Overview of Factors Affecting the Risk of Re-identification in Canada

Khaled El Emam (PhD)

University of Ottawa
CHEO RI, 401 Smyth Road
Ottawa, Ontario K1H 8L1

May 8, 2006

This report was written for the Access to Information and Privacy Division of Health Canada.

Table of Contents

Introduction

The objectives of this report are twofold. First, to present a summary of the most important factors that would have an impact on the risk of re-identification when releasing so-called anonymized data in Canada, and second, provide some guidance on how to manage that risk. The primary audience is the set of individuals who would be responsible for deciding what data to release. Secondary audiences who would find this report potentially of value include: data analysts to get a better understanding of data protection issues, members of research ethics boards to help assess the privacy risks in protocols, and administrators to formulate policies concerning the release and handling of data collected by their organizations.

We start off by providing some definitions to make clear what the nature of this risk is. This is followed by a review of the factors. The report concludes with a set of recommendations, which constitute good practices for minimizing this risk. The report deliberately eschews the mathematics for evaluating risk, and attempts to convey the principles and concepts through examples.

Definitions

1. Data Release Context

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[*].

The most common ways in which data can be released is in raw data format (also known as *microdata*), or in tabular format[†]. Our focus here will be on the release of microdata.

The released microdata, s, is considered to be a sample from a larger 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 [‡].

The nature of how s was collected, the number of individuals in s, and the sampling fraction of s are not going to be critical here.

A must ensure that the risk of re-identification of the individuals in the released data s is exceedingly small. The exact threshold used to for deciding that the risk of re-identification is small will not be of concern at this point.

The remainder of this report will focus on the factors that would have an impact on the re-identification risk for the individuals in s, and will conclude with some measures that can be taken by A for reducing this risk.

2. Variable Types

When evaluating the risk of re-identification it is useful to classify the variables in s in terms of that risk. There are three classes of variables:

> **Identifying variables.** These are variables that can directly identify individuals, such as 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 de-identified. 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, combined with a telephone number then the individual can be easily identified.

[*] 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.

[†] 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).

[‡] 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 will not exist.

Quasi-identifiers. These are variables that do not directly identify an individual, but can play an important role in indirect re-identification. There are two ways in which quasi-identifiers can be used for re-identification: (a) by linking to external databases containing identifying variables (record linkage), and (b) if the values on these variables are unique.

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.

Since identifying variables make re-identification trivial, we will not consider these. Non-identifying variables play no role in re-identification, and therefore we will not consider these either. The key set of variables for re-identification are the quasi-identifiers.

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.

3. The Attacker

Our analysis of the risk of re-identification assumes that there potentially exists an attacker who does get access to s and attempts to re-identify the individuals in that data set. The attacker may be engaging in a completely legitimate commercial activity – therefore the term *attacker* should not be construed to mean that something inappropriate is being performed.

4. Types of Attack

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

a) 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 geographical information in the record to determine which town or city that individual lives in, and re-identify them.

b) The attacker may have a list of known individuals, but is not sure if these individuals are in s. S/he wishes to re-identify them in s if they exist there because the records in s contain additional information about these individuals. For example, a spouse in a divorce case may wish to determine financial or health information by re-identifying their adversary in a publicly released database.

c) The attacker may know that a specific individual participated in the study that generated the data in s. S/he would then attempt to re-identify that individual in the data set using characteristics s/he knows about that individual. For example, a parent may look for their child's record knowing that they participated in a sexual behavior survey.

d) The attacker wishes to re-identify all individuals in s. For example, the attacker may be a marketing company that will customize the advertising materials to be sent based on age and gender.

5. Re-identification Risk Exposure

The objective of A is to manage the overall risk of re-identification in order to focus resources on data releases that are potentially the most problematic. The overall risk of re-identification in s is a function of three parameters: (a) the probability of an attacker attempting to re-identify individuals in s, (b) the probability of successful re-identification of individuals in s, and (c) the consequences of successful re-identification. The overall risk is termed the *risk exposure*.

The three parameters are not independent of each other. For example, if the probability of successful re-identification is high then that provides an incentive for an attacker to try. An attacker is not likely to attempt an attack if it is known that the probability of success is exceedingly small.

If the probability of re-identification is high and the consequences of re-identification are severe, but the party E who receives the data is trustworthy, has signed a data sharing agreement with A, and agrees to be audited on a regular basis, then one can consider the risk exposure to be relatively low because the likelihood of an attempt is very small. However, if the data s is being released for public use then the probability of an attack attempt is much higher and hence the risk exposure would rise.

If the probability of re-identification is very low but the consequences of re-identifying any individual in s can be personally devastating for the re-identified individual and/or damaging to A, then the risk exposure is high. On the other hand, if the probability of re-identification is very high but the consequences for all involved are trivial, then risk exposure would be low.

The extent of effort and actions taken to reduce the risk of re-identification should be consistent with risk exposure to ensure that resources are put into activities that will provide the most privacy protection.

Factors Affecting Risk Exposure

In this section we will present a list of factors that would have a substantial impact on risk exposure. One of these factors is uniqueness. Since uniqueness is a large topic it is addressed in a sub-section of its own, and this is followed by a review of the more general factors.

1. Uniqueness

An individual is said to be unique in a data set if the values on certain combinations of quasi-identifiers are unique. If the data set is s then the individual is *sample unique*. If the data set is U then the individual is *population unique*. The greater the number of quasi-identifiers in a data set the greater the probability that an individual will be unique in the data set.

If an individual is population unique then they are by definition sample unique. If an individual is sample unique they *may be* population unique. However, if an individual is not sample unique then by definition they are also not population unique.

In general, uniqueness increases the probability of re-identification on two fronts:

- Uniqueness makes it easier to re-identify individuals in s through record linkage with an external database.
- Unique individuals in s are more vulnerable to being recognized.

We will discuss these two scenarios further below.

1.1 Record Linkage

An attacker can attempt to match records in s with records from an *identification database*. The matching would be made on the quasi-identifiers.

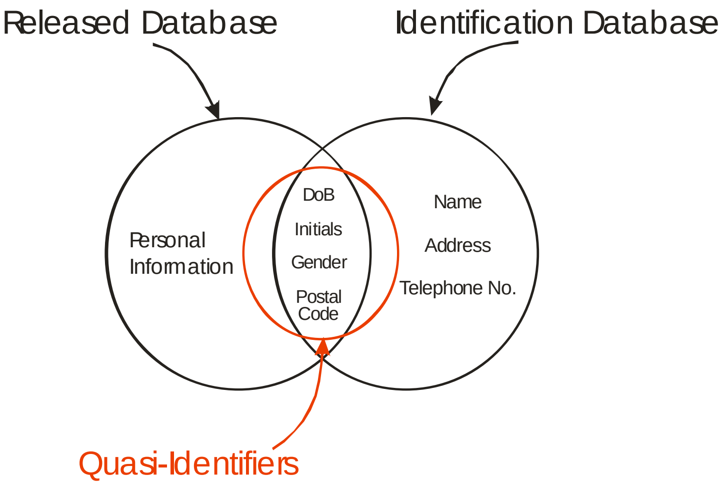


Figure 1: Illustration of how the released database can be linked with an identification database.

Figure 1 shows how this 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 this kind of record linkage to re-identify individuals conducted by Sweeney in the US. 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.

In Canada there is less data that is available for the construction of an identification database. However, it can still be done. An identification database can be constructed in a number of ways:

- publicly available information from government bodies and professional associations (this is discussed further in the appendix),
- 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.

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.

Thus far we have kept the definition of the population vague. That was deliberate because the population will depend on the nature of the data set being released. However, if the population is say all residents of Ontario or all Canadians, then in practice it will be very difficult for an attacker to get hold of an identification database that is equivalent to a population database.

It will also be quite common in practice that not all of the individuals in s are in D. But it will not be possible to find out who these individuals who 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.

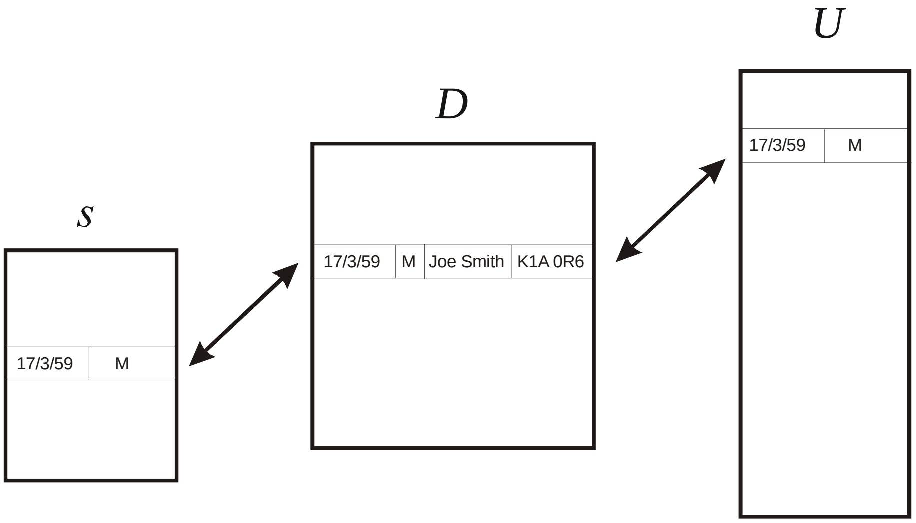


Figure 2: *Example of record linkage through an identification database where an individual is population unique. The two quasi-identifiers in this example are date of birth and gender. The identification database also has the individual's name and home postal code. We assume that the individuals in* D *are a subset of the individuals in* U *and that some of the individuals in* s *will not be in* D*. Because* s *and* D *do not have exactly the same individuals, it is not possible to re-identify some individuals in* s *.*

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 and uniqueness . If an individual in s is population unique and that individual exists in D, then it is almost certain that that individual will be re-identified using record linkage. This is illustrated in Figure 2. Because the individual is population unique then by definition that individual would be unique in both D and s . Population uniqueness makes the verification that the match is correct trivial.

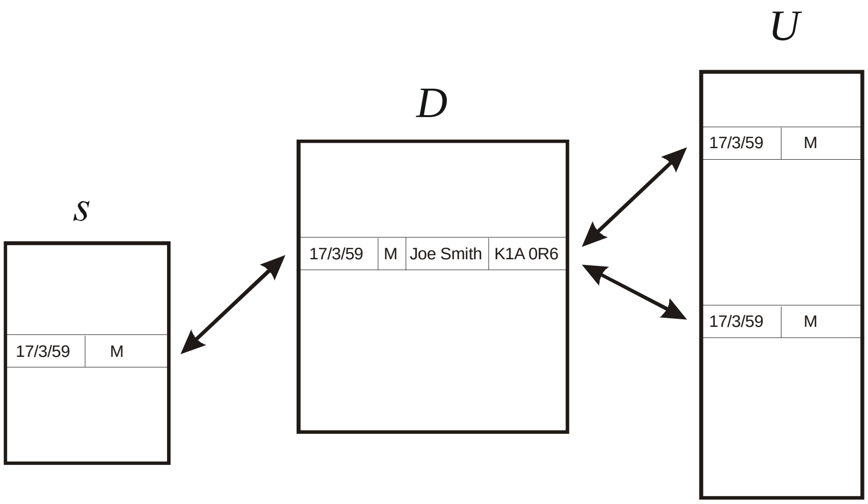


Figure 3: *A re-identification scenario where the individual is not population unique.*

Another scenario is shown in Figure 3 whereby the individual is sample unique but not population unique. Here the attacker will match the records in D and s, 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).

In summary then the probability of a correct match is dependent on:

- whether the individual is in the identification database,
- whether the individual is population unique, and
- the ability to verify that the match is correct through other sources.

Minimizing the number of quasi-identifiers in s will make it less likely that a record is population unique. Also, since verification is a time consuming exercise, the lack of population uniqueness would increase the burden of correct re-identification.

1.2 Inference of Quasi-Identifiers

Even if the records in s do not have sufficient information for successful record linkage because some quasi-identifiers have been removed, a sophisticated attacker may attempt to infer the values for the missing quasi-identifiers:

- Through knowledge of statistical relationships between variables in s, the attacker may be able to estimate other quasi-identifiers for individuals in s and hence increase their probability of re-identification. For example, by measuring the distance between the postal code of a university campus and neighboring postal codes, one can estimate the age of individuals by proximity of residence to the university. Individuals with postal codes closest to the university are likely to have ages in the late teens and early twenties.

- The attacker may use direct inference of removed quasi-identifiers. For example, if the record does not include gender but does include laboratory results, then gender can be inferred from the medical tests that have been performed or from the results of the medical tests (for example, prostate-related tests or pregnancy tests).

The inference of missing quasi-identifiers makes the individual more likely to be re-identified.

1.3 Individual Uniqueness

Not all individuals in s will have the same risk of re-identification. For example, an individual who is unique when considering only two quasi-identifiers is at a much higher risk of re-identification than an individual who is unique on say six quasi-identifiers. There are two reasons for this:

1. Everything else being equal, constructing an identification database with two quasi-identifies is likely to be easier than constructing one with six quasi-identifiers. Therefore, the former individual is more likely to be re-identifiable through record linkage.
2. Many data sets will have measurement and coding errors. The former individual is much more susceptible to being re-identified when there are errors in the data than the latter individual because fewer variables important for re-identification (the quasi-identifiers) are affected by the error.

Furthermore, individuals unique on many different *combinations* of quasi-identifiers will be at a higher risk. For example, an individual who is unique on three combinations: (i) date of birth and gender, (ii) date of birth and initials, and (iii) date of birth and home postal code, will be more vulnerable than an individual unique on only one combination, such as date of birth and gender.

1.4 Sample Uniqueness

If an individual is sample unique then they are at risk of re-identification through record linkage because this means that they are potentially also population unique. Sample uniqueness can be problematic in other ways:

- Identifying unique professions, such as the mayor of a town where the town is known (either because s only covers the town or geographical information is included as part of the records in s). Similar risk exists with the only dentist or family physician in a particular geographical region.
- Outliers can identify individuals in the community that are easy to trace. For example, individuals with exceptionally high income, those with a particularly high number of children, or those that are very old.
- Values on variables that are sample unique and are unusual given societal norms. For example, a 16 year old widow or a 19 year old PhD graduate would be unusual.

Traceability of such sample unique individuals increases the more precise the geographical information in the record (see Section 2.9).

1.5 Geographical Variables

The most commonly used geographical variable is the six character postal code. Geographical information makes an individual more easily traceable and could also increase their uniqueness. In this section we will provide an overview of postal codes and illustrate how they have an impact on the risk of re-identification.

The postal code system is an administrative construct designed for the sole purpose of making mail delivery efficient. The first three characters of the postal code refer r the Forward Sortation Area (FSA). The last three characters are known as the Local Delivery Unit (LDU). There are many LDUs in each FSA. The number of FSAs and LDUs in Canada is summarized in Table 1.

Province/Territory	# LDUs	# FSAs
Alberta	76,756	150
British Columbia	112,738	188
Manitoba	23,928	64
New Brunswick	57,286	110
Newfoundland	10,382	35
North Western Territories	506	3
Nova Scotia	25,290	76
Nunavut	29	3
Ontario	268,839	521
Prince Edward Island	3,148	7
Quebec	202,284	413
Saskatchewan	21,514	48
Yukon	933	3

Table 1: *The number of Local Delivery Units and Forward Sortation Areas in each province and territory as of April 2006.*

The average number of households covered by an LDU is approximately 15, but can range from zero to 7000 households[§]. This large variation occurs because some LDUs contain only businesses and no households.

[§] Statistics Canada: *2001 Census Dictionary*. Catalogue No. 92-378-XIE.

If a record in s has an FSA field then that individual will be part of a bigger population than if the record has an LDU field. Being part of a larger population means that the chance of being unique is smaller. Therefore, in general it is better to use FSAs rather than LDUs when including geographic information in a record.

The size of the Canadian population in each FSA by province and territory is shown in Figure 4. It is clear that there is wide variation and the median number of individuals per FSA is very different across the provinces and territories. For example, the median in Ontario is 20,207 and in Quebec it is 16,138. The largest FSA is in Quebec with more than 132,000 individuals (“J0K”).

In general, it is easier to be unique in a smaller FSA than a larger one. The median population size per FSA in New Brunswick is approximately 4,400 individuals, which is the smallest median across the provinces. Therefore, if s has an FSA variable, then New Brunswick residents are more likely to be unique than Ontario residents, on average. If we add a gender variable then we can expect to halve that number. Therefore, the FSA and gender can narrow an individual down to a subset of just over 2000 individuals in New Brunswick.

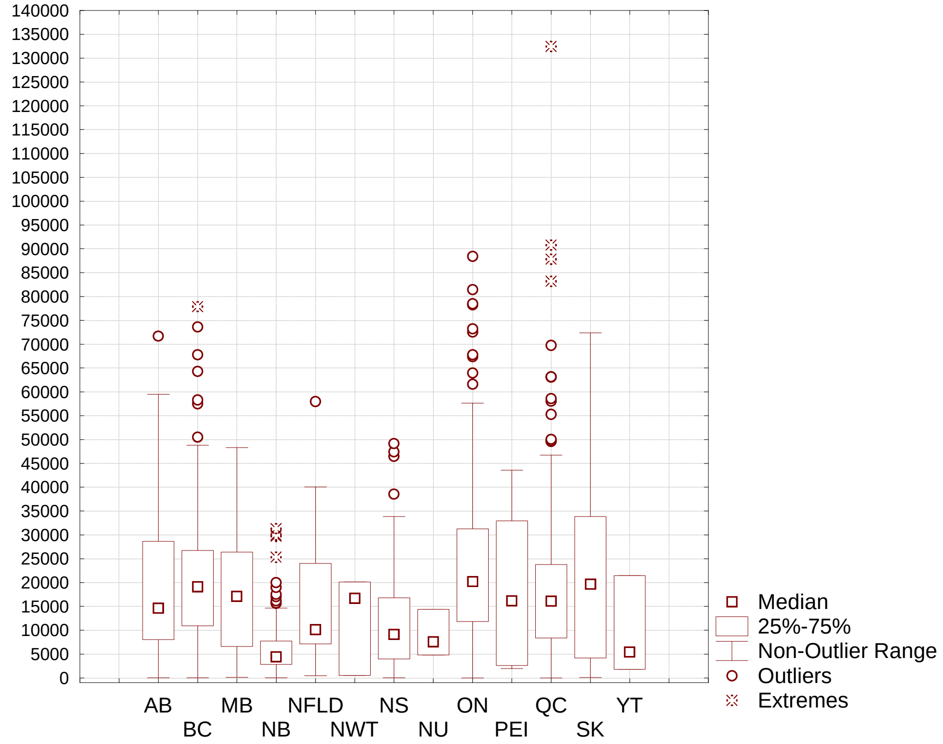


Figure 4: *Box and whisker plots showing the number of individuals per FSA across all of the provinces and territories.*

Whether the geographical unit is in an urban or a rural area will also make a difference. As can be seen by comparing the populations per FSA in Figure 5 (urban) and Figure 6 (rural) FSAs, the rural ones tend to have a larger population. For example, in rural Ontario FSAs have a median of 30,681 individuals compared with a median of 19,592 for urban ones. Again the variation is quite large in both types of communities.

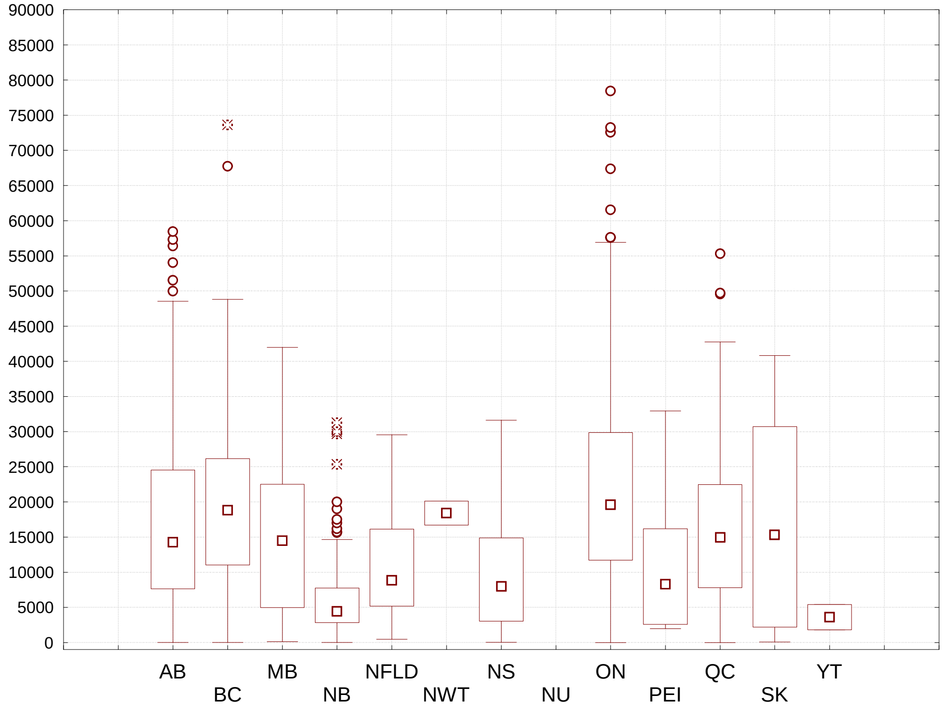


Figure 5: *Box and whisker plot showing the population per FSA for urban FSAs only across provinces and territories.*

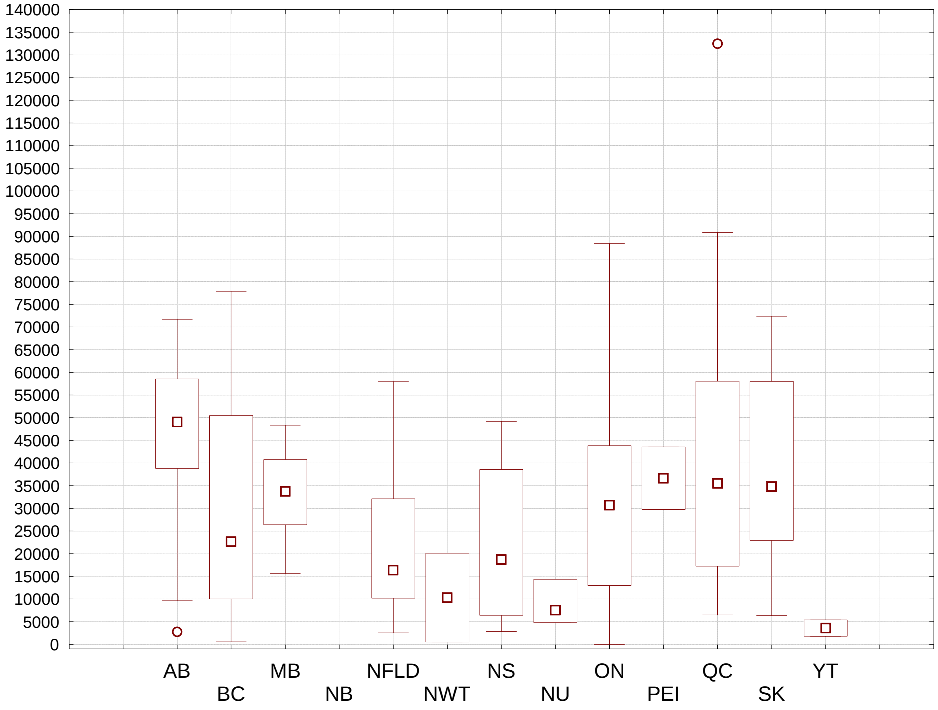


Figure 6: *Box and whisker plot showing the population per FSA for rural FSAs only across provinces and territories.*

The addition of quasi-identifiers to a record that has geographical information can increase the record's uniqueness. As an example we will consider physicians in Ottawa. Ottawa has 23 FSAs. The data are shown in Figure 7. The median number of male physicians in Ottawa per FSA is 24 and the median for females is 18. The median number of individuals per FSA for Ottawa is over 19,000. Therefore, the three quasi-identifiers [FSA, occupation, gender] narrowed down the median number of matches considerably (from more than 19,000 to 24 and 18 for male and female physicians respectively). In some FSAs there are only two female physicians ("K1P" and "K2E"). These two would be quite easy to re-identify if included in s .

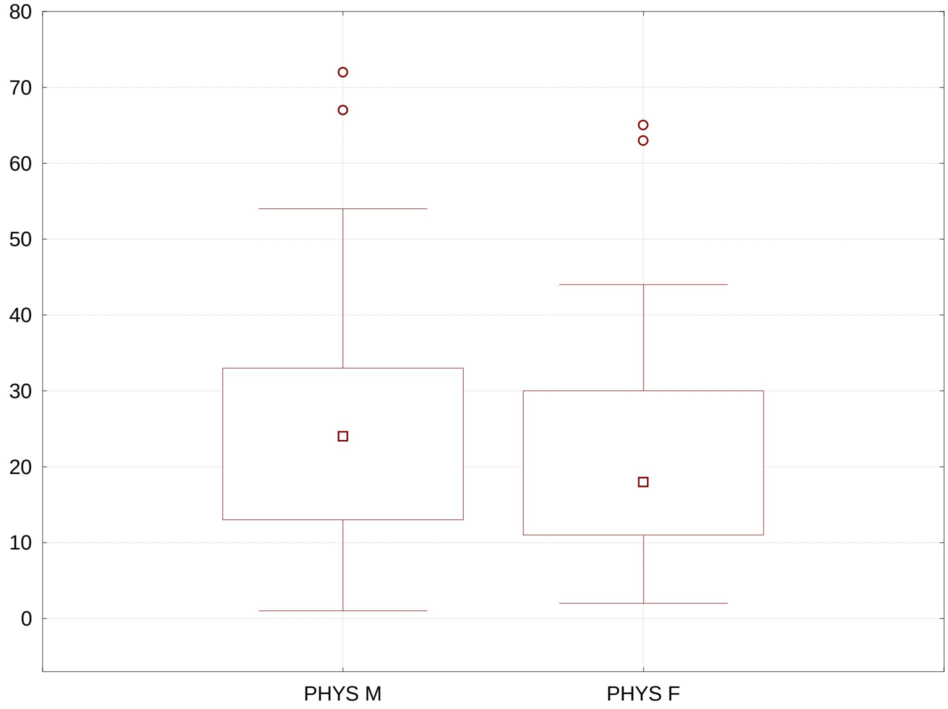


Figure 7: *Comparison of the number of physicians per FSA in Ottawa grouped by gender.*

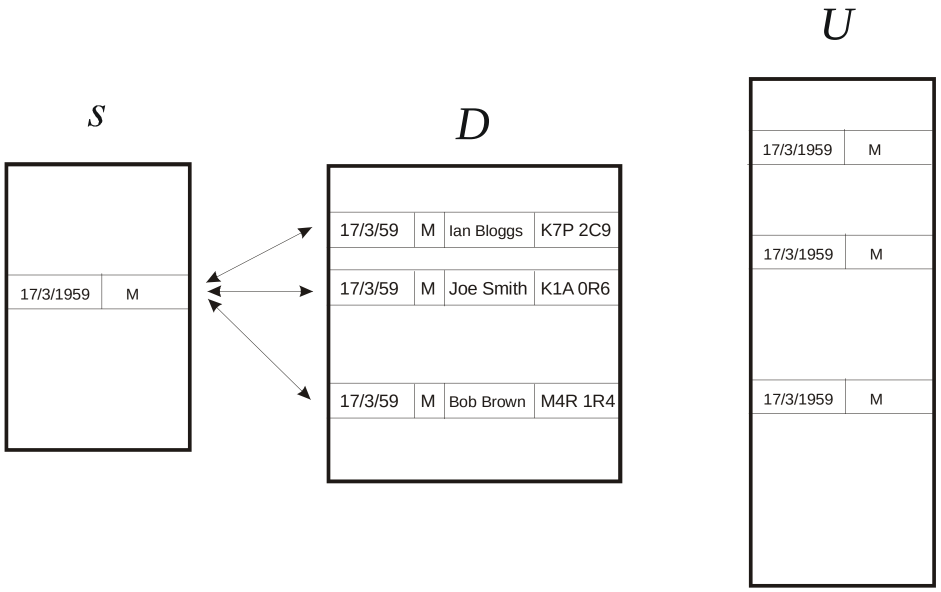
In summary then, we have shown a number of factors that determine the impact of geographical information (namely FSA) on uniqueness: the province or territory since the FSA sizes are not all equal, whether the location is rural or urban, and on other quasi-identifiers included in the record, such as occupation and gender.

1.6 Household Variables

Household variables may characterize the household itself or the members of the household. All members of the household are linked through the household variables because they will have the same values on all variables that characterize the household itself (e.g., the number of children and occupation of the head of the household). If one individual in the household is re-identified then the remainder of the household members would potentially be re-identified. The greater the number of members of the household, the greater the probability of re-identifying any one of them. Members of the household can also be re-identified if the household itself is unique among all households. Therefore, there are more opportunities for re-identification in household microdata sets compared to individual microdata sets. Consequently, it is important to ensure that households and their members are not unique on quasi-identifiers in a released data set.

1.7 Non-uniqueness

Thus far we have explored the different ways in which uniqueness can increase the probability of re-identification. However, from an attacker's perspective uniqueness may not even be necessary – it depends on the motive for the attack.



U
s
D
17/3/1959 M
17/3/59 M Ian Bloggs K7P 2C9
17/3/59 M Joe Smith K1A 0R6
17/3/59 M Bob Brown M4R 1R4
17/3/1959 M
17/3/1959 M
17/3/1959 M

Figure 8: *A match between the sample and the identification database identified three matches. In some cases it may not matter which one of the three matches in* D *is the correct one.*

Let's assume that the attacker wishes to re-identify an individual who has a certain diagnosis in s to send that individual some marketing materials on available therapies. Consider the example in Figure 8. The attacker may not be able to determine which of the three matching individuals in D is the correct one. This, however, may not matter as the attacker can send the marketing materials to all three matching individuals in D. The incremental cost of two more mailings may be so small that the additional cost does not act as an incentive to get correct matches.

Consider another scenario where s has data on 16 year olds in a particular geographical area and of a particular socio-economic status showing that they have all experimented with drugs. Even if no individual 16 year old is re-identified in s, the lack of variation on drug use among 16 year olds makes it relatively easy for say a parent to infer that their 16 year old has experimented. In this case specific records in s have not be associated with specific individuals, but attributes about these individuals have been disclosed from s.

2. General Factors

2.1 Existence of a Data Sharing Agreement

A data sharing agreement between A and E provides a framework to restrict what E can legitimately do with the data, how the data is stored and accessed, how long the data should be kept before it is destroyed, how it is destroyed, controls over who would have access to the data, and penalties for non-compliance with the agreement. Normally such agreements would also prohibit or restrict record linkage attempts.

Data released for public use or under access to information requests would not have a data sharing agreement in place.

2.2 Ability to Audit

If A is able to audit E on a regular basis to ensure that good data management, data protection, and access control practices and mechanisms are in place, then the overall risk can be reduced substantially. The scope of audits should cover all information technology that would be used to store and exchange data as well.

2.3 Means of Re-identification

This characterizes the extent to which the attacker has the financial and technical means to launch an attack on s. For example, the construction of some identification databases requires the collection of information from government registries record-by-record, and each record search has a fee. The construction of large identification databases under these circumstances can be costly and very time consuming. In general, the more resources that an attacker has the greater the chance that the attack will be successful.

2.4 Opportunity to Attack

This characterizes the ease with which the attacker would have access to the data set s itself. For example, if s will be made freely available through a download on a web site or is released for general public use, then the opportunity for an attack is very high. If s is being given to a handful of known researchers then the opportunity is very limited.

2.5 Alternative Approaches

This characterizes the extent to which there exist other methods to re-identify the individuals in s or achieve the attacker's goals. For example, if the attacker wishes to re-identify s to launch a marketing campaign, then there may be alternatives that are more certain to get the same list of individuals.

2.6 Variable Sensitivity

The sensitivity of the variables in a data set will vary, with some being potentially very sensitive. Examples of sensitive variables include: criminal history, mental health and addiction information, HIV status, and information on sexual behavior. Sensitive variables increase the risk exposure because the consequences of re-identification can be severe.

Sensitive variables can be non-identifying or can be quasi-identifiers. In general, additional care needs to be exercised if s contains any sensitive information

2.7 Characteristics of the Population

The population about which data in s is being released would have an impact on the consequences of re-identification. For example, the consequences of re-identification are likely to be more severe if s contains data exclusively on politicians, judges, and members of the police services. Alternatively, the data set may be about the general population but includes a *Profession* variable that allows one to construct a subset of these professions.

2.8 Visibility

When the value of a variable is generally known or can be ascertained easily, for example, gender, then that variable is considered to be more visible. High visibility increases the probability that an individual can be re-identified. For example, if s contains data on a diagnosis and a particular illness is asymptomatic then visibility would be low. However, if it is obvious that a person has an illness by just looking at them then the visibility is high.

2.9 Traceability

If the amount of effort needed to trace an individual from their value on a particular variable in s is small, then that variable increases the probability of re-identification. For example, variables related to place of residence or place of work make an individual more easily traceable.

3. Mapping Of Factors

In Table 2 we map each of the ten factors that were discussed above to the three dimensions of risk exposure. It is generally recommended that actions taken address all three dimensions to minimize the risk exposure.

	Probability of attempt to attack	Probability of successful attack	Consequences of successful attack
Uniqueness		X	
Visibility		X	
Traceability		X	
Data Sharing Agreement	X		
Ability to Audit	X		
Means	X		
Opportunity to Attack	X		
Alternatives	X		
Variable Sensitivity			X
Characteristics of Population			X

Table 2: *Mapping the ten risk factors with the three dimensions of risk exposure.*

Recommendations

The recommended strategy for minimizing risk exposure consists of addressing each of the three dimensions of risk exposure. We will focus only on actions that A can take to reduce risk exposure.

1. Minimizing Consequences of Attack

This set of recommendations is driven by the recognition that we will never reduce the probability of attempted attacks to zero and that it is unlikely that the probability of a successful attack will be zero for all records, then there will always be a residual non-zero probability of re-identification. Therefore, it is always necessary to make efforts to reduce the consequences of a successful attack:

- To the extent that it will not negatively effect the utility of s, remove the sensitive variables before the data is released.
- Remove variables that can identify susceptible subsets of individuals in s. For example, variables such as *Profession* and *Place of Work* may be candidates if there are values in the data set for susceptible groups (such as judges and abortion clinic doctors).

2. Deterrents For An Attempt to Attack

Provide strong deterrents for an attempt to attack. This is particularly important if the consequences of an attack are severe and if the probability of a successful attack are high. An effective approach is to make it more difficult to construct an identification database:

- On a regular basis examine the possible data sources for constructing an identification database and launch mock attacks on s. This will inform A about changing risks and provide guidance for additional actions.
- Whenever there is a data sharing agreement in place, ensure that there is a provision to audit E.
- For data releases with no data sharing agreements, extensive effort should be placed on reducing the probability of a successful attack.

3. Reduce the Probability of a Successful Attack

The agency A should modify the original data set s to minimize the probability of a successful re-identification attack, taking into consideration the consequences on the uses of the data (for example, one can remove all quasi-identifiers, but that will likely make the data useless for most users). Possible actions include:

- Identifying variables should not be included in s. It would be under exceptional circumstances with appropriately strong data sharing agreements that A would release data with this kind of information.
- Minimize the number of quasi-identifiers in s, in particular variables that can potentially be used for record linkage with external databases.
- If it is possible, it is always better to release microdata that is outdated than current. This makes it more difficult to match with external sources. A common delay is two years.
- Because order may be revealing (for example, all respondents from a particular area are clustered together in the released database), it is better to randomly permute the order of records in the released database.
- If certain combinations of quasi-identifiers are known to have a high risk of re-identification, it may be better to calculate the derived values that users would need and release these rather than the quasi-identifiers. For example, if users of s are interested in the distance between residence and work, then calculate that value and remove the residence and work postal codes from s.
- To the extent possible, do not reveal any steps taken by A to make re-identification less likely.
- The list of individuals for whom s pertains should not be released. For example, if s is data from a survey, then do not release another list indicating who the respondents to the survey were. If the sampling frame was obtained from another organization, then release a random sub-sample of s.
- Use top or bottom coding to remove extreme values on continuous variables. This will make it more difficult to identify outliers. For example, instead of including all ages, have an age category 65+ and group all those above that threshold.
- Geographical information makes it easier to trace individuals, especially when combined with other variables, such as profession and initials (which by themselves are innocuous). In general, geographical information should be released in the most general form possible given the expected uses of s (e.g., three letter postal code rather than full postal code).
- When releasing three letter postal codes it is important to consider the risk posed by other quasi-identifiers in the data set, the province, and whether the location is urban or rural.
- Check for indirect indicators of geography, such as distance of residence from a nuclear reactor or distance from an international airport.
- Rank the individuals in s in terms of the smallest number and length of quasi-identifier combinations that makes them unique and remove the highest risk individuals from s.

- Avoid releasing dates of birth. If such information is needed then generalize date of birth to age.
- Reduce the number of sample uniques through generalization. For example, convert age to age groups so that there is a minimum number of individuals in each age group.

More complex perturbations can be made to s to mask the data and make it more difficult to re-identify individuals, for example, the addition of noise, and the removal of individual values.

Appendix: Availability of and Access to Public Data

In this appendix we will discuss some of the practical issues involved in constructing an identification database. We will use the Ontario jurisdiction as a specific example to illustrate various points.

An identification database can be constructed from one or more sources. Some of the sources may be government and others may be commercial. A data source is useful for constructing an identification database if: (i) it contains quasi-identifiers about the individuals in the database to be released, *and* (ii) it contains some identifying information about those individuals. The utility of a data source will vary depending on the s being attacked because that will determine which quasi-identifiers are relevant.

Some examples of data sources that can be used for constructing an identification database are:

- **Telephone directories.** Examples of telephone directories are <www.canada411.com>, <www.whitepages.ca>, and <www.yellow.ca>. These are available on-line on the internet and accessed on a record-by-record basis. The whole directory can be purchased in bulk as well.
- **Court records.** These can be purchased in electronic format from commercial data brokers.
- **Registries of members of professional organizations.** Many such organizations will publish a complete list of their members on the internet, for example, the College of Physicians and Surgeons of Ontario and Professional Engineers Ontario. Commercial brokers also provide professional lists, such as LexisNexis, Martindale, and WestLists.
- **Public records/registries released by government bodies.** These are registries that are made available to te general public, for example, the land title registry and private property security registration database.[**]
- **Information in newspapers.** Newspaper announcements (e.g., births and deaths) and articles typically include identifying information and quasi-identifiers. Copies of newspapers are maintained at libraries.
- **Commercial data brokers.** They will sell microdata with individual records, although the accuracy of that data is suspect because of the way it is

[**] It is very difficult to get a comprehensive list of all data released to the general public. The privacy offices at government ministries, who would normally oversee data releases, do not maintain such lists.

constructed. Example data brokers are: Americanada, Prospects Influential, Nation Reach, and InfoCanada[††].

In the remainder of this appendix we will address practical and legal considerations when constructing identification databases.

Differing Coding Schemes

The underlying assumption of record linkage is that the coding schemes for variables in different databases are the same and can be mapped together. Examples of codes are diagnostic codes, drug codes, discharge codes, and expense codes. It is not always the case that they are same in different data sources and coding scheme incompatibilities (or data formats in general) can make the construction of identification databases difficult.

Bulk Release Versus Record-by-record Release

Data custodians may release information record-by-record or in bulk. In general, government agencies will not release data in bulk without restrictions on its use. One assumes that one reason is to make it more difficult to construct an identification database. The Privacy Commissioner of Ontario has maintained that government bodies cannot be compelled to provide the data in bulk to someone (e.g., an attacker) if they do not routinely provide the data in bulk to the public. However, there is nothing stopping the attacker from reconstructing a larger database by extracting data record-by-record from a public register.

Paper Versus Electronic Release

The data custodians may provide data on paper or electronically. Of course electronic access, especially over the internet, substantially reduces the barriers to the construction of an identification database. However, depending on the means and motives of the attacker, collecting paper records and then converting them to an electronic version may be worth the effort and cost.

Restrictions on Use

A variety of government bodies collect personal information (including common quasi-identifiers) specifically for the purpose of creating a record that is available to the general public (e.g., the land registry). Many public registries do not have restrictions on what the general public can access this information for. For instance, there are no provisions authorizing only those uses of public register information that are consistent with the purpose for which the information was compiled.

Attacker Must Have Prior Information

Many public registries require the attacker to provide some information about the individual before providing access to their record. One would assume that this is intended to make it more difficult browse through the complete listing. For example,

[††] For a more detailed description of the data brokerage industry in Canada, the types of information collected and how that data is collected, please refer to the following study: “A Report on the Canadian Data Brokerage Industry” by the Canadian Internet Policy and Public Interest Clinic, April 2006.

the Private Property Security Registration requires a search to be done by name. However, the White Pages and professional organizations provide names of a large percentage of the residents of say Ontario and comprehensive lists of their members respectively. Therefore, these kinds of pre-requisites for a search in the public registries are not necessarily a hindrance for constructing an identification database.

Further Reading

The following are key references providing additional information on the issues presented in this report.

1. *CIHR Best Practices for Protecting Privacy in Health Research*: CIHR; 2005.
2. Confidentiality and Data Access Committee. *Checklist on Disclosure Potential of Proposed Data Releases*: Office of Management and Budget; 1999.
3. Doyle P, Lane J, Theeuwes J, Zayatz L. *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*: Elsevier Science; 2001.
4. Jabine T. *Procedures for Restricted Data Access*. Journal of Official Statistics. 1993;9(2):537-589.
5. Willenborg L, de Waal T. *Statistical Disclosure Control in Practice*: Springer-Verlag; 1996.
6. Willenborg L, de Waal T. *Elements of Statistical Disclosure Control*: Springer-Verlag; 2001.

About The Author

Dr. El Emam is an Associate Professor at the University of Ottawa, Faculty of Medicine and a Canada Research Chair in Electronic Health Information. Previously he was a senior research officer at the National Research Council of Canada, where he was the technical lead of the Software Quality Laboratory, and prior to that he was head of the Quantitative Methods Group at the Fraunhofer Institute for Experimental Software Engineering in Kaiserslautern, Germany. In both of these latter roles he was working on the development of predictive models of software quality, and on developing and evaluating audits of software processes to ensure good project outcomes. In 2003 and 2004, Khaled was ranked as the top systems and software engineering scholar worldwide by the *Journal of Systems and Software* based on his research on measurement and quality evaluation and improvement, and ranked second in 2002 and 2005. He holds a Ph.D. from the Department of Electrical and Electronics Engineering, King's College, at the University of London (UK).