, but the new numbers will pass the validity check for credit cards.

The new randomized data set can then be disclosed after the above randomizations with confidence that the identifying information has been anonymized and that this is not reversible.

2.7 Record linkage

An attacker can attempt to match records in s with records from an *identification database*. The matching would have to be made on the basis of quasi-identifiers. Therefore, even if the identifying variables are coded or anonymized, re-identification may still be possible.

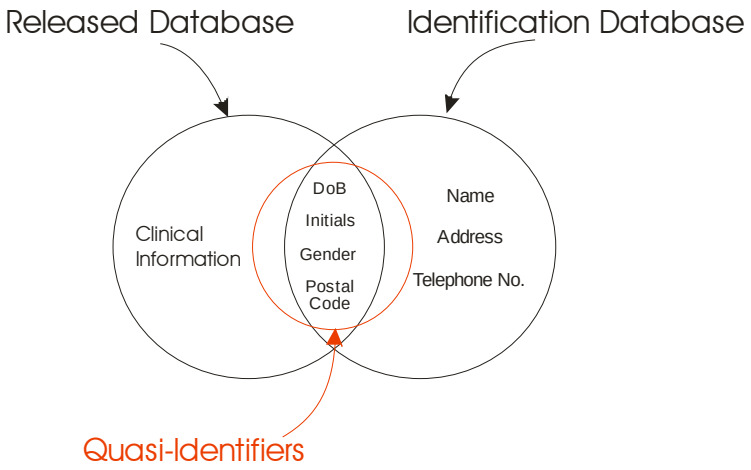


Figure 7: Illustration of how the released database can be linked with an identification database.

Figure 7 shows how record linkage would occur. Because s and the identification database have the quasi-identifiers in common (in this example they are the date of birth, initials, gender, and postal code), then an attacker could match the records in both databases. The s database does not have any identifying information, but the identification database does have identifying information (such as name, telephone number, and home address). If the record linkage is successful, we can associate the identifying information with the individuals in s and re-identify them.

There have been some explicit studies demonstrating the potential of record linkage to re-identify individuals conducted by Sweeney in the US [75]. She found that three variables: [5-digit ZIP code, gender, date of birth] can uniquely identify 87% of the US population by linking to public data sources. The variable set: [place, gender, date of birth] can uniquely identify 53% of the US population (where place is the city, town, or municipality). This means that if someone has access to s containing these three variables, then it would be possible to re-identify the subjects by performing the record linkage with publicly available information at a reasonable cost. This can be

done in the US because voter lists and other sources of personal data are publicly available, and these contain date of birth, gender, and zip code, as well as name and telephone number.

Another matching experiment in the US linked an anonymized database of Chicago homicide victims over three decades with the Social Security Death Index and was able to re-identify 35% of victims [76]. A matching study conducted in Germany found that the risk of re-identification was very small [90]. In this particular case an attempt to re-identify the sub-population of German scientists and academics was made. Yet another matching experiment conducted in the UK was only able to match a small fraction of the individuals in the database [91]. Clearly, the success of re-identification attacks is jurisdiction-dependent.

In practice, an identification database can be constructed in a number of ways:

- publicly available information from government bodies and professional associations,
- data already available to E from other sources (for example, a researcher with data available to him from another project),
- the circle of acquaintances of E, which is the set of individuals from the population about which the attacker knows the values of the quasi-identifiers,
- commercial organizations that sell databases about members of the population,
- mining the internet for information that individuals post about themselves (e.g., resumes or personal web pages),
- inadvertent access to data, such as the purchase of surplus or second hand computer equipment with data remaining in them, or
- illegal activities, such as theft of computers with data on them, or backup tapes during transit.

Even if data is not made available publicly and is not readily available for sale, if it exists somewhere one can argue that bribes, blackmail, and theft make such data still accessible under some circumstances. Therefore, arguably, the mere existence of a data source already poses some risk of a re-identification attack.

The identification database, which we will call D, may have the same individuals as U (in which case it is another population database) or may have a subset of the individuals in U. Only the individuals that are in both s and D are at risk of re-identification. The risk of re-identification is higher if D has all of the members of the population as U because then all members of s are by definition at risk of re-identification.

If an attacker knows who participated in the data collection effort—for example, s/he knows who took part in a survey—then the attacker can ensure that D has exactly the same individuals as s. In this scenario, all individuals in s are at risk of being re-identified.

It may also be the case, in practice, that not all of the individuals in s are in D, but it will not be possible to find out who the excluded individuals are. Therefore, to be prudent it is better to make the assumption that all individuals in s will be in D and manage risks accordingly. The sub-setting relationships are shown in Figure 8.

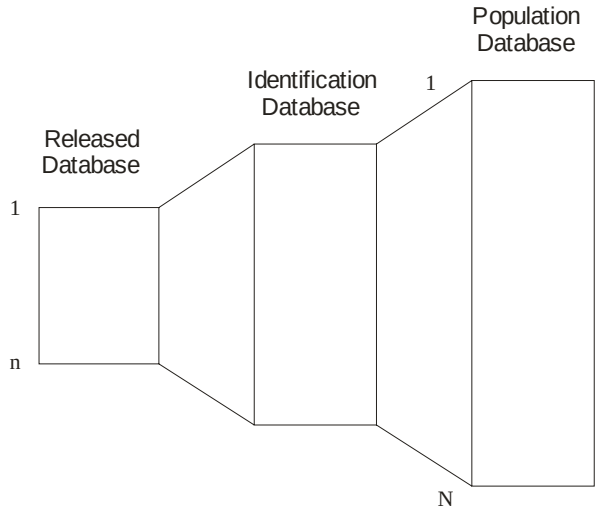


Figure 8: The released database is assumed to be a subset of the identification database, which is assumed to be a subset of a population database.

It also becomes quite complicated if it is necessary to analyze the relationship between D and U. In practice, we have to assume that $D = U$ because record linkage would only occur with D anyway, even if there exists a larger population U.

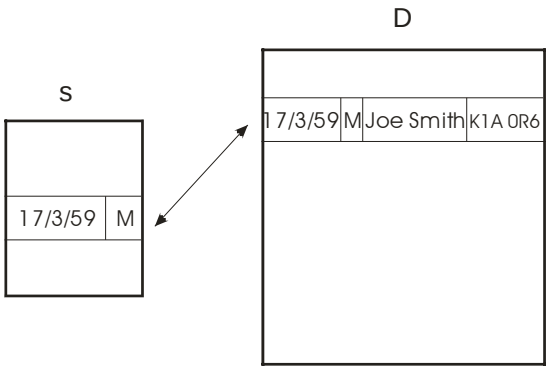


Figure 9: Example of record linkage through an identification database where an individual is population unique. The two quasi-identifiers in this example are date of birth and gender. The identification database also has the individual's name and home postal code.

To re-identify an individual, two things must happen: (i) the record in s must be matched with a record in D, and (ii) it has to be verified that the match is correct. Let us consider some scenarios with record matching.

If an individual in s is population unique and that individual exists in D, then it is almost certain that the individual will be re-identified using record linkage. This is illustrated in Figure 9. Because the individual is population unique then, by definition, that individual would be unique in both D and s.

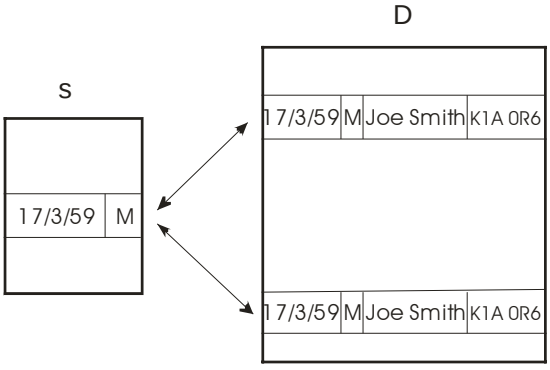


Figure 10: A re-identification scenario where the individual is not population unique.

Another scenario is shown in Figure 10 whereby the individual is sample unique but not population unique. Here the attacker will match the record in s with a record in D, but it will not be known, without additional verification effort whether the match is correct. Therefore, if an individual is not population unique then the probability of an incorrect match is at least 0.5 (i.e.,

assuming a minimum of two matches, there is at least a one in two chance that one of the matches is incorrect).

There are a few considerations when evaluating the success of record linkage when records are not population unique. If we take the example of a match with two other records, a random assignment means that there is a 0.5 chance of getting it correct. For a database s with 10,000 individuals who match doubles, that means that 5,000 will be matched correctly by chance. The attacker will not know which 5,000 are matched correctly without additional verification. However, there are many scenarios where knowing which ones are the correctly matching records would not be a significant deterrent:

- The cost of verification may be low. For example, the cost of making 5,000 phone calls to verify which ones are correct matches may be a relatively small cost for the attacker.
- When using the re-identified data, the costs of getting 50% wrong is not a deterrent if the benefit of getting 50% right is substantial.
- It may not be necessary to verify if a particular match is correct. For example, because the incremental cost of sending an email is essentially zero, an attacker who is trying to identify lawyers with diabetes for an email campaign could send emails to all double and triple, etc. matches that are found. Therefore, if two names in the identification database match to every name in s, the verification cost may not actually matter in such scenarios.

Therefore, while uniqueness puts a record at a higher risk of re-identification, the risk may not be ignorable even if a record is not unique.

3 Construction of Identification Databases

The availability of identification databases is critical to being able to launch re-identification attacks using record linkage. While there have been re-identification experiments in other nations, such as the US [75, 76], the UK [91], and Germany [90], there have been no attempts to construct identification databases in Canada. In this chapter we describe a series of studies to investigate ways to construct identification databases using *public sources* in the province of Ontario. Later in this chapter we generalize our findings outside Ontario by making the case that the methods can be replicated relatively easily in other parts of Canada.

A public data source is defined as providing data that is available to the general public, for free or for a reasonable fee with a reasonable amount of effort to get access. If a formal application is needed to access the data, a data-sharing agreement needs to be signed which restricts data use, or a protocol describing how the data will be used is necessary then that was not considered a public source.

3.1 Methods

3.1.1 Identifying public data sources

All twenty nine Ontario government ministries were contacted. We identified staff in the freedom of information and privacy (FOIP) office in each ministry, if one existed. In all ministries, with one exception, the FOIP office was contacted and we conducted a telephone interview with staff about the data that they release and the procedures for us to get that data.

A sample of commercial information brokers in Canada claiming to sell population databases were contacted to determine the type of data that they hold, the sources of data, how the databases they sell were constructed from the sources, and conditions of disclosure. After examination of their web sites we followed-up with phone calls to verify the information and get additional details. The commercial information brokers contacted were: Americanada, Prospects Influential, Nation Reach, and InfoCanada.

Two sources of genealogical data were examined as well: the Ottawa public library and the National Archive Center. Genealogical data include birth, baptism, death, marriage, adoption, and divorce data. In both of these locations staff on-site were interviewed to determine the types of data that are available and how that data was released.

Professional societies frequently release comprehensive member lists. In some instances work addresses and gender are also provided. We examined a sample consisting of the College of

Physicians and Surgeons of Ontario (CPSO), Law Society of Upper Canada (LSUC), Professional Engineers of Ontario, College of Physiotherapists of Ontario, and the College of Occupational Therapists of Ontario. For all the above professional societies, the membership lists were available on the web. Commercial brokers also provided lists of professionals, such as LexisNexis, WestLists, LawyerLocate, and Martindale. For commercial organizations, the data holdings were advertised on the web sites. We followed up with phone calls to ensure the accuracy of the information on the web and to fill in any missing details in our understanding of their data holdings.

We also contacted Statistics Canada and examined the information in the various products from the 2001 census data set. In particular we focused on tabulations giving gender and age, and on microdata releases. Additionally, we contacted Elections Canada to confirm restrictions on the use of voter lists, interviewed senior members of a national political party to understand the process for managing voter data, and interviewed a convenience sample of volunteers in election campaigns to understand how voter lists are used in practice.

3.1.2 Creating identification databases

An identification database consists of two elements: (i) quasi-identifiers, and (ii) identifying information. There are two general methods that can be used for constructing an identification database:

- A **direct method** where a public source will have both elements needed for an identification database. An example of that would be a voters list.
- An **indirect method** where we first find a source with the identifying information about individuals, and then these are linked with other sources that have the quasi-identifiers.

We followed both methods to creating an identification database.

3.2 Results

3.2.1 Direct method

The privacy offices at government ministries do provide oversight on the release of data. However, they are unable to control all possible releases, and therefore only intervene when there is a complaint, an access to information request, or when they are asked for assistance from one of the departments. Most of the privacy offices that were contacted stated that they do not sell, share with external parties, or make available to the public any personal information. The exceptions are described further below.

The commercial information brokers we contacted linked publicly available Statistics Canada census data with telephone directory data. Because of the aggregations performed on census data that are released, information such as age is only approximate. In addition, these methods would still not produce complete population databases because not everyone has a telephone

registered in their name. A recent independent study has confirmed that this is the approach used when commercial brokers utilize public data [92].

Birth and death notices are available from the General Registrar of Ontario. However, it is necessary to prove a relationship to the individual about whom data is being requested to get access to that information. Driver's license information requires the name and the driver's license number in advance to be able to make an information search request. In both of these cases it is therefore not possible to construct a database for record linkage.

The voter lists are made available to candidates or their party representatives. These lists include the name, address, and date of birth of eligible voters. That information is to be used solely for the purposes of an election, including raising funds. Party members participating in an election campaign are bound by the party oath in terms of protecting that information. Volunteers on election campaigns who are not party members would not be bound by an oath, and would not normally sign a confidentiality agreement. Therefore, there are ways to get the voter list for the purposes of a re-identification attack, but that would require deceptive practices and such use would likely go against the Elections Act.

Some commercial brokers may collect data sets directly from the public through surveys, subscription lists, or they may purchase these from retailers (e.g., loyalty card users or warranty card information). These data sets may contain the quasi-identifiers we are interested in as well as identifying information. However, these do not include all members of the population.

We were therefore unable to construct an identification database for the whole population using the direct method.

3.2.2 Indirect method

We were able to construct an identification database using the indirect method. However, it was not possible to do so for the whole population, but instead for specific sub-populations: professionals and home owners.

3.2.2.1 Professional identification databases

We constructed identification databases for physicians, lawyers, and federal civil servants in Ontario. The list of physicians is published by the College of Physicians and Surgeons of Ontario (CPSO), the list of lawyers is published by the Law Society of Upper Canada (LSUC), and the list of civil servants is published in the Government Electronic Directory Service (GEDS).

3.2.2.1.1 CPSO and LSUC

It is possible to link the list of members (which includes name, practice/firm address, and gender) with the Ministry of Government Services' Personal Property Security Registration (PPSR) data

and the Canada 411 telephone directory data (both available on the Internet, the former for a fee) to identify the home postal code and date of birth (see Figure 11).

The PPSR allows individuals to register a notice of security interest or lien on personal property (e.g., cars, boats, and furniture) that was used as collateral to obtain a loan. It also allows individuals to make enquiries to find out if a notice of security interest or lien has been filed.

We created a random sample data set of 236 physicians and 189 lawyers across Ontario with the quasi-identifiers under study. This represents a 1% sampling fraction of all registered physicians who are still active and practice in Ontario (23,506) and all practicing lawyers in Ontario (18,728). The variables in our identification database were: full name, gender, graduation date (CPSO only), date of birth, address for place of work (practice/firm), home address, and home telephone number.

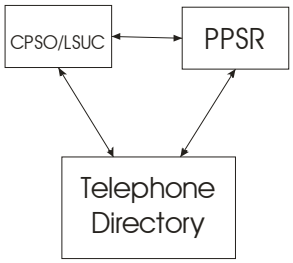


Figure 11: The three main source databases used to construct an identification database for a professional sub-population.

	CPSO	LSUC
Ability to get home postal codes (source: PPSR and telephone directory)	60%	45%
Ability to get practice/firm postal codes (source: CPSO/LSUC)	100%	100%
Ability to get date of birth (source: PPSR)[III]	40%	45%
Ability to get gender (source: CPSO/genderizer for LSUC data)	100%	100%
Ability to get initials (source: CPSO/LSUC)	100%	100%

Table 2: Ability to get various data elements on physicians and lawyers, with the source of the data (n=236 for CPSO and n=189 for LSUC).

[III] As demonstrated in Chapter 4, it is possible to predict quite accurately the year of birth from the graduation year.

Table 2 shows the success rates in getting the quasi-identifiers for D on these two professions. Name (and initials), practice postal codes and gender are available from the CPSO, so we can therefore obtain these for all physicians. Name and firm postal codes are available from LSUC. Since the LSUC does not publish gender in their public listing, genderizing software (see the analysis of the accuracy of such tools in Chapter 4) was used to estimate gender for the lawyers from their first names. We were able to determine the home postal code and date of birth from the PPSR for both professions. Additional verification of identity and home postal code was performed by checking against the Canada 411 web site (on-line telephone directory). To verify that matches were correct, we also consulted the land registry in some instances to confirm addresses. Records were flagged for additional manual investigation under two conditions: (1) Ontario Maps and the Euclidean distance between the longitude and latitude of the work and home postal codes were used to verify that these were plausibly close to each other (less than 100km), and (2) for physicians, the graduation date and the date of birth had to have a span of at least 25 years.

As evident in Table 2, it was not always possible to get the date of birth (40% and 45% success rates for physicians and lawyers respectively) and the home postal code (60% and 45% success rates for physicians and lawyers respectively). There was also a gender difference. We were able to get the home postal code for 49% of all female physicians (vs. 63% of all males), the date of birth for 29% of all female physicians (vs. 45% of all males), the home postal code for 40% of all female lawyers (vs. 48% of all males), and the date of birth for 40% of all female lawyers (vs. 48% of all males).

3.2.2.1.2 GEDS

GEDS is an on-line database of a subset of the federal civil service. The total federal civil service has 386,630 employees distributed as shown in Table 3.

GEDS has more than 170,000 records [93], which is less than 50% of all civil servants. There are a number of reasons for this:

- Organizations choose whether or not to list their employees.
- Employees in some cases have to take specific action to be listed in GEDS. Many employees do not know that they have to do this.
- Some employees are exempted because of the nature of their work.
- Similarly, some organizations such as DND and CSIS would not be listed.

Therefore, GEDS is not a comprehensive listing of federal employees. However, within the scope of GEDS, we selected 40 employees working in Ontario-based health care related departments

and attempted to obtain their missing quasi-identifiers from the PPSR and the telephone directory as described earlier.

Province/Territory/Other	Number of employees
Newfoundland	7,158
PEI	3,611
Nova Scotia	23,823
New Brunswick	14,610
Quebec	80,104
Ontario	159,652
Manitoba	16,601
Saskatchewan	9,637
Alberta	28,046
BC	38,081
Yukon	580
NWT	1,185
Nunavut	313
Outside Canada	3,229

Table 3: Distribution of Canadian federal civil servants (2006 figures) [94].

We were able to obtain the home address for 50% (20/40), home telephone number for 40% (16/40), gender for 100% (40/40), and date of birth for 22.5% (9/40) of the GEDS entries that we selected. While this is not a random sample, as was the case for the CPSO and LSUC, it does indicate the potential for the construction of D for matching with public servants in GEDS. A complete record with all of the quasi-identifiers was obtainable for 9 out of the 40 individuals.

3.2.2.2 Homeowners

In this section we show how we can construct D for a given postal code. Postal codes are often released in s. If a record in s is for a homeowner, then how easy would it be to re-identify that individual ?

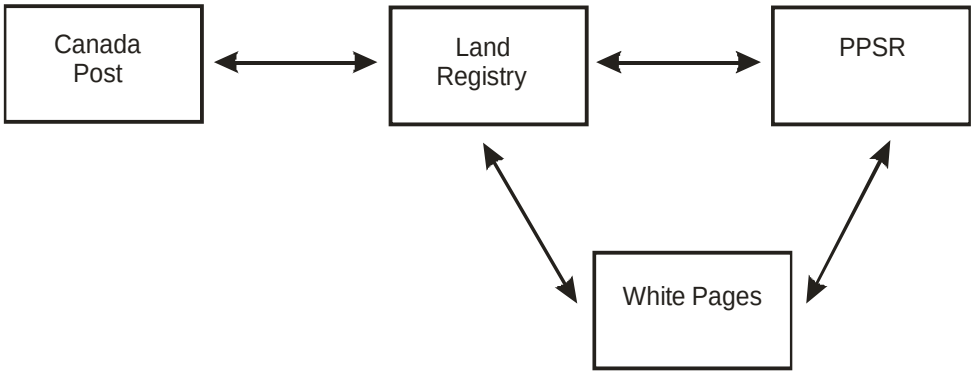


Figure 12: The four main source databases used to construct an identification database for a home owner sub-population.

To construct D for a postal code, we start by getting a list of all the properties within a particular postal code from Canada Post. Canada Post makes available a reverse lookup function on its website that allows visitors to enter a postal code and retrieve a listing of all addresses in that postal code. Satellite imagery is available for many parts of the country – particularly urban areas – through Google Maps. By entering a postal code on Google Maps, one can determine whether a given postal code or address is industrial, commercial, mixed use, or primarily residential in nature. For each property, a search in the land registry would identify present and past homeowners by name. Once we have a comprehensive list of all homeowners in a particular postal code, we follow the indirect method described above by doing searches in the PPSR database. The overall linking that is required is shown in Figure 12.

For our analysis two urban postal codes were selected from demographically similar neighborhoods of Ottawa and Toronto. The two urban postal codes examined yielded 35 residential addresses. All but one property was owned by individuals. The results in terms of the percentage of individuals for whom we were able to get useful quasi-identifiers are presented in Table 4.

	Ottawa	**Toronto**
Ability to get initials	93%	100%
Ability to get date of birth	33%	40%
Ability to get home telephone number	80%	50%
Ability to get gender (using genderizing software)	87%	95%

Table 4: The ability to get quasi-identifiers for home owners for two Ontario urban postal codes.

3.3 Discussion

An important pre-requisite for a record linkage attack is the ability to construct an identification database. We found that it is not possible to construct an identification database for the whole population of Ontario. We were unable to do that using public sources, with either the direct or indirect methods.

In Canada, the ability to access and use information is qualified by legislative restrictions designed to protect the privacy of individuals. This information may consist of what otherwise may be considered “public data” in other countries, e.g., drivers license databases or public information.

In some instances, population databases are available for access but have certain data elements removed. For example, in Ontario, personal information is collected by the Ministry of Transportation under the authority of section 205 of the *Highway Traffic Act*. The information forms part of a public record and is used for the administration of the Ministry's driver, vehicle and carrier programs. However, while residence address information is collected, it is not considered part of the public record and is not available to the general public. A further qualification is that only "authorized" requestors who have been approved and have entered into a contractual agreement with the Ministry may obtain residence address information for certain limited purposes. These purposes do include research by educational or research organizations. This limited degree of access is safeguarded by application of public sector privacy legislation in Ontario - the *Freedom of Information and Protection of Privacy Act* (FIPPA). The federal government and each of the 13 provincial/territorial jurisdictions in Canada have similar legislation designed to protect the privacy of individuals and to protect personal information held by government bodies.

Under such laws, "personal information" is broadly defined to generally mean recorded information about an identifiable individual, including "any identifying number, symbol or other particular assigned to the individual". Once it has been determined that a record contains personal information, these types of statutes generally prohibit the disclosure of this information, except in certain circumstances. One instance where disclosure may occur is where “personal information collected and maintained specifically for the purpose of creating a record available to the general public”, which is the case with the PPSR database that we used.

The preceding discussion was directed to *government* holdings of information. The use of *publicly available information* held by non-public sector entities is governed by private sector privacy legislation that exists in Canada. At the federal level and in those jurisdictions that do not have comprehensive personal information protection statutes, the legislation in question is the *Personal Information Protection and Electronic Documents Act.* British Columbia, Alberta and

Quebec have their own statutes that place restrictions on the collection, use and disclosure of personal information by non-public sector entities.

Generally, the provincial statutes governing non-public sector entities apply to publicly available information, making the use of such information subject to a consent requirement. Use without consent is permitted for certain prescribed sources of information. The federal statute permits the collection, use and disclosure of publicly available information but then defines "publicly available information" by regulation. These include names, addresses, and telephone numbers in a telephone directory; name, title, address, and telephone number that appear in a professional or business directory available to the public; and personal information that appears in a registry collected under a statutory authority.

In our case it was possible to construct identification databases for professionals whose associations or employers publish their membership lists. We found that it is more difficult to construct an identification database for adult females. It would also not be possible to perform a similar exercise on youth because youth would not have any loans that are registered, would not have property registered in their names, and would not have telephone numbers in their names. Therefore, their names would not appear in any of the publicly available data sources that we investigated. Also, it would not be possible to do so for professional associations that do not publish their membership lists.

It was also possible to construct identification databases for specific postal codes. This means that if the database s has full postal codes, it would be relatively easy to get a list of all unique postal codes in s and construct an identification database for each one which could then be matched with variables in s. As we have shown, it is possible to get dates of birth and gender information for a significant percentage of home owners. With such quasi-identifiers it would be relatively easy to re-identify individuals in s.

There does exist a financial deterrent to constructing identification databases. At the time we did this study, there were 23,506 physicians registered in Ontario who were still active and practicing in the province. To attempt to construct a complete identification database with records containing names, addresses (including postal codes), gender and date of birth for all physicians in Ontario would cost at least C$188,048 because of the PPSR search fee (which is C$8 per search). Similarly there were 18,728 registered lawyers, making the minimal cost for attempting to construct an identification database C$149,824. An attacker would need to construct a complete identification database for re-identification rather than just samples (as we did).

For homeowners, the land registry search fee is around $30, plus there is the PPSR fee of $8 per search. However, constructing an identification database for a specific postal code with, say, 20 homeowners would be quite small: $760. Therefore, there is no real financial deterrent for re-

identification of homeowners if s contains postal codes. For the financial deterrent to be effective then, s should contain at most the FSA, which is the first three characters of the postal code. As can be seen from Figure 13, the number of households who live per FSA is quite large. For example, the median number of households per Ontario FSA is just above 8000. Thus, the median cost of creating a homeowner identification database for an Ontario FSA would be $240,000, which is a significant financial deterrent. The smallest FSAs are in New Brunswick where the median is just under 2000 dwellings.

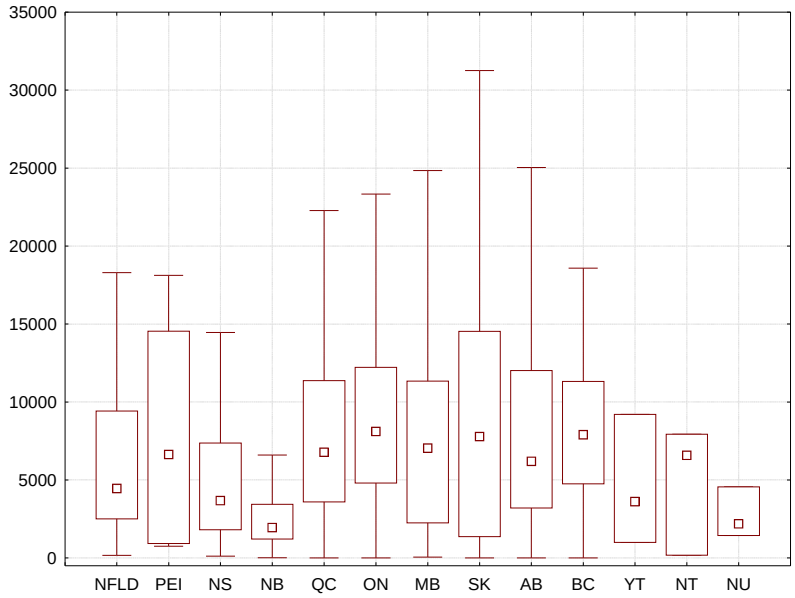


Figure 13: Box and whisker plots showing the number of dwellings per FSA across all of the provinces and territories. The figure shows the median, inter-quartile range (IQR), and whiskers at 1.5 the IQR. Outliers and extreme values have been removed.

4 Inference Attacks

Sometimes the identification database D may not contain all of the quasi-identifiers that are needed to successfully re-identify individuals in s. Even in such circumstances, it may be possible to predict the values of the missing quasi-identifiers accurately using other information that exists in D. Similarly, it may be possible to infer missing variables in s from other variables, creating new quasi-identifiers that would make it easier to match with an identification database.

In this chapter we investigate how easy it is to predict some common quasi-identifiers from other variables that one would find in either s or D. Specifically, we evaluate how easy it is to: (1) predict gender from first name, (2) predict the year of birth from the graduation year, and (3) predict one postal code from another postal code each indicating, for example, different addresses for the same person or two different parties in a transaction.

4.1 Inference of gender

Gender is an important quasi-identifier as it is often found in s. In some cases the D database may not have gender, but it would have the first name of the individuals. We evaluated the accuracy of using genderizing software to predict gender from the first name.

4.1.1 Methods

The data set that we tested with was the list of 23,506 practicing physicians in Ontario, for which we knew the correct gender. A search for genderizing software was performed on MedLine (no date limit), Journal of Marketing (2002 to present), Marketing (January 1996 to present), as well as web searches on Yahoo and Google. The search terms used were: (genderizer or genderizing or genderizing) and (software or tool or API). Nine products were identified as well as the gender list provided by the US Census Bureau. Of the products, a number of them used the same underlying API. We contacted the vendors for the remaining products and were only able to successfully contact and purchase four products. The Census Bureau list is available for free.

Each of the four products, as well as the Census Bureau list, was used to predict the actual gender for the list of physicians. Prediction accuracy was evaluated using standard classification accuracy measures: overall percentage of accurate classifications, precision, recall, and the f-measure.

4.1.2 Results

The results are shown in Table 5. While the accuracy measures are quite high overall and tend to be quite close to each other, Personator with Genderbase had the best results for this data set.

This is the tool that we use in Chapter 3 to predict the gender in the lawyers data set (LSUC). Given that this data set consists of heterogeneous Canadian professionals working in an Anglophone environment, it is reasonable to generalize to other similar groups. However, we cannot make broader generalizations to professionals, for example, in Francophone areas (e.g., Quebec).

	Male	Female
	ParseRat (overall accuracy=0.81)	
Precision	0.988	0.989
Recall	0.818	0.80
F-measure	0.89	0.88
	Personator (overall accuracy=0.81)	
Precision	0.98	0.99
Recall	0.82	0.79
F-measure	0.89	0.88
	Personator with Genderbase (overall accuracy=0.89)	
Precision	0.98	0.98
Recall	0.9	0.87
F-measure	0.94	0.93
	MAILERS+4 (overall accuracy=0.78)	
Precision	0.988	0.997
Recall	0.78	0.77
F-measure	0.87	0.87
	US Census Bureau (overall accuracy=0.77)	
Precision	0.98	0.996
Recall	0.77	0.78
F-measure	0.86	0.88

Table 5: Results of evaluating the accuracy of various tools for predicting gender from the first names. The *overall accuracy* shown in the table is the simple proportion of overall predictions that were accurate. Precision, recall, and the f-measure are standard measures of binary classification accuracy.

4.2 Inference of year of birth

We often found D databases with the graduation year of individuals, but no information about their age. However, often age information is important for re-identification. In this analysis we evaluate the accuracy in predicting the year of birth from the graduation year.

4.2.1 Methods

In the data we obtained for the College of Physicians and Surgeons of Ontario, we had the graduation year for all physicians, and the year of birth for a subset of the 1% sample that we looked up in the PPSR database (recall that we were not able to get date of birth information for all of the sample).

Our initial (ordinary least squares) regression model attempted to control for gender based on the assumption that the relationship between age and graduation year would be gender specific. The model also allowed for the possibility of an interaction effect as follows:

$$YearOfBirth \sim GradYear + Gender + \left(GradYear \times Gender\right) \quad \text{Eqn 4}$$

The terms in this model were removed if they were found not to be significant.

4.2.2 Results

As can be seen from the results in Table 6, there is no need for the interaction effect as that was not statistically significant. When we built the model controlling for gender only, gender was also not statistically significant as seen in Table 7. Therefore, we removed that as well.

Parameter	Estimate (p-value)
Intercept	250 (p=0.022)
GradYear	0.86 (p<0.0001)
Gender	36.43 (p=0.734)
GradYearXGender	-0.0187 (p=0.73)
R^2=0.85 (p<0.0001)	

Table 6: Ordinary least squares regression model for predicting year of birth with an interaction effect (n=89).

Parameter	Estimate (p-value)
Intercept	275 (p=0.0007)
GradYear	0.85 (p<0.0001)
Gender	-0.619 (p=0.2339)
R^2=0.85 (p<0.0001)	

Table 7: Ordinary least squares regression model for predicting year of birth without an interaction effect (n=89).

The final model is a simple bivariate one, but it demonstrates a very high goodness of fit as shown in Table 8.

Parameter	Estimate (p-value)
Intercept	265 (p=0.001)
GradYear	0.85 (p<0.0001)
R^2=0.85 (p<0.0001)	

Table 8: Final ordinary least squares regression model for predicting year of birth (n=89).

Using a leave-one-out approach (iterate n times and at each iteration remove 1 different observation, build the model with the n-1 observations, and predict the removed observation) we predicted the year of birth from the graduation year. The leave-one-out approach alleviates the optimism inherent in using the same data set for building a model and testing it with the same data set. A graph showing the relationship between the actual year of birth and the estimated year of birth is shown in Figure 14.

The root mean square error (RMSE) for this prediction is 4.4 years. However, there were three physicians, all born in 1953, who graduated at ages between 44 and 46. These were higher graduation ages than the rest of the sample. If we remove these three individuals the RMSE becomes 3.4 years. In general, however, the graduation year can be used to predict year of birth within a relatively narrow range of error.

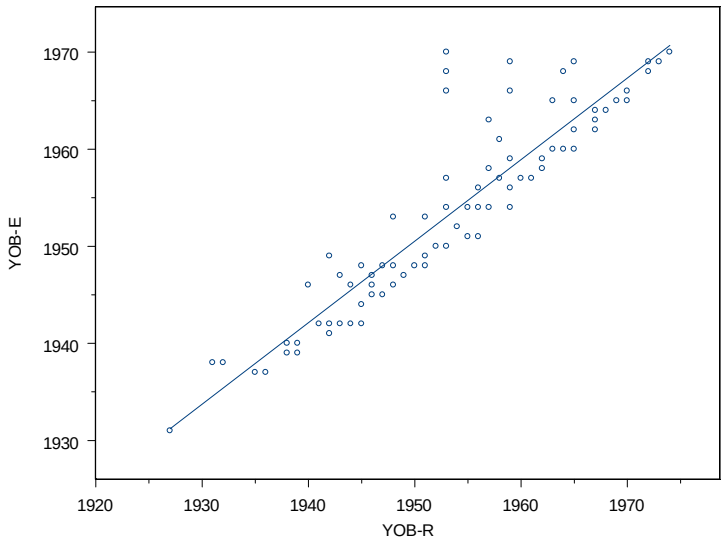


Figure 14: Relationship between the real year of birth (YOB-R) and the estimated year of birth (YOB-E) using leave-one-out.

4.3 Inference of postal code

Geographical information is often critical in care settings as well as in research. Many databases with PHI will contain multiple postal codes. For example, the postal code for the place of work and the residence of an individual, or the postal code of a patient's home and their family physician's practice. The former occurs when a record pertains to an individual, and the latter when the record pertains to a transaction (e.g., a prescription record).

One may remove one of these postal codes in the belief that personal information has been removed. For instance, as in the second example above, one may remove the patient's postal code but keep the physician's postal code, believing that the address information for the patient has been removed. The question we examine in this section is whether it is possible to predict one postal code from another ?

Consider the scenario where, for a particular profession we have the postal code for an individual's home and work. If there is a regularity in how far people live from their work for that profession, then we should be able to predict their home postal code from their work postal code. If such predictions can be done accurately, then removal or non-existence of the home postal code in a database would not be an impediment to predicting it from other geographical information in the record (such as the work address).

In this study we evaluate how accurate it would be to predict one postal code from another in Ontario, Nova Scotia, and Alberta.

4.3.1 Methods

To compute the distance between any two postal codes, we calculate the Euclidean distance in kilometers from the centroid of the postal codes. The centroids are defined by their longitude and latitude.

Let us assume that for any record we have a postal code x and we wish to predict another postal code y. For any postal code x, we consider a circle of radius Q kilometers around its centroid. If the distance between x and any other postal code is less than Q then we consider that to be within Q. We can consider different values of Q. To predict y we would have to select one of the postal codes within Q at random. For example, if our value of Q is 0.1km and within that distance from x there are 20 other postal codes, and assuming that y is within that 0.1km, our best guess will have a chance of 0.05 of selecting y (which is the correct selection).

The conservative assumption we make is that y is always within Q, and compute the probability of a correct prediction on that basis. A random sample of 5000 urban postal codes were analyzed and all rural postal codes were analyzed. For each of the 5000 postal codes we determined the number of other postal codes within Q for different values of Q. In the case of rural postal codes we computed the number of other rural postal codes at various distances for all rural postal codes (since there are fewer rural postal codes it was not necessary to take a sample).

4.3.2 Results

The results for the urban postal codes are shown in Figure 15 for Ontario, Figure 16 for Nova Scotia, and Figure 17 for Alberta.

For urban postal codes the probability of selecting the right postal code if it was within 100m ranges from 0.2 to 0.33 (median) across the three provinces. However, the variation is quite large at such a short distance. This probability drops down quite rapidly as we move away, ranging from 0.04 to 0.05 at 300m, and around 0.02 at 500m for all three provinces. The variation also drops significantly as the distance increases. Beyond 300m it is very unlikely that one would be able to select the correct postal code. But even for predicting postal codes within 100m, the probability of getting it right is quite low, and there. will be high uncertainty in that prediction.

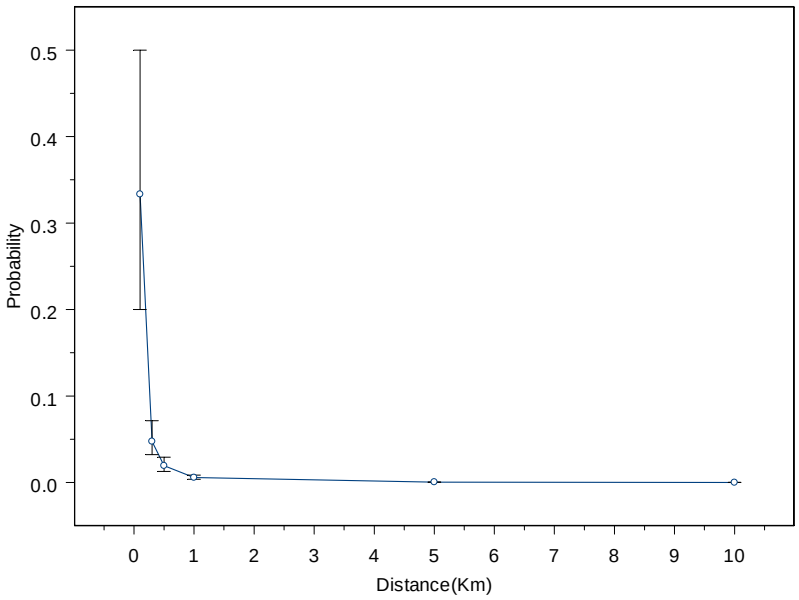


Figure 15: Distance between any two urban postal codes in Ontario (based on a simple random sample of 5000 urban postal codes) vs the probability of correctly guessing one from the other. The graph shows the median and the interquartile range for various distances. The distances are as follows 0.1km, 0.3km, 0.5km, 1km, 5km, and 10km.

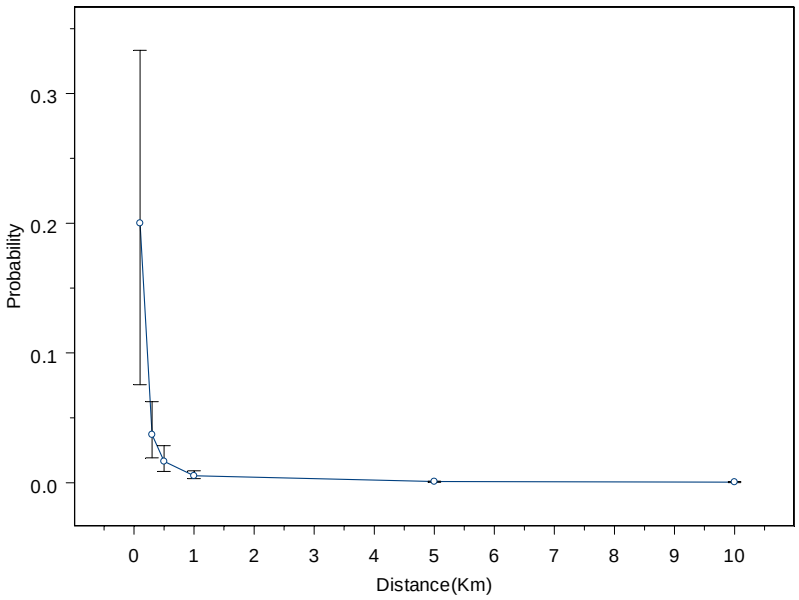


Figure 16: Distance between any two urban postal codes in Nova Scotia (based on a simple random sample of 5000 urban postal codes) vs the probability of correctly guessing one from the other. The graph shows the median and the interquartile range for various distances. The distances are as follows 0.1km, 0.3km, 0.5km, 1km, 5km, and 10km.

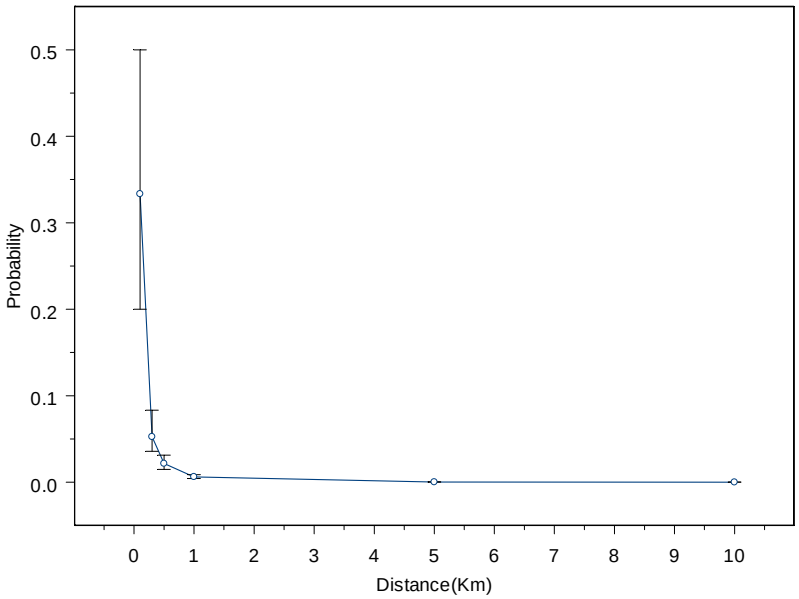


Figure 17: Distance between any two urban postal codes in Alberta (based on a simple random sample of 5000 urban postal codes) vs the probability of correctly guessing one from the other. The graph shows the median and the interquartile range for various distances. The distances are as follows 0.1km, 0.3km, 0.5km, 1km, 5km, and 10km.

The graphs for rural postal codes are shown in Figure 18 for Ontario, Figure 19 for Nova Scotia, and Figure 20 for Alberta. The probabilities for a successful prediction under this assumption are different across the provinces. In Ontario the probabilities are quite high up to 5km, and remain relatively high at 10km. In Nova Scotia the ability to predict accurately starts to drop off after 1km. In Alberta the ability to predict postal codes remains high up to 10km and relatively high at 15km.

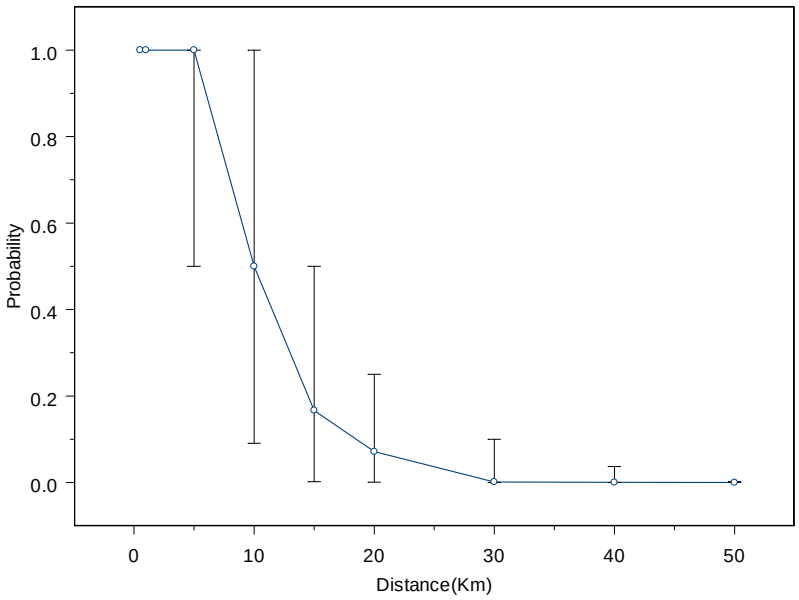


Figure 18: Distance between any two rural postal codes in Ontario vs the probability of correctly guessing one from the other. The graph shows the median and the interquartile range for the following distances: 0.5km, 1km, 5km, 10km, 15km, 20km, 30km, 40km, and 50km.

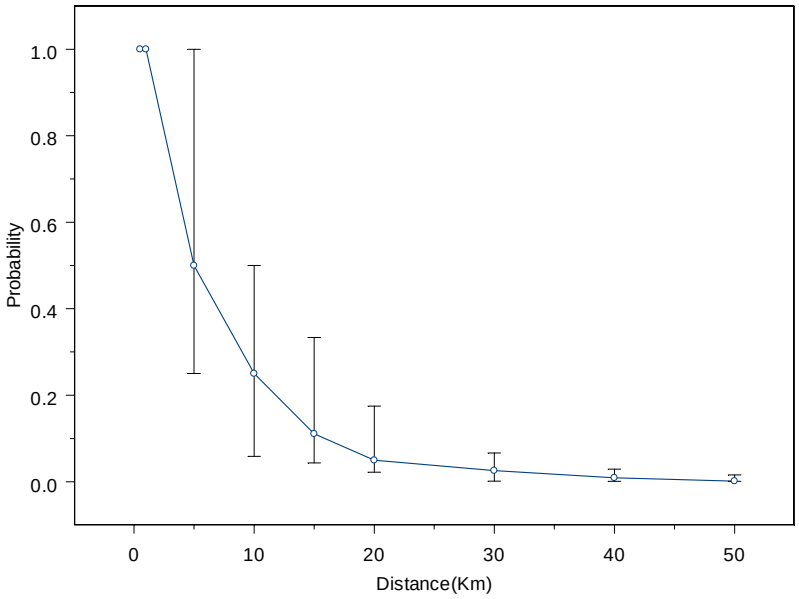


Figure 19: Distance between any two rural postal codes in Nova Scotia vs the probability of correctly guessing one from the other. The graph shows the median and the interquartile range for the following distances: 0.5km, 1km, 5km, 10km, 15km, 20km, 30km, 40km, and 50km.

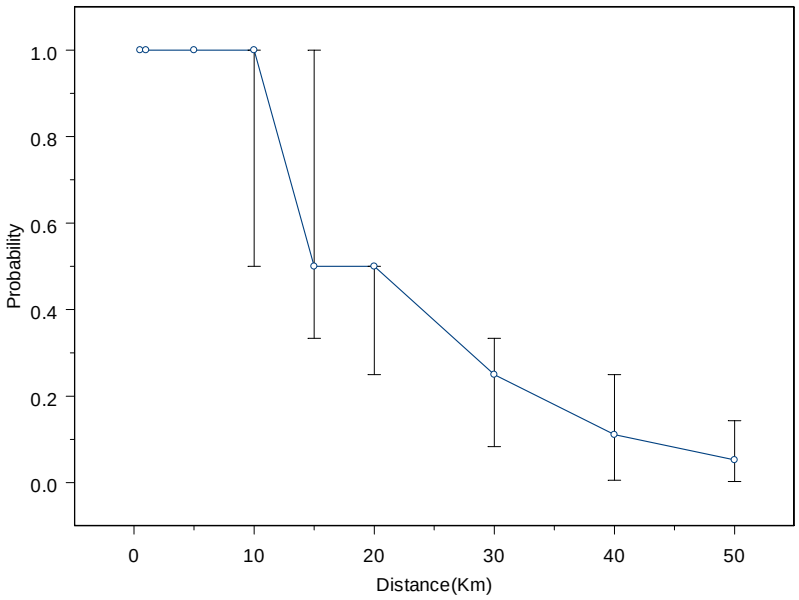


Figure 20: Distance between any two rural postal codes in Alberta vs the probability of correctly guessing one from the other. The graph shows the median and the interquartile range for the following distances: 0.5km, 1km, 5km, 10km, 15km, 20km, 30km, 40km, and 50km.

The implications of our findings differ depending on the province and whether one is referring to rural or urban postal codes. Predicting urban postal codes from other postal codes will not result in accurate results. This applies even more strongly as the distance increases. For rural postal codes one would be able to predict accurately for short distances only in Nova Scotia, but would be able to do so up to 5km-10km in Ontario and 10km-15km in Alberta.

We can consider a specific example. In Figure 21 we can see the distribution of distances between the work postal code and home postal code for a random subset of physicians in the CPSO database. The work postal code was obtained from CPSO and the home postal code was obtained from the identification database described in Chapter 3. In this dataset all of the physicians but one were in urban postal codes. Just under 20% of the physicians worked within 100m from their home. For the majority the work-home distances were larger. Therefore, we can conclude that the prediction of home postal code from work postal code would not be fruitful as it would be quite incorrect for most physicians. The overall probability of a randomly selected physician working within 100m from where s/he lives and being able to predict their home postal code from their work postal code accurately is approximately 0.06 (for urban postal codes). This number goes down rapidly as one increases the distance from 100m.

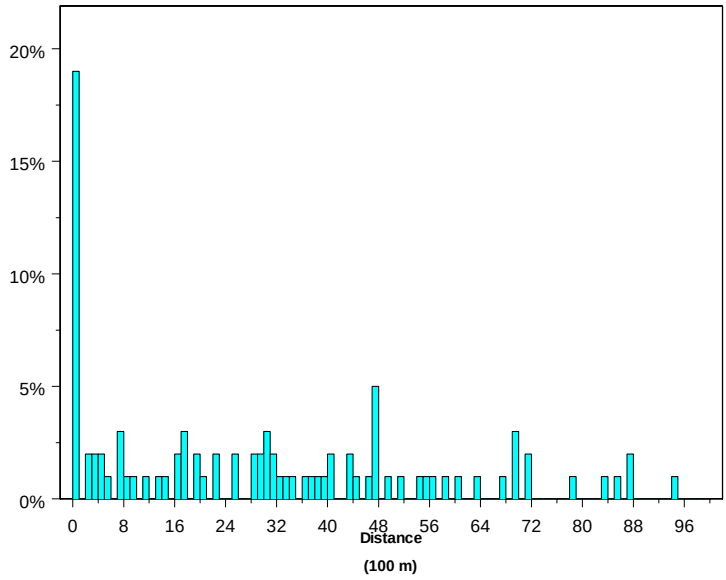


Figure 21: The actual distance from work to home (measured as distance from the centroids of the postal codes) for a subset of physicians listed in the CPSO database (n=130; only the first 10km shown in this graph).

One limitation of this analysis is that we used the centroids of postal codes, which does introduce errors. Also, we used the Canada Post and the Canadian Medical Association definitions of a rural postal code, which is only one of multiple possible definitions [95].

4.4 Discussion

In this chapter we have shown that other variables in D and s can be used to predict important quasi-identifiers relatively accurately. Specifically, we could accurately predict the gender of individuals from their first name and predict their year of birth (and hence age) from their graduation year. However, we found that it would not be possible to predict postal codes accurately from other postal codes for urban pairs unless they are known to be quite close (around 100m), but even then the uncertainty will be quite high. Prediction is quite likely to be accurate for rural pairs at short distances, but the exact cut-off will depend on the province. These results suggest that inference attacks can fill in the gaps for some types of variables and one has to analyze carefully these possibilities before deciding on the utility of an identification database for a re-identification attack.

5 Measuring the Risk of Re-identification

Now that we have seen the different ways in which identification database information can be constructed, one may be prompted to ask, “So what?”. What are the actual risks of re-identification once we are able to build these databases? In this chapter we describe a study that attempted to answer this question.

5.1 Methods

The list of quasi-identifiers that were evaluated in this study are shown in Table 9.

Date of Birth (DoB)	Forward Sortation Area
DoB – month & year	City
Year of Birth	Region
Gender	Initials
Postal Code	

Table 9: List of nine quasi-identifiers evaluated in this study.

The measure of the risk of re-identification that we use is grounded in the matching process that an attacker would likely use in order to re-identify an anonymized data set. Our measure of re-identification risk assumes that an attacker is attempting to re-identify *all* of the individuals in the research database. The attacker is doing so by matching the individuals in s with records in D on the quasi-identifiers. We predict the probability that a randomly selected individual can be matched successfully.

The estimation method we use is Data Intrusion Simulation (DIS) [96, 97]. DIS predicts the conditional probability that a unique match of a record in D with a record in s is a correct match: $P\left(correct\,match \mid unique\,match\right) = P\left(cm \mid um\right)$.

It should be noted that we do not actually need a complete s database nor a complete D database to estimate re-identification risk. All that is needed is a sample identification database, as shown in Figure 22, containing *only* the quasi-identifiers and identifying variables. No actual clinical or lab data is required to perform the risk analysis.

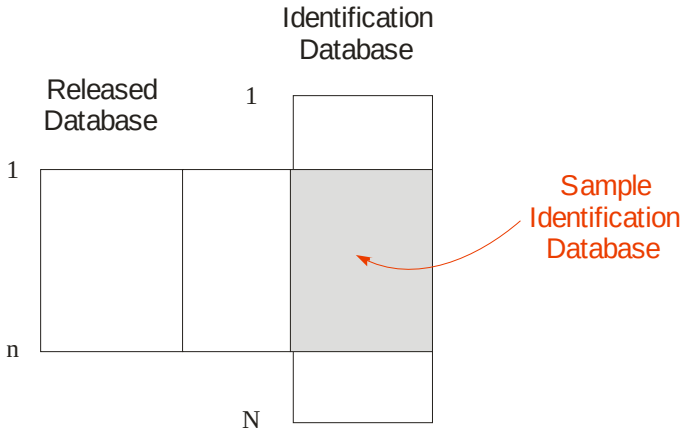


Figure 22: A sample identification database (shown shaded) is all that is needed for Data Intrusion Simulation.

A simulation described in Appendix B illustrates the robust performance of DIS under a range of sampling fractions. Other measures of re-identification risk that have been proposed do not produce accurate results for small sampling fractions, and are not specific to a type of attack [98, 99].

Although there are no generally accepted re-identification thresholds, one can easily make the case that any probability of a successful attack greater than 0.01 would be unacceptable (for a large database, a probability of successful attack as high as 0.01 would compromise the privacy of a relatively large number of individuals). We will therefore use that as a threshold for interpreting the risk results.

In our evaluation, three parameters were varied: (1) the data set, (2) the sample size, and (3) the quasi-identifier combinations evaluated. These are defined as follows:

- We constructed two identification databases to see whether the risk findings carry across them (the CPSO and LSUC data sets described in Chapter 3).
- For each combination of quasi-identifiers we decrement the sample size by one observation chosen at random from n to 30, and determine whether $P(cm|um)$ is below the threshold at the reduced sample size. This process was iterated one hundred times for each sample size and the average number of times that the risk was below the threshold was taken as the result for that sample size. If the risk is below the threshold then we consider the quasi-identifier combination as “safe” (i.e., one that ensures low re-identification risk quite often). We then look at the frequency of quasi-identifiers that are

considered “safe” across all sample sizes. If a quasi-identifier is “safe” more than 50% of the time then it ensures that the risk is below the threshold across sample sizes.

- We considered all possible individual and 2, 3 and 4 –fold combinations of the different quasi-identifiers.

5.2 Results

In Table 10 we see the results of our analysis. The table shows the percentage of times that a particular combination of quasi-identifiers was found to be “safe” (i.e., below the 0.01 risk threshold) as we varied the sample sizes across the two data sets. In total, 143 quasi-identifier combinations were evaluated. We only show those quasi-identifiers and their combinations that had percentages higher than 50%. If a quasi-identifier is not “safe” at least 50% of the time then we can make the case that it is not stable. This means that if the quasi-identifier combination was above the risk threshold more than 50% of the time, it was therefore sensitive to sample size.

The findings indicate that Gender, Region, and the Year of Birth are all relatively stable across sample sizes and data sets, as well as the combination of Region and Gender. This means that the inclusion of these quasi-identifiers in a released data set does not increase the risk of re-identification.

The Gender and Year of Birth combination was low risk 80% of the time only for the CPSO data set. Consequently we consider it unstable across data sets.

Safe Combinations	**Percentage of Time the Quasi-Identifiers Were Below the Threshold**	
	CPSO	**LSUC**
Gender	100%	100%
Region	93%	65%
DoB – Y	94%	85%
Gender + Region	85%	82%
Gender + DoB – Y	80%	--

Table 10: As sample sizes are varied from 30 to the maximum, this table shows the percentage of times that a quasi-identifier or a combination of quasi-identifiers was considered “safe” more than 50% of the time.

5.3 Discussion

We found that only a small subset of the quasi-identifiers represented a consistently low risk of re-identification across both sample size changes and data set changes. Most quasi-identifiers were not stable. In terms of formulating heuristics for the anonymization of data, the following quasi-identifiers were low risk (out of the set that we evaluated):

- Region alone
- Gender alone
- Year of birth alone
- The combination of Gender and Region

A corollary of this result is that all other quasi-identifiers individually, and all other combinations are not safe.

5.4 Generalization of findings

Our data sets were constructed for an Ontario population. We have investigated the ability to construct similar identification databases in Canada. The two main data sources were the PPSR and telephone directory. There is an online telephone directory for every province. In Appendix C we have listed the PPSR sources for all provinces and territories. These would allow the construction of similar identification databases holding similar types of quasi-identifiers.

5.5 Limitations

In our study we used a particular measure of the risk of re-identification. This measure assumes a particular attack scenario on s and our conclusions are limited to that attack scenario.

We also made the assumption that all individuals in s have the same probability of re-identification. Future work should consider record level re-identification risk. For instance, by knowing which specific records are high risk, they can be targeted for disclosure control actions. This will result in fewer distortions to a data set.

The threshold for high risk that we chose was arbitrary. There are no precedents for defining acceptable risk of re-identification for the release of personal health information, therefore the risk threshold will have to evolve as our understanding of acceptable risk evolves. Furthermore, acceptable risk is not static. Society may grow to accept higher risk in return for specific conveniences or personal benefits. Conversely, acceptable risk may decrease if there is a perception of abuse by custodians, or if there is a sharp rise in medical identity theft.

There may be a profession with its members listed whose distribution of quasi-identifiers has many unique observations (e.g., predominantly of a single sex or very sparsely distributed geographically). In such a case, the “safe” quasi-identifiers identified here may no longer be safe.

Furthermore, our list of “safe” quasi-identifiers may not be so for larger sample sizes. As seen in Appendix B, risk is proportional to sample size. Up to the 1% sampling fraction the “safe” quasi-identifiers we have identified were below the threshold, but that may not be the case for larger sampling fractions.

6 Personal Information on the Web

As we have seen in previous chapters, there exists a wealth of public information that can be used for the construction of an identification database. This information is critical for successful re-identification attacks on anonymized databases. Each record in an identification database contains quasi-identifiers and identifying information. To help formulate recommendations on managing re-identification risk, we need to understand the extent to which Canadians object to having such identifying information about them made publicly available.

We noted earlier that Canadians seem to be quite concerned with the privacy of their personal information. The surveys summarized in Chapter 1 evaluate attitudes and reactions to personally hypothetical scenarios (for example, if an individual objects to their identifiable or de-identified PHI being shared with academic researchers). There is evidence of a considerable gap between people's attitudes and actual behaviors when it pertains to their privacy: consumers will reveal personal information for relatively minor personal gains and conveniences [100, 101]. People just do not behave in the way they say they will. Therefore, we need to find out what personal information Canadians will actually self-disclose in a public forum, and whether that information is useful for the construction of identification databases.

6.1 Methods

The public forum we selected for this study was the public web (i.e., no login accounts or access controls are required to view the information). The behavior we wished to detect was the posting of personal information. It is expected that for individuals to voluntarily disclose personal information, they would need to perceive some gain, convenience or benefit. We therefore selected job seekers. Job seekers would expect that the posting of personal information in their on-line resumes would increase the chances of finding a job, which is a significant gain.

Google searches were conducted for pages from Canada matching the search terms "cv cv.html" and "resume resume.html". Google returned a maximum of 1000 results in each case. A web crawler was used to crawl the links from the search result pages, one level deep, and save the content. Only those pages to which we could connect and from which we could download the content were saved. All of the saved pages from the two searches were manually filtered to determine whether or not they were legitimate CVs. The legitimate CVs were then manually examined to determine whether each contained the following pieces of personal information: 1) Name, 2) Address, 3) Postal Code, 4) Telephone number, and 5) An age indicator (date of birth or year of graduation).

6.2 Results

A total of 677 complete CVs were collected. Of the 677 resumes and vitae collected, job seekers' first and last names were found on 628 of the documents (92.76%). Two resumes were found with a first name only. 160 resumes were found to contain a home address, but only 154 displayed a postal code. In 28 instances (4%), only the city of residence was given; and in 180 cases (26.5%), only a professional address was included. This geographic information, in addition to a name, would be sufficient to complete a search for a full address on an online telephone directory, providing that most people live and work in the same city. A great number of resumes also included an indication of age: a birth date, a year of birth, or a graduation date from a secondary/post-secondary institution. While only 98 (14.5%) resumes contained an actual date and/or year of birth, 507 (74.9%) included a graduation date. This is significant as one's graduation date could be used to calculate, approximately, his/her age and year of birth as shown earlier in this report. Combined then, 605 resumes (89.4%) contained some indication of age. Lastly, 186 (27.5%) of 677 provided a home telephone number. A summary is shown in Figure 23.

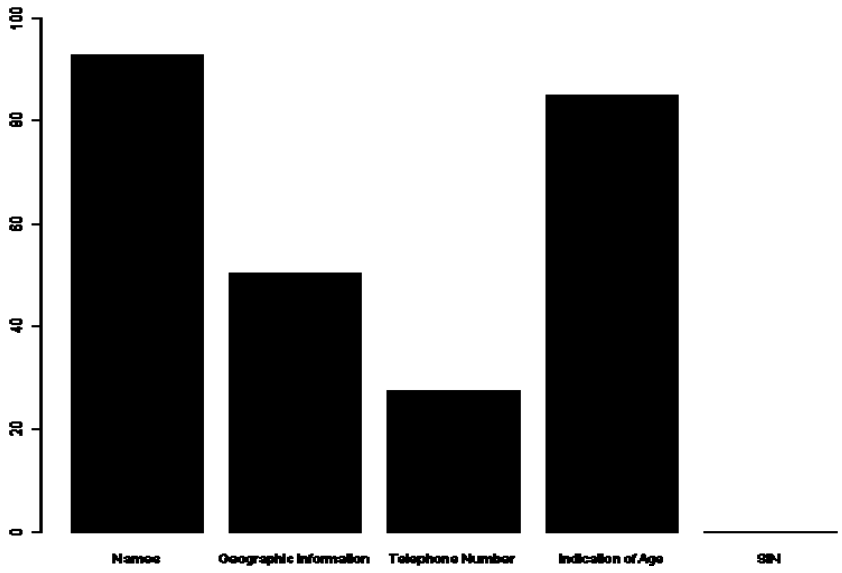


Figure 23: Percentage of CVs containing identifying information and quasi-identifiers for the Canadian data set.

6.3 Discussion

A significant percentage of Canadians are quite willing to post personal identifiers and quasi-identifiers on the public web. The types of variables that they make publicly available, as we have demonstrated earlier in this report, are useful for constructing identification databases that are needed for re-identification.

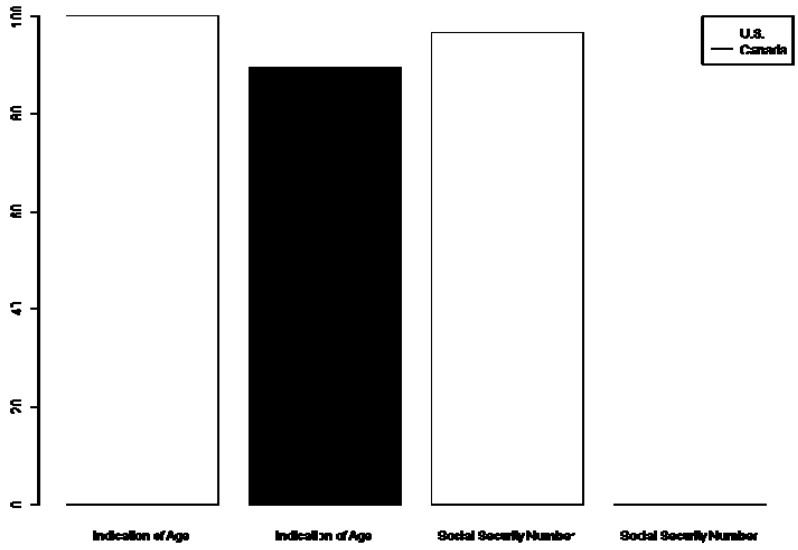


Figure 24: Comparison of results from the US and Canadian studies.

Given the re-identification risks from publicly posting such information, job seekers may be doing so because of [102]: incomplete information (e.g., they do not understand the risks), bounded rationality (even if they know all of the potential risks, they do not have the capacity to perform the time-dependent stochastic analysis to evaluate all of the actual risks), and other psychological distortions.

In a previous study of job seekers conducted in the US, sensitive personal information was sought out online in resumes and CVs posted by job seekers. Four pieces of identifying and quasi-identifying data were sought out —1) Name, 2) Social Security number (SSN), 3) Address and 4) Date of birth—as they are sufficient to obtain a credit card online in the U.S., and can also be used for other forms of identity theft [103, 104]. These results show that the majority of resumes and CVs examined contained these vulnerable pieces of information. In comparison to the U.S. results, Canadians appear to be somewhat more cautious about the types of personal information that they post online. No resumes were found to contain Canadian SIN numbers, while 93% of the resumes in one US study, and all resumes in a second contained SSN [103,

104]. Also, when comparing findings of dates of birth, the US findings are again much higher than the Canadian. One hundred percent of U.S. resumes contained DOB [103, 104], while only 14.5% Canadian resumes provided a DOB and 79.9% gave another indication of age.

6.4 Privacy trade-offs

Job seekers who post their resumes on-line are trading some of their privacy for a possibly greater likelihood of getting a job. That would be considered a significant gain.

Similarly, some of the public registries that we used where identification information is posted (e.g., the PPSR and the Land Registry) also provide significant benefits to individuals when they borrow money and when they wish to buy/sell property. Without similar public registries, such transactions would be more costly and time consuming for individuals. Arguably then, by their actions a large minority of Canadians may be willing to post date of birth information and home postal code information for personal gain.

Some of the public registries that do hold personal information do not bring direct benefit to the individuals in them, but rather bring benefits to businesses or the community as a whole. It is an empirical question whether Canadians will be willing to trade their privacy for a common gain or a benefit to someone else?

6.5 Limitations

Because the US study was conducted a few years prior to ours, and assuming that privacy awareness increases over time, any increase in the public's awareness of privacy issues could confound the comparison of the Canadian findings with the US findings. It is possible then that the US results would be similar our Canadian results if the study was replicated today.

7 Personal Information and Data Remnants

Individuals will often trade elements of their privacy for some personal convenience or gain. Such a trade-off occurs when disposing of their old computer equipment. People may give away old computer equipment to a good cause. These machines may be used for some time and then sold as second hand in the used equipment market. Or individuals may sell their old equipment to resellers directly. Under these scenarios, people either feel that they did something good or they gain financially. Plus, they would then not have to deal with the issue of how to properly dispose of the equipment if they wanted to throw it away or destroy it.

Stolen computers containing PHI may also be laundered through the second hand resale market. Therefore, it is also plausible that some of these machines were not directly or indirectly placed on the second hand market by their owners.

Unfortunately, the disk drives in old computers may contain a considerable amount of personal information. In this study we examine the data remnants in second hand disk drives that are sold in Canada, and determine the extent to which personal information, and PHI specifically, is readily accessible on that equipment. To our knowledge, there have been no such studies performed in Canada, and no studies that have attempted to assess the extent to which PHI can be obtained in that way [105, 106].

7.1 Background

Data recovery is a process by which an individual may retrieve data from a previously deleted data storage medium. This data may be a file, or set of files that have been simply deleted from a system, or an entire volume of data that has been removed from a magnetic disk storage device via the formatting or repartitioning of the entire disk. A general misconception exists amongst the common user of computer systems that once a file has been deleted, the data is forever destroyed. This is not necessarily the case.

In order to properly understand the best methods and practices for data recovery, one must study and understand the fundamental structures of file systems utilized in magnetic disk storage devices. In essence, the structure of a magnetic disk storage device file system can be viewed as being composed of the following major sections (from the beginning to the end of the structure):

Boot Sector

Reserved Sectors

File Allocation Table #1

File Allocation Table # 2

DATA (Containing the Root Directory)

The first section of the file system is the Boot Sector. This contains the executable code that is passed to the central processing unit by the BIOS at boot time after POST has been completed. The Boot Sector also contains information about the physical structure of the disk and handles the data access at boot time. The executable code within the Boot Sector then launches the Operating System contained in the DATA section of the disk.

The File Allocation tables on a disk contain information which track the allocation of sectors as well as the grouping of all the clusters on a disk. As an example, this would have an entry stating that a cluster is located at a specified physical location of a disk. A cluster can also be referred to as an allocation unit, and is the smallest unit of disk space that can be allocated to a single file. Two File Allocation Tables exist in order to provide redundancy on a disk should one fail.

A Root Directory or Root Folder (also referred to as the System Area) is located within the DATA area of a disk and is referenced within the File Allocation Table. This Root Folder contains a record of the mapping of files to clusters on the hard disk. These file clusters are contained in the remaining areas of the magnetic disk.

When a file or folder is deleted from the magnetic disk storage device by the operating system, the first character of the file or directory name is changed to Sigma - ASCII symbol 229 (0xE5) and all pointers for this file are removed from the Root Folder leaving the data in the clusters to be overwritten by the operating system at some later date. The reason why this has been done in operating systems is to increase their speed and efficiency when working with files. When a magnetic disk storage device is formatted, the Root Directory is purged, but the data remains on the drive as per a file deletion. In the case of repartitioning, the File Allocation Tables AND the Root Directory are purged, but all data still remains within the clusters of the disk until it is overwritten.

7.2 Methods

Our objective was to purchase 60 functional second hand disk drives from across the country (we ensured that all non-functional drives were returned and replaced). Due to cost, we could only purchase stand alone drives rather than full second hand computers from which the drives would be physically extracted.

7.2.1 Identifying vendors

Second hand vendors were identified from telephone directories, contacts and experts in the computer industry, Canadian vendors listed on eBay, local business directory searches, and a Google search to find “used computer equipment in Canada”. The results were reviewed to form a list of potential vendors. These vendors were then contacted via telephone and/or email for more information on their inventory. We submitted orders to those vendors who had used hard drives between 10GB and 40GB in their inventory, and who were able to ship this equipment to Ottawa. A few local companies who were contacted requested that the equipment be purchased in person, to which we complied.

7.2.2 Data recovery

The key to recovering data from a magnetic disk storage device is found in the fact that all data remains within the sectors of the device until it is overwritten. Simply put, all one would have to do is parse through the data in each cluster of a DATA section of a magnetic disk storage device in order to reconstruct missing data. In the case of simple file deletion, all one would have to do is parse through the DATA clusters looking for clusters containing the Sigma character. Once a cluster is found which contains a Sigma character, one can inspect the data within the cluster. Information encoded within this data region can be used to recreate the cluster chain (the location of all the parts of the file across the disk). Once a cluster chain is determined, the file can be reconstructed on another storage medium.

If a partition has been deleted, the method is much easier for data recovery. In this scenario, all that one would have to do is iterate through each cluster in order to extract the cluster chain and then reconstruct the files on another storage media. This same method can also be utilized for formatted magnetic disk storage media.

Manually recovering data from a magnetic disk storage device can be a tedious and time consuming ordeal. Fortunately this process can easily be automated. Several commercial as well as freeware software packages have been developed in order to facilitate data recovery. *Recover My Files* is a tool in this category and can recover files that have been deleted in all the aforementioned scenarios. This utility contains an easy to use graphical user interface with a variety of options for file recreation such as burning to a DVD. This is the tool that we used.

7.2.3 Data analysis

All data from the recovered drives were stored on DVDs. A search of the files on each DVD was performed in order to isolate files which may contain personal information. The DVDs were searched for Microsoft Word documents, Excel documents, Power Point documents, Outlook files, and Adobe (PDF) documents. The results of the searches were manually reviewed and a summary of the information found was completed for each disk. This summary also included the serial number of the drive, the name of the vendor from which it was purchased, the date of

purchase, the vendor's statement of the condition of the drive (cleaned or not), and whether the drive contained any inappropriate/obscene material (e.g., pornography). Drives that were flagged as containing inappropriate/obscene material were dealt with through a special protocol.

7.3 Results

The sixty drives were purchased from 12 different vendors across Canada, distributed by province as shown in Table 11. It was much easier to identify vendors who sold stand alone disk drives in Ontario than in any of the other provinces and territories, hence the heavy weighting for that province. Some vendors were non-responsive or did not want to transact with us once they understood the possible implications of the study we were performing.

Province	Number (percent)
Ontario	42 (70%)
Quebec	5 (8.3%)
Alberta	12 (20%)
British Columbia	1 (1.7%)

Table 11: Distribution of drives purchased by province.

The status of the drives once they were received is summarized in Table 12. Repartitioning and formatting are two common approaches for manipulating the drives. In practice, much data can be recovered despite their use. The DOD 5220.22-M standard is a Department of Defense standard providing specifications for clearing and sanitizing electronic data storage devices [107]. There are some commercial and publicly available tools that implement that standard. It is not possible to extract the data from such drives.

Only 5 drives had no action taken to remove the data. For thirty five drives, there was some weak attempt to remove the data. All drives from Alberta were sanitized using the DOD standard, as were five from Ontario.

Province	Repartitioned	Formatted	DoD 5220.22-M	Data Available	Blank
Ontario[IV]	19 (45%)	11 (26%)	5 (12%)	4 (10%)	3 (7%)
Quebec		5 (100%)			
Alberta			12 (100%)		
British Columbia				1 (100%)	

Table 12: The status of received drives distributed by province.

We were able to retrieve data from 39 drives (one of the repartitioned drives from Ontario had no data on it). Five of the drives had pornographic data on them. In total, we extracted 57 DVDs of data from the various drives.

Vendor Statement About Wiping Drives	Count
"Like New Condition"	1
Verbally stated that the drives were formatted	1
"Recertified to factory settings"	1
None	6

Table 13: The claims made by the vendors of the drives from which we were able to extract data.

The nine vendors from whom we bought drives that had data on them did not actually make any claims that the data would be removed in a secure way (see Table 13). It would be up to the individual who provided the drive to enquire further about the vendor's practices.

Province	Owner PI	Owner PHI	Other's PI	Other's PHI
Ontario	26/33	3/33	12/33	6/33
Quebec	5/5	0/5	3/5	0/5
British Columbia	1/1	0/1	0/1	0/1

Table 14: The types of data available on the drives.

[IV] Four of the drives bought from Ontario belonged to US-based entities, two of them were state government departments, one was a municipal department, and one belonged to an individual.

A summary of the type of data that was uncovered in these drives is shown in Table 14. The vast majority of drives with data had personal information about their owners. Examples of personal information found include:

- Personal budgets, salary information, tax returns and completed tax filing forms.
- Personal letters regarding personal relationships.
- Information on life insurance policies, and inheritances.
- Payroll records of employees, including addresses, dates of birth, and Social Insurance Numbers.
- Email correspondence regarding employees and their actions.
- Police record checks.
- Divorce documents.

Very few drives had personal health information. A considerable percentage had personal information about other people apart from the owner. Note that we excluded contact lists and resumes (which are equally likely to be posted on the public web) as personal information on others. Only Ontario drives had personal health information about other people. One of these drives had very sensitive mental health information about a significant number of individuals. Examples of personal health information include:

- Psychological assessments of adults and children, correspondence related to custody cases involving children, affidavits, and social history of abuse victims.
- Medical certificates.
- Letters regarding alcohol addiction of other individuals (not the owners of the drive).
- Reports from an RN about other individuals' health problems, cases of abuse, children's health, and medication lists.
- Correspondence regarding the placement of adults and children in long-term care facilities.

Some of the drives also had confidential corporate information rather than personal information. The leaking of this type of information may prove detrimental to the affected businesses. Examples of confidential data found include:

- Internal policies and procedures, contracts, executive briefs, minutes of meetings for large and small corporations.
- Correspondence related to business deals and corporate financial documents.

- Letterhead templates for corporations and government agencies.

7.4 Discussion

In a previous data remnants study done in the US [106], 158 drives were bought. Of these, 129 were successfully imaged. Approximately 9% were wiped. It was possible to extract data from many of the remaining drives. Our results show that the extent to which data is recoverable is quite dependent on the province. No data was recoverable from drives purchased from Alberta vendors because of the secure delete approach that was used. However, most of the Ontario drives were recoverable. In comparison to international data (see Table 15), the only even marginally comparable jurisdiction to Alberta is the UK where almost half of the drives were wiped clean.

	UK & Australia (2005)	UK (2006)	Australia (2006)	Germany (2006)	North America (2006)
Total Drives	116	200	53	40	24
Faulty Drives	13 (11%)	87 (43%)	3 (6%)	30 (75%)	12 (50%)
Wiped[v]	17 (16%)	55 (49%)	18 (36%)	4 (40%)	1 (8%)
Had Personal Info	51 (49%)	35 (31%)	9 (18%)	3 (30%)	7 (60%)

Table 15: Summary of findings from an international data remnants study [105].

Previous studies did not make the same distinctions we did about the type of data exposed, but in terms of overall exposure of personal information, North America has the highest rates at 60% (see Table 15). We found that almost 79% of the Ontario drives revealed personal information about the owner and 37% revealed personal information about other people. By international standards, these numbers can be considered quite high.

Not a large percentage of drives had owner PHI or PHI about others. This is likely an indication that not much PHI is electronic yet. Most of the PHI that we found was in correspondence rather than in electronic records per se. One would expect that as EMRs become more widely deployed in Canada, more PHI will be available to patients and hence the risk of their disclosure would increase.

There is clearly a need for organizations and individuals, certainly in Ontario and to some extent Quebec, to take actions to reduce the risk of personal data leaking from second hand disk drives. While there are a large number of techniques that an individual and organization can potentially employ to protect personal data [108], a number of them require specialized equipment or resources and are therefore not practical for most users. There are two general approaches that can be pursued: encryption of drive data and secure delete technology.

At the outset, when a user gets a computer they can use encryption technology. Encryption can be used to create specific virtual drives, and all sensitive information can be stored there. Unless the password used is weak or the encryption algorithm is compromised, it would be extremely difficult to extract that information. However, this is generally not enough. Many programs will store their data, temporary files, cached files, backup files and registry values outside the encrypted virtual drive. Quite a significant amount of information can be left in these files. Most users would not know to change the settings of their applications to only use the encrypted drive, and sometimes that option is not available. Therefore, if one really wants to protect their data, this would probably not be the best approach unless they possess a great deal of technical expertise (to change the setting of the applications and force them to use the encrypted drive).

The best encryption technology to use is whole disk encryption that is invoked before the operating system, during system boot. This ensures that all data on the drive (temporary, backup, and data) is encrypted. Fortunately, this type of technology is becoming more generally available in common operating systems and hardware. Therefore, one would expect that, in the next few years it will be much more widely deployed and would eliminate the risks we identified.

The second technology is secure delete. This allows the one to delete all of the data on the drive so that it is not recoverable. Secure delete by itself, however, is not enough. One needs to perform a more general disk wipe. Software for wiping disks usually performs a secure delete as well as removing all of the temporary, backup, and cached files from the system.

A recent study noted that commercial software for wiping disks tends to be quite unreliable [109]. In one case, the software did not even attempt a secure delete because of a software bug. The difficulty of wiping software is that it needs to determined where each application keeps its information. This is difficult to do for a very large number of applications that change often. It has been argued that because the market for privacy tools is small (and hence the vendors have limited resources), such vendors will not be able to keep up to date with the application and operating system changes [109]. Therefore, while the use of wiping software is reassuring, it may not actually be sufficient to protect personal data on disk drives.

[V] As a percentage of those that were not faulty.

8 Recommendations

In this section we summarize some main points and formulate recommendations based on the results of our studies. It should be noted, however, that our focus has been on the re-identification risk from record linkage. There are other types of risks that would not be addressed even if one were to implement all of our recommendations.

8.1 Anonymization process

The overall decision process for anonymizing a data set is shown in Figure 25. The various steps in this process can be informed by the results of the studies described in this report.

The first step is to prepare the data and the context for interpreting it. This involves two things: (1) determine the risk threshold, and (2) perform data cleansing.

The risk threshold is affected by three things: (1) the probability of an attack attempt, (2) the consequences of an attack, and (3) the impact of anonymization on the data. Therefore, it is important to gather information about the context of the data disclosure in order to be able to decide on a threshold. Example information that would be useful would be:

- Who will have access to the data once it is disclosed ?
- Will those who have access to the disclosed data sign any data sharing agreements, are they auditable, are any agreements enforceable in the jurisdictions where they operate, and how much would it cost to enforce these agreements ?
- Who are the likely attackers and what would be their motives for an attack ?
- What is the likelihood of an attack attempt by any of these attackers ?
- Would the perception that the data is re-identifiable cause harm to the organization, even if there was little risk of re-identification ?
- Will multivariate analyses be performed on the data by the end-uses or simple univariate summaries ?

A risk threshold could at the outset be characterized as “low”, “medium-low”, “medium”, “medium-high”, or “high”. This would, in practice, provide sufficient guidance for some of the trade-offs that will be necessary at a later stage.

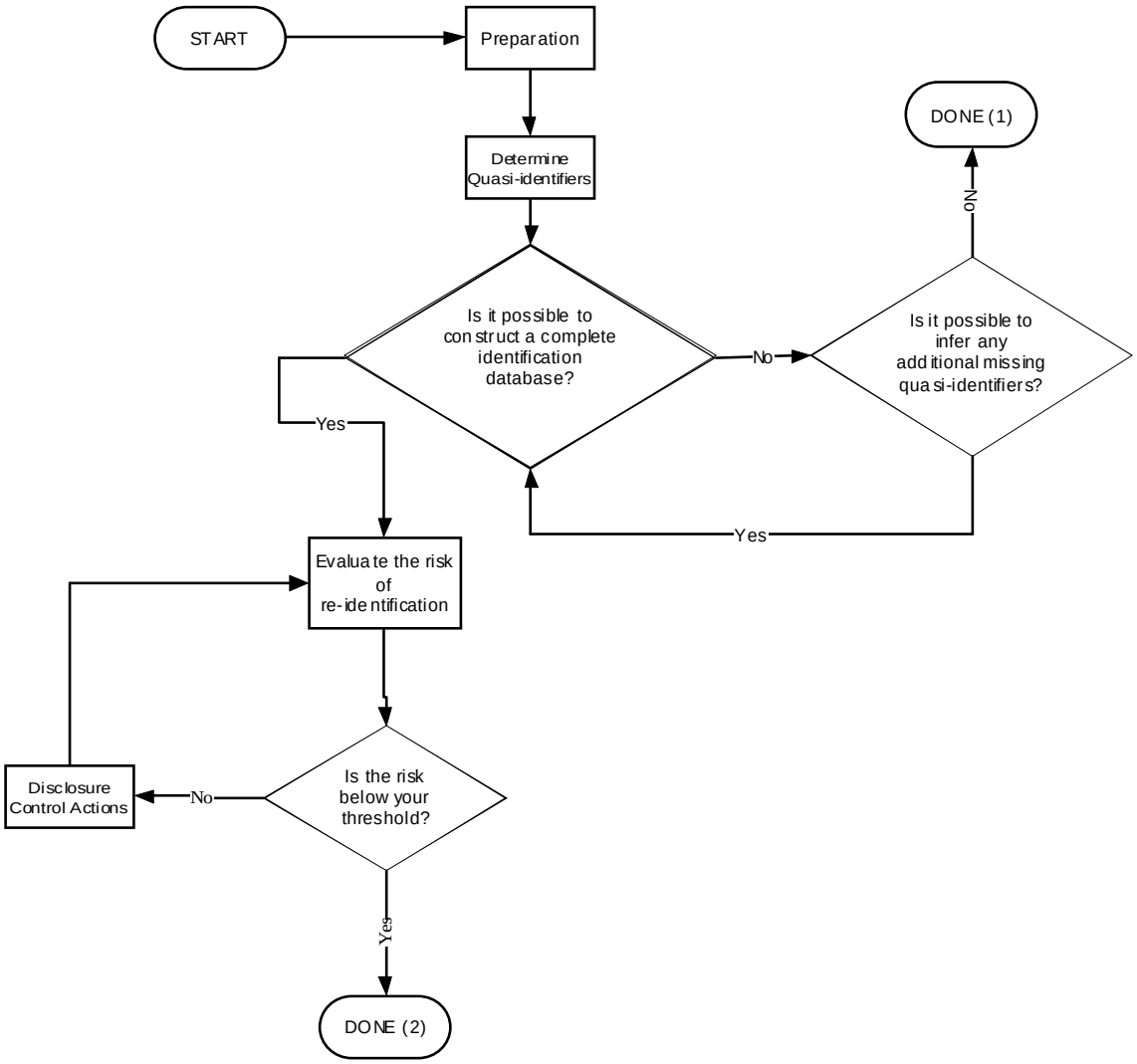


Figure 25: The recommended decision process for anonymizing a data set.

Data cleansing in practice means that A needs to deal with the identifying variables. As discussed in Chapter 2, there are multiple ways to deal with identifying variables: coding, removal, and randomization. Depending on the situation, the analyst must select one of these techniques to reduce the risk of re-identification from identifying variables.

After the completion of data preparation, we should have a data set without identifying variables and an understanding of the risk threshold.

The next step is to determine if there are any quasi-identifiers in the data set. There are some obvious quasi-identifiers, such as: postal code or other geographic information, date of birth, gender, initials, and profession. Any other variable that can conceivably exist in an identification database should be flagged as a quasi-identifier. It is therefore important to also have an

understanding of the different ways in which an identification database can be constructed for s and whether an attacker would be able to construct such a database. In our analysis we looked at public sources and found the PPSR registry, telephone directories, the land registry, the postal code database from Canada Post, and membership listings for professional associations to be good sources of public information to construct an identification database. The data custodian needs to know these sources and if they are relevant to s, otherwise inadvertent disclosures may be made.

There are other sources of information that can make it into an identification database that are privately owned and that can be bought. It would be prudent to examine these as well. Within the budget available for this research program we were not able to do so, but for highly sensitive data it may in some cases be worth the investment to purchase some of these databases to get an understanding of the types of data that they contain and the types of populations that they represent.

One important thing to consider when looking at identification databases or sources of information that can be used to construct identification databases is the accuracy of the data. It is not always the case that these sources have accurate information. For example, the data may be old and no longer reflective of reality, it may be extrapolated or estimated data, or there may be data entry errors. Different data sets may use incompatible coding schemes making any mapping between them uncertain. Even inaccurate data can cause harm. However, for the purposes of re-identification if the data is highly inaccurate then it would pose less of a risk.

Another consideration is the economic deterrent. Even if it is plausible that specific sources can be combined and a comprehensive identification database constructed, it may still not be economically feasible for an attacker. Therefore, the custodian A needs to have an understanding of the possible attackers and their resources to come up with a sound judgment about which identification databases are likely.

Once the possible identification databases that can be constructed have been determined, it is then necessary to see how many of the quasi-identifiers in s exist in these identification databases. It is also necessary to look at possible inference attacks. There may be quasi-identifiers in s that do not exist in the identification databases, but they can be inferred from other variables that exist in the identification database. Similarly, the quasi-identifier set in D may be extended through inference attacks, and this needs to be assessed. In our evaluations we found that year of birth can be predicted from year of graduation, and gender from first name. We also determined that it is highly unlikely that one can accurately predict one postal code from another. Therefore the existence of another postal code in the record will not pause a high risk of inference.

At this stage we should know which actual and predicted quasi-identifiers in s and D can be used for record linkage.

The next step is to perform a baseline assessment of re-identification risk. An initial qualitative assessment based on heuristics is a good start, but this needs to be followed with a more quantitative approach.

Many heuristics for suppressing and generalizing variables treat quasi-identifiers individually. For example, one heuristic may say that you need to generalize the full date of birth to the year of birth. But because there exist other quasi-identifiers in the data set, such a generalization may not be enough to reduce the risk appreciably. In fact, all combinations of quasi-identifiers need to be considered to decide which combination is safe to keep and or generalize. Based on our studies (which considered four quasi-identifiers simultaneously), the following general heuristics can be utilized for risk assessment and risk reduction:

- If s has full postal codes for individual residences then these should be removed or generalized unless the risk threshold is *high*. Postal codes make re-identification quite easy. An initial generalization of a postal code is to the FSA.
- If the profession of a record in s is known and that profession lists its members, then do not release initials, unless the risk threshold is *medium-high* or *high* then the initials would be acceptable.
- If the profession of a record in s is known and that profession lists its members, then do not release FSA and either remove it or generalize it to the region, unless the risk threshold is *medium-high* or *high* then the FSA is acceptable.
- If the profession of a record in s is known and that profession lists its members, then do not release date of birth and either remove it or generalize it to year of birth, unless the risk threshold is *medium-high* or *high* then the date of birth or month of birth are acceptable.
- If the profession of a record in s is known and that profession lists its members, and there are exactly two quasi-identifiers in the data set and they are the following pair {gender and region}, then that is acceptable irrespective of the risk threshold.
- If the profession of a record in s is known and that profession lists its members, and there are exactly two quasi-identifiers in the data set and the risk threshold is *medium-low* or *low*, then one of them should be removed if they are not the following pair {gender and year of birth}, with the exception of the above rule.

- If the profession of a record in s is known and that profession lists its members, and there are more than two quasi-identifiers, then reduce the number of quasi-identifiers to two or less if the risk threshold is *medium*, *medium-low* or *low*, and then apply the above rules.

If there are other quasi-identifiers not covered by the above heuristics then a more formal quantitative risk assessment would be required.

Once baseline risk assessment is complete, if the risk of re-identification is deemed to still be high then additional action needs to be taken to reduce it. This means the application of quantitative disclosure control techniques is required. Quantitative disclosure control techniques include sub-sampling, generalization and suppression. These techniques require access to the actual data. As we see in Figure 25, the application of disclosure control techniques is an iterative process that may require multiple passes to converge to a satisfactory risk level below the threshold.

8.2 General Considerations

Some general issues and recommendations that came out of this research that do not fall specifically within the decision process above are summarized below:

1. **Identification Databases.** One way to manage the risk of re-identification due to record linkage using quasi-identifiers is to reduce the sources of identification databases. In practice, this would be very difficult to do because some of the data sources are necessary for other legitimate reasons and their removal would be quite disruptive for business and society in general. Furthermore, individuals do gain personally from having these registers available and they may not mind to their personal information being available on them (this would be consistent with other privacy trade-offs). Hence Canadians may object if these gains are eliminated. The approach that we have recommended above is that the data custodians develop an understanding of the types of identification databases that can be constructed and incorporate that information in their anonymization process.

2. **Public Education.** It is necessary to educate the public about the types of information that can identify them. This will allow the public to make more informed decisions about the trade-offs they are making when they give up some privacy for a personal convenience or gain.

3. **Drive Encryption.** Organizations should adopt drive encryption technology that is readily available to protect the personal information and PHI that they hold. Individuals should accelerate the adoption of more modern hardware and operating systems that provide a seamless implementation of this type of technology. Failing the ability to encrypt the disk

drives, then the drives should be properly erased before resale or destroyed when they are no longer being used.

8.3 Future Research

Our results have also identified some opportunities for additional research. These will provide more information that can be used within the decision process described above.

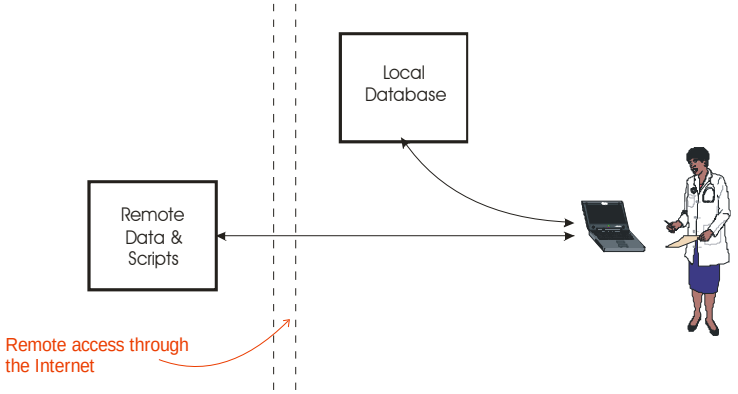
1. **Trade-offs.** We need to develop a better understanding of the trade-offs that Canadians are willing to make with their quasi-identifiers. Quasi-identifiers allow re-identification when they appear in identification databases. In our research we found evidence of broad self-disclosure of identifying and quasi-identifying variables in exchange for potential personal gains. Since that is the case, then maybe Canadians do not mind if their quasi-identifiers are stored in public registries, despite the privacy risk, if there are personal gains to them ? But it is also important to study societal gains in addition to personal gains. For example, will Canadians be willing for their information to be disclosed if that will eventually result in a benefit to our community or society at-large (but not necessarily of direct benefit to them personally) ?

2. **Inference Attacks.** More research needs to be performed to identify other possible inference attacks. We believe we have only scratched the surface with the three evaluations reported upon here. A better understanding of such risks will help to directly inform the anonymization decision process described above.

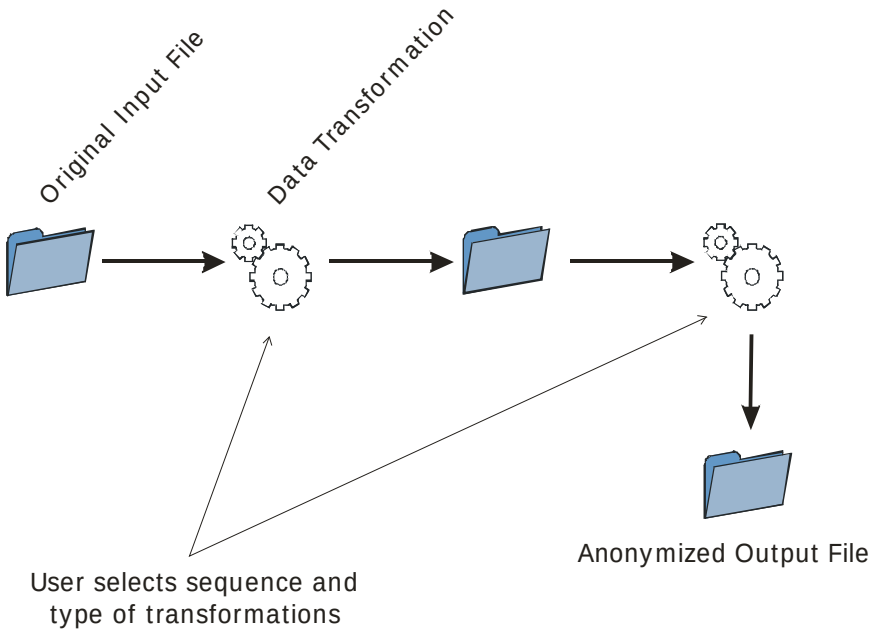
3. **Risk-based Anonymization.** While they are a good start, the use of simple heuristics and rules of thumb to anonymize data may result either in one being over-cautious or not being cautious enough – it is difficult to tell without doing a quantitative risk assessment on the data set itself. This is particularly true when multiple quasi-identifiers are involved. The risk of re-identification is also dependent on the sample size, which is not captured by simple heuristics. Therefore, additional research needs to be done on optimal quantitative techniques for risk-based anonymization.

9 Appendix A: The PrivacyAnalytics Tool

The PrivacyAnalytics tool is a desktop application that reads data and performs sequential data transformations on it to anonymize it. The basic architecture of PrivacyAnalytics makes it a regular desktop application and a client of web services. A local database is used to store and process the data sets that are being anonymized, and a remote connection provides additional generic data that are used by some of the tool's functions. Some of the analysis algorithms are also loaded from a remote site (at http://www.ehealthinformation.ca/). No data that is undergoing anonymization ever travels the internet.



The basic concept behind PrivacyAnalytics is that it reads in a source file and then transforms it repeatedly according to user instructions. The transformations available in the current version of the tool randomize the identifying variables.



The tool also provides capabilities to evaluate quantitatively the risk of re-identification for a particular data set.

10 Appendix B: DIS Simulation

In this appendix we report on a simulation to demonstrate and evaluate the characteristics of Data Intrusion Simulation. We used data on the 23,506 physicians listed by the College of Physicians and Surgeons of Ontario as our population. We created random samples of various sizes from that population.

For the simulation we drew samples varying in size from 100 to 3000 individuals. A series of three quasi-identifiers were evaluated individually : gender, work postal code, and the work forward sortation area. Each sample size was drawn 1000 times and in each case the estimate of the probability of successful re-identification was calculated from the sample. Since we have the population data set as well, it was possible to compute the actual probability using the population data set. The bias of the predicted probability is computed by comparing it with the actual probability.

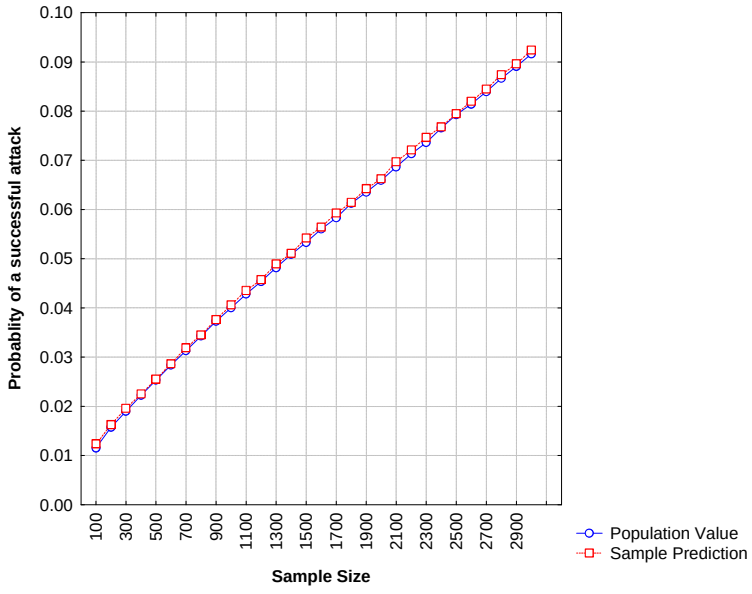


Figure 26: The probability of a successful re-identification attack (y-axis) for various sample sizes (x-axis) based on a Monte Carlo simulation. The results are for the forward sortation area of the work postal code. The graph shows the predicted and actual values based on the sample and the population respectively. These were averages across 1000 iterations for each sample size.

Figure 26 shows the results for the work forward sortation area. It can be seen that the bias is quite small for the full range of sampling fractions studied. The highest sampling fraction was just under 13% and the smallest sampling fraction was 0.04%. The magnitude of bias ranges from 0.00015 to 0.0016. For our purposes this bias is quite small and indicates that the predicted probability is quite robust even for small sampling fractions. Similar results were obtained for the other quasi-identifiers examined.

It should also be noted that the probability of re-identification increases with sampling fraction. This is consistent with the common recommendation made in the disclosure control community to minimize the size of the sample that is released because that has a lower risk of re-identification. Therefore, one approach to reduce the risk of re-identification is to release a smaller data set.

11 Appendix C: Personal Property Security Registries

The following is the list of locations to obtain PPSR information across Canada for the construction of identification databases.

Province	**URL**
British Columbia	https://www.bconline.gov.bc.ca
Alberta	Available from authorized registry agents
Saskatchewan	http://www.isc.ca
Manitoba	https://direct.gov.mb.ca/ppr/
Ontario	https://www.personalproperty.gov.on.ca/ppsrweb/en/enquiry/cc_enquiry.jsp
Quebec	http://si2.rdprm.gouv.qc.ca/index.asp
New Brunswick	https://www.web11.snb.ca/snb7001/e/2000/2700e_6.asp
Nova Scotia	http://www.acol.ca/Services/PPR/NS/menu.html
Prince Edward Island	http://www.acol.ca/Services/PPR/PE/menu.html
Newfoundland and Labrador	http://www.acol.ca/Services/PPR/NF/menu.html
Northwest Territories	http://www.acol.ca/Services/PPR/NT/menu.html
Nunavut	http://www.acol.ca/Services/PPR/NU/menu.html

12 Acronyms

API	Application Programming Interface
BIOS	Basic Input/Output System
CIHR	Canadian Institutes of Health Research
CPSO	College of Physicians and Surgeons of Ontario
CSIS	Canadian Security Intelligence Service
CV	Curriculum Vitae
DIS	Data Intrusion Simulation
DND	Department of National Defence
DoB	Date of Birth
DOD	Department of Defense
EMR	Electronic Medical Record
FIPPA	Freedom of Information and Protection of Privacy Act
FOIP	Freedom of Information and Privacy (Office)
FSA	Forward Sortation Area
GEDS	Government Electronic Directory Service
IRB	Institutional Research Board
IPC	Information and Privacy Commissioner (of Ontario)
IQR	Inter-Quartile Range
LSUC	Law Society of Upper Canada
OECD	Organization for Economic Co-operation and Development
PHI	Personal Health Information
POST	Power-On Self Test
PPSR	Private Property and Security Registry
REB	Research Ethics Board
RMSE	Root Mean Square Error
RN	Registered Nurse
SIN	Social Insurance Number
SSN	Social Security Number (US)
YOB	Year of Birth

13 Acknowledgements

This program of research was funded by the following organizations over the last 12 months: The Office of the Privacy Commissioner of Canada, Ontario's Ministry of Research and Innovation, the Canadian Foundation for Innovation, Bell Canada, and the Ontario Research Network for Electronic Commerce. Some initial work related to this program of research was funded through a Student Internship Program grant from Ontario Centers of Excellence. We are very grateful for these organizations' support of our work.

The following people have contributed to our work during the project: Sam Jabbouri, Michael Power, and Youenn Drouet. Also, Rebecca Johnston, Mary Lysyk and Joan Roch reviewed earlier drafts of this document. We would like to thank all of them for their time and effort.

The program of research described in this report has been approved by the Research Ethics Board of the Children's Hospital of Eastern Ontario Research Institute.

14 References

[1] Irving R. 2002 Report on Information Technology in Canadian Hospitals. Canadian Healthcare Technology 2003.

[2] HIMSS. Healthcare CIO Results. Healthcare Information and Management Systems Society Foundation 2004; February.

[3] Andrews J, Pearce K, Sydney C, Ireson C, Love M. Current State of Information Technology Use in a US Primary Care Practice-based Research Network. *Informatics in Primary Care* 2004;12:11-18.

[4] Bower A. The diffusion and value of healthcare information technology. RAND Health 2005.

[5] Fonkych K, Taylor R. The state and pattern of health information technology adoption. RAND Health 2005.

[6] HarrisInterractive. Health information privacy (HIPAA) notices have improved public's confidence that their medical information is being handled properly. Available at: http://www.harrisinteractive.com/news/allnewsbydate.asp?NewsID=894. Accessed 4th April, 2005.

[7] California Health Care Foundation. Medical privacy and confidentiality survey 1999.

[8] Grimes-Gruczka T, Gratzer C, The Institute for the Future. Ethics survey of consumer attitudes about health web sites. California Health Care Foundation 2000.

[9] Willison D, Kashavjee K, Nair K, Goldsmith C, Holbrook A. Patients' consent preferences for research uses of information in electronic medical records: Interview and survey data. *British Medical Journal* 2003;326:373.

[10] Mitchell E, Sullivan F. A descriptive feast but an evaluative famine: Systematic review of published articles on primary care computing during 1980-97. *British Medical Journal* 2001;322:279-282.

[11] Dixon P. Medical Identity Theft: The Information Crime that Can Kill You. The World Privacy Forum, 2006. Available at: http://worldprivacyforum.org/pdf/wpf_medicalidtheft2006.pdf. Archived at: http://www.webcitation.org/5OTRVDqAe.

[12] van Heusden P. Applying software validation techniques to Bioperl. *Bioinformatics Open Source Conference*, 2004. Available at: http://www.open-bio.org/bosc2004/presentations/Heusden_SW_validation_BOSC_2004.pdf. Archived at: http://www.webcitation.org/5NslvhFRA.

[13] Powell J, Buchan I. Electronic health records should support clinical research. *Journal of Medical Internet Research* 2005;7(1):e4.

[14] El Emam K, Jabbouri S, Sams S, Drouet Y, Power M. Evaluating common de-identification heuristics for personal health information. *Journal of Medical Internet Research*, 2006; 8(4):e28. Available at: http://www.jmir.org/2006/4/e28/. Archived at: http://www.webcitation.org/5OTRan1nS.

[15] Brand R. Overseas antidote: medical services are moving offshore, raising privacy issues. *Rocky Mountain News*. May 21, 2005.

[16] Report HI-040001-1: A hospital in a rural centre. Office of the Privacy Commissioner of Ontario, 2005. Available at: http://www.ipc.on.ca/images/Findings/up-HI_040001_1.pdf. Archived at: http://www.webcitation.org/5OEjzUUW9.

[17] Report HI-050007-1: A Private laboratory. Office of the Privacy Commissioner of Ontario, 2004. Available at: http://www.ipc.on.ca/images/Findings/up-HI_050007_1.pdf. Archived at: http://www.webcitation.org/5OEkCVKtR.

[18] Report HI-050011-1: A community care access centre. Office of the Privacy Commissioner of Ontario, 2004. Available at: http://www.ipc.on.ca/images/Findings/up-HI_050011_1final.pdf. Archived at: http://www.webcitation.org/5OEkSR7no.

[19] Report: HI-050015-1: A nursing services company. Office of the Privacy Commissioner of Ontario, 2004. Available at: http://www.ipc.on.ca/images/Findings/up-HI_050015_1.MW.pdf. Archived at: http://www.webcitation.org/5OEksXyg1.

[20] Report HI-050022-1: A Community Care Access Centre. Office of the Privacy Commissioner of Ontario, 2006. Available at: http://www.ipc.on.ca/images/Findings/up-HI_050022_1.MW.pdf Archived at: http://www.webcitation.org/5OF8kcKpY.

[21] Report HI-050021-1: An Audiology Clinic. Office of the Privacy Commissioner of Ontario, 2006. Available at: http://www.ipc.on.ca/images/Findings/up-HI_050021_1.MW.pdf Archived at: http://www.webcitation.org/5OF9DvVbu.

[22] Report HI-050031-1: A Hospital in an Urban Setting. Office of the Privacy Commissioner of Ontario, 2006. Available at: http://www.ipc.on.ca/images/Findings/up-HI_050031_1_final_report.pdf Archived at: http://www.webcitation.org/5OFAad8cN.

[23] Report HI-050019-1: A Municipality's Public Health Unit. Office of the Privacy Commissioner of Ontario, 2006. Available at: http://www.ipc.on.ca/images/Findings/up-HI_050019_1.pdf Archived at: http://www.webcitation.org/5OFAnwyZT.

[24] Report HI-050016-1: A City Hospital. Office of the Privacy Commissioner of Ontario, 2006. Available at: http://www.ipc.on.ca/images/Findings/up-HI_050016_1.pdf Archived at: http://www.webcitation.org/5OFBoGGWO.

[25] Report HI-050004-1: A Public Laboratory. Office of the Privacy Commissioner of Ontario, 2005. Available at: http://www.ipc.on.ca/images/Findings/up-HI_050004_1.pdf Archived at: http://www.webcitation.org/5OFQ8ezN4.

[26] Report HI-050039-1: A Designated Rehabilitation Assessment Centre Office of the Privacy Commissioner of Ontario, 2006. Available at: http://www.ipc.on.ca/images/Findings/up-HI_050039_1.MW.pdf. Archived at: http://www.webcitation.org/5OFXmB1Ip.

[27] Order HO-001. Office of the Privacy Commissioner of Ontario, 2005. Available at: http://www.ipc.on.ca/images/Findings/up-ho_001.pdf Archived at: http://www.webcitation.org/5OFYQWkf3.

[28] Order HO-002. Office of the Privacy Commissioner of Ontario, 2006. Available at: http://www.ipc.on.ca/images/Findings/up-HO_002.pdf. Archived at: http://www.webcitation.org/5OFZz2FXq.

[29] Order HO-003. Office of the Privacy Commissioner of Ontario, 2006. Available at: http://www.ipc.on.ca/images/Findings/up-ho_003.pdf. Archived at: http://www.webcitation.org/5OI9nKfYU.

[30] Report HI-050044-1: A Psychologist Working for a School Board. Office of the Privacy Commissioner of Ontario, 2006. Available at: http://www.ipc.on.ca/images/Findings/up-HI_050044_1_Report.pdf. Archived at: http://www.webcitation.org/5OF7qHLuW

[31] Report HI-050047-1: A Physiotherapy and Rehabilitation Centre. Office of the Privacy Commissioner of Ontario, 2006. Available at: http://www.ipc.on.ca/images/Findings/up-I_050047_1.pdf. Archived at: http://www.webcitation.org/5OF67opQF.

[32] Robeznieks A. Privacy fear factor arises. *Modern Healthcare* 2005;35(46):6.

[33] Mandl K, Szolovits P, Kohane I. Public standards and patients' control: How to keep electronic medical records accessible but private. *British Medical Journal* 2001;322:283-286.

[34] Cheng T, Savageau J, Sattler J, DeWitt A, Thomas G. Confidentiality in health care: A survey of knowledge, perceptions, and attitudes among high school students. *Journal of the American Medical Association* 1993;269(11):1404-1408.

[35] New poll: Doctors lie to protect patient privacy. Association of American Physicians and Surgeons, 2001; 2006(24th July). Available at: http://www.aapsonline.org/press/nrnewpoll.htm. Archived at: http://www.webcitation.org/5NzvR0MX2.

[36] Rethinking the information highway. EKOS 2003.

[37] Saravamuttoo M. Privacy: Changing attitudes in a tumultuous time. Sixth Annual Privacy and Security Workshop, 2005.

[38] Gostin L, Turek-Brezina J, Powers M, Kozloff R. Privacy and security of health information in the emerging health care system. *Health Matrix: Journal of Law-Medicine* 1995;5(1):1-36.

[39] Hodge J, Gostin L, Jacobson P. Legal issues concerning electronic health information. *Journal of the American Medical Association* 1999;282(13):1466-1471.

[40] Woodward B. The computer-based patient record and confidentiality. *New England Journal of Medicine* 1995;333(21):1419-1422.

[41] Goldman J. Testimony before the subcommittee on health of the committee on ways and means on "Patient Confidentiality" 1998.

[42] Leape L, Bates D, Cullen D, Cooper J, Demonaco H, Gallivan T, Hallisey R, Ives J, Laird N, Laffel G, Nemeskal R, Petersen L, Porter K, Servi D, Shea B, Small S, Sweitzer B, Thompson T, Viet M. Systems Analysis of Adverse Drug Events. *Journal of the American Medical Association* 1995;274(1):35-43.

[43] Ash J, Berg M, Coiera E. Some Unintended Consequences of Information Technology in Health Care: The Nature of Patient Care Information System-related Errors. *Journal of the American Medical Informatics Association* 2004;11(104-112).

[44] Johnson N, Mant D, Jones L, Randall T. Use of Computerised General Practice Data for Population Surveillance: Comparative Study of Influenza Data. *British Medical Journal* 1991;302:763-765.

[45] Wilton R, Pennisi A. Evaluating the Accuracy of Transcribed Computer-stored Immunization Data. *Pediatrics* 1994;94:902-906.

[46] Davidson B, Lee Y, Wang R. Developing Data Production Maps: Meeting Patient Discharge Data Submission Requirements. *International Journal of Healthcare Technology and Management* 2004;6(2):223-240.

[47] Garg A, Curtis J, Halper H. Quantifying the impact of IT security breaches. *Information Management & Computer Security* 2003;11(2):74-83.

[48] Campbell K, Lawrence G, Loeb M, Zou L. The economic cost of publicly announced information security breaches. *Journal of Computer Security* 2003;11:431-448.

[49] Approaches to security breach notification: A white paper. Canadian Internet Policy and Public Interest Clinic 2007.

[50] Lenard T, Rubin P. An economic analysis of notification requirements for data security breaches. The Progress and Freedom Foundation 2005.

[51] Becker C, Taylor M. Technical difficulties: Recent health IT security breaches are unlikely to improve the public's perception about the safety of personal data *Modern Healthcare* 2006;38(8):6-7.

[52] Ponemon L. National survey on data security breach notification. White & Case LLP; conducted by the Ponemon Institute LLC 2005.

[53] Organisation for Economic Co-operation and Development. Science, Technology and Innovation for the 21st Century 2004.

[54] Organisation for Economic Co-operation and Development. Promoting Access to Public Research Data for Scientific, Economic, and Social Development: OECD Follow Up Group on Issues of Access to Publicly Funded Research Data 2003.

[55] Fienberg S, Martin M, Straf M. Sharing Research Data. Committee on National Statistics, National Research Council 1985.

[56] Hutchon D. Infopoints: Publishing raw data and real time statistical analysis on e-journals. *British Medical Journal* 2001;322(3):530.

[57] Are journals doing enough to prevent fraudulent publication ? *Canadian Medical Association Journal* 2006;174(4):431.

[58] Draft Policy on Access to CIHR-funded Research Outputs. Canadian Institutes of Health Research. Available at: http://www.cihr-irsc.gc.ca/e/32326.html. Archived at: http://www.webcitation.org/5OFJXMExj.

[59] Melton III L. The threat to medical-records research. *New England Journal of Medicine* 1997;337(13):1466-1470.

[60] Woolf S, Rothemich S, R J, Marsland D. Selection bias from requiring patients to give consent to examine data for health services research. *Archives of Family Medicine* 2000;9:1111-1118.

[61] McKinney P, Jones S, Parslow R, Davey N, Darowski M, Chaudry B, Stack C, Parry G, Daper E. A feasibility study of signed consent for the collection of patient identifiable information for a national pediatric clinic audit database. *British Medical Journal* 2005;doi:10.1136/bmj.38404.650208.AE.

[62] Tu J, Willison D, Silver F, Fang J, Richards J, Laupacis A, Kapral M. Impracticability of informed consent in the registry of the Canadian Stroke Network. *New England Journal of Medicine* 2004;350(14):1414-1421.

[63] Armstrong D, Kline-Rogers E, Jani S, Goldman E, Fang J, Mukherjee D, Nallamothu B, Eagle K. Potential impact of the HIPAA privacy rule on data collection in a registry of patients with acute coronary syndrome. *Archives of Internal Medicine* 2005;165:1125-1129.

[64] 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. Centre for Health Services and Policy Research and Canadian Policy Research Networks 2005.

[65] Jacobsen S, Xia Z, Campion M, Darby C, Plevak M, Seltman K, Melton L. Potential effect of authorization bias on medical records research. *Mayo Clinic Proceedings* 1999;74(4):330-338.

[66] Nelson K, Rosa E, Brown J, Manglone C, Louis T, Keeler E. Do patient consent procedures affect participation rates in health services research ? *Medical Care* 2002;40(4):283-288.

[67] Al-Shahi R, Vousden C, Warlow C. Bias from requiring explicit consent from all participants in observational research: prospective, population based study. *British Medical Journal* 2005;331:942-.

[68] Junghans C, Feder G, Hemingway H, Timmis A, Jones M. Recruiting patients to medical research: Double blind randomised trial of "opt-in" versus "opt-out" strategies. *British Medical Journal* 2005;331(940-).

[69] Ward H, Cousens S, Smith-Bathgate B, Leitch M, Everington D, Will R, Smith P. Obstacles to conducting epidemiological research in the UK general population. *British Medical Journal* 2004;329:277-279.

[70] McCarthy D, Shatin D, Drinkard C, Kleinman J, Gardner J. What is the effect of state legislation requiring patient consent for use of medical records in research ? Research Findings 5(1): Center for Health Care Policy and Evaluation 1999.

[71] Personal data for public good: Using health information in medical research. Academy of Medical Sciences 2006.

[72] Gordis L, Gold E. Privacy, confidentiality, and the use of medical records in research. *Science* 1980;207(11):153-156.

[73] Order HO-004. Office of the Privacy Commissioner of Ontario, 2007. Available at: http://www.ipc.on.ca/images/Findings/up-3ho_004.pdf. Archived at: http://www.webcitation.org/5OFOzaj1O.

[74] El Emam K. Data Anonymization Practices in Clinical Research: A Descriptive Study. May 2006; Health Canada, Access to Information and Privacy Division, 2006. Available at: http://www.ehealthinformation.ca/documents/HealthCanadaAnonymizationReport.pdf. Archived at: http://www.webcitation.org/5OTRpWIJC.

[75] Sweeney L. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems* 2002;10(5):557-570.

[76] Ochoa S, Rasmussen J, Robson C, Salib M. Reidentification of individuals in Chicago's homicide database: A technical and legal study. Massachusetts Institute of Technology 2001.

[77] Willison D. Academic REBs and governance of privacy, confidentiality and security in database research. First Workshop on Electronic Health Information and Privacy, 2005.

[78] Nair K, Willison D, Holbrook A, Keshavjee K. Patients' consent preferences regarding the use of their health information for research purposes: A qualitative study. *Journal of Health Services Research & Policy* 2004;9(1):22-27.

[79] Power A, Pullman D. Sorry, you can't have that information: Stakeholder awareness, perceptions and concerns regarding the disclosure and use of personal health information, in *Electronic Health Information and Privacy Conference*. Ottawa, 2006.

[80] OIPC Stakeholder Survey, 2003: Highlights Report. GPC Research 2003.

[81] Saxena N, MacKinnon M, Watling J, Willison D, Swinton M. Understanding Canadians' Attitudes and Expectations: Citizens' Dialogue on Privacy and the use of Personal Information for Health research in Canada. Canadian Policy Research Networks 2006.

[82] Lysyk M. Electronic Health Information and Privacy: What Canadians Think - A Review of Public Opinion research 2001-2005, in *Electronic Health Information and Privacy Conference*. Ottawa, 2006.

[83] Lambert D. Measures of disclosure risk and harm. *Journal of Official Statistics* 1993;9(2):313-331.

[84] Marsh C, Skinner C, Arber S, Penhale B, Openshaw S, Hobcraft J, Lievesley D, Walford N. The case for samples of anonymized records from the 1991 census. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* 1991;154(2):305-340.

[85] Elliot M, Dale A. Scenarios of attack: the data intruders perspective on statistical disclosure risk. *Netherlands Official Statistics* 1999;14(Spring):6-10.

[86] Lysyk M, El Emam K, Lucock C, Power M, Willison D. Privacy Guidelines Workshop Report. CHEO Research Institute and the University of Ottawa 2005.

[87] El Emam K. Overview of Factors Affecting the Risk of Re-identification in Canada. Access to Information and Privacy Division, Health Canada, 2006. Available at: http://www.ehealthinformation.ca/documents/HealthCanadaReidReport.pdf. Archived at: http://www.webcitation.org/5OTRtsfUw.

[88] Knoppers B, Saginur M. The Babel of genetic terminology. *Nature Biotechnology* 2005;23(8):925-927.

[89] El Emam K, Sams S. Anonymization case study 1: Randomizing names and addresses. Electronic Health Information Laboratory, 2007. Available at: http://www.ehealthinformation.ca/documents/PACaseStudy-1.pdf. Archived at: http://www.webcitation.org/5OT8Y1eKp.

[90] Blien U, WIrth H, Muller M. Disclosure risk for microdata stemming from official statistics. *Statistica Neerlandica* 1992;46(1):69-82.

[91] Elliot M. Disclosure Risk Assessment. In: Doyle P, Lane J, Theeuwes J, Zayatz L, editors. *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*: Elsevier, 2001.

[92] A report on the Canadian data brokerage industry. Canadian Internet Policy and Public Interest Clinic 2006.

[93] GEDS FAQ. Available at: http://direct.srv.gc.ca/cgi-bin/direct500/TE?FN=faq.htm.

[94] Public sector employment, wages and salaries, by province and territory Available at: http://www40.statcan.ca/l01/cst01/govt62a.htm.

[95] Pong R, Pitblado J. Don't take geography for granted ! Some methodological issues in measuring geographic distribution of physicians. *Canadian Journal of Rural Medicine* 2001;6(2):103-112.

[96] Elliot M. A new approach to the measurement of statistical disclosure risk. *International Journal of Risk Management* 2000;2(4):39-48.

[97] Skinner G, Elliot M. A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society (Series B)* 2002;64(Part 4):855-867.

[98] Chen G, Keller-McNulty S. Estimation of identification disclosure risk in microdata. *Journal of Official Statistics* 1998;14(1):79-95.

[99] Zayatz L. Estimation of the percent of unique population elements on a microdata file using the sample. US Bureau of the Census 1991.

[100] Spiekermann S, Grossklags J, Berendt B. E-privacy in 2nd generation e-commerce: Privacy preferences versus actual behavior. 3rd ACM Conference on Electronic Commerce, 2001; 38-47.

[101] Hann I-H, Hui K-L, Lee T, Png I. Online information privacy: Measuring the cost-benefit trade-off. 23rd International Conference on Information Systems, 2002.

[102] Acquisti A. Privacy in electronic commerce and the economics of immediate gratification. ACM Conference on Electronic Commerce, 2004; 21-29.

[103] Sweeney L. AI technologies to defeat identity theft vulnerabilities. AAAI Spring Symposium on AI Technologies for Homeland Security, 2005.

[104] Sweeney L. Protecting job seekers from identity theft. *IEEE Internet Computing* 2006:74-78.

[105] Jones A, Valli C, Sutherland I, Thomas P. The 2006 analysis of information remaining on disks offered for sale on the second hand market. *Journal of Digital Forensics, Security, and Law* 2006;1(3):23-36.

[106] Garfinkel S, Shilat A. Rememberance of data passed: A study of disk sanitization practices. *IEEE Security and Privacy* 2003:17-27.

[107] DoD 5220.22-M: National Industrial Security Program Operating Manual (NISPOM). Department of Defense, 2006. Available at: http://download.ehealthinformation.ca/Cite/522022m.pdf. Archived at: http://www.webcitation.org/5OIWVOeq4.

[108] Clearing and declassifying electronic data storage devices. Communications Security Establishment, 2006. Available at: http://www.cse-cst.gc.ca/documents/publications/gov-pubs/itsg/itsg06.pdf. Archived at: http://www.webcitation.org/5OIbcPcnJ.

[109] Geiger M, Cranor L. Scrubbing stubborn data: An evaluation of counter-forensic privacy tools. *IEEE Security and Privacy* 2006;4(5):16-25.