GUIDELINES WORKSHOP
December 1, 2005

ANONYMIZATION OF ELECTRONIC HEALTH INFORMATION DATA

Session Lead: Carole Lucock, LLB, LLM, University of Ottawa

BACKGROUND

This session draws on the work of Dr. Latanya Sweeney, who demonstrated that by linking three shared variables (data of birth, a portion of a ZIP code, and gender) from two sets of data (voter list and medical data), apparently anonymous medical data could be re-identified. This study relied on the general availability of a matching data source, which in this case was a voter list that could be purchased for $20.[1] Dr. Sweeny uses the term *quasi-identifiers* for those variables that, while not explicit like a name or address, can nevertheless, in combination with an external data source, be used to re-identify data. In her work on *k*-anonymity,[2] Dr. Sweeny notes that data-holders who wish to release data anonymously often do not know what data sources are available to the data recipient and are therefore unaware of which quasi-identifiers in their data set are risky. Consequently, release of data could be re-identified through the use of quasi-identifiers. Although Dr. Sweeney focused on externally (publicly) available data-sources, her insight with respect to the ability to re-identify through the use of quasi-identifiers would be equally applicable to the combination of two (or more) private data sets.

Dr. Khaled El Emam has tried to replicate Dr. Sweeney's research in Ontario using her three variables, data of birth, postal code and gender. He has found that there is no comparable data set that is externally available to enable the same re-linking in the case of medical data. His study did find, however, that readily available information for doctors and lawyers (at a cost) did permit replication of Dr. Sweeny's work.

Dr. El Emam has also conducted a qualitative study on how persons engaged in clinical research perceive privacy risks. Through interviews with 20 persons – investigators, study coordinators, Research Ethics Board (REB) members and IT personnel – Dr. El Emam found that while REBs may require anonymization there is no systematic or evidence-based approach concerning how this will be achieved. For example, although data limitation (data with some variables eliminated) was the method used for anonymization, knowledge of which variables to remove or

---

[1] L. Sweeney, *Uniqueness of Simple Demographics in the U.S. Population*, LIDAPWP4. Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA: 2000.

[2] L. Sweeney. *k*-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems,* 10 (5), 2002; 557-570, online: <http://privacy.cs.cmu.edu/dataprivacy/projects/kanonymity/kanonymity.pdf>.

which variables were high-risk was lacking and there was wide variation among practices. In general, decisions were made on the basis of intuition and hearsay rather than justified according to evidence. He also found that no one used statistical methods extensively. Dr. El Eman's findings are supported by the study of REB chairs and coordinators by Don Willison and others, which found considerable variation in the ability to recognize the potential for re-identification through the combination of variables.

LEGAL AND POLICY DIMENSIONS

The following provides a cursory overview of the guidance found in Canadian legislation with respect to anonymization or de-identification and raises issues in connection with these statutory schemes, particularly as concerns clarity and implications. This is followed by three examples of approaches taken to what is considered anonymous information, which are included to further inform consideration of the questions for this session.

1.1 The complex legislative landscape

The privacy legislative landscape in Canada is, to say the least, very complex and far from uniform. This is particularly relevant in the health information context because in different Canadian jurisdictions different rules apply, and it cannot be assumed that health information is protected according to a common (or similar) legislative regime.

Beginning with public sector legislation, which is found in all fourteen jurisdictions (often combined with general, access to government information provisions), and which varies from jurisdiction to jurisdiction in terms of scope. For example, some jurisdictions include institutions like hospitals, regional health authorities or universities under its coverage, while others do not.

Next we now have federal legislation (*PIPEDA*)[3] that applies nationally to the private sector when engaged in commercial activities, which probably includes key players in the health sector including physicians and pharmacists unless they are captured under other, more recently enacted legislative regimes. We also have general, private sector legislation in three provinces,[4] which, like *PIPEDA* may have application to health information for some purposes. And finally we have health-sector specific legislation in four provinces,[5] which more directly addresses all components of health care information whether in public or private settings.

1.1.1 Provisions of privacy laws

What all of these statutes have in common is that they apply to certain types of information and not to others. Information that is linked to identity is generally

[3] *Personal Information Protection and Electronic Documents Act*, S.C. 2000, c.5.
[4] Alberta, B.C. and Quebec.
[5] Alberta, Manitoba, Saskatchewan and Ontario

what these laws seek to protect and information that doesn't meet this criterion would likely not be subject to their provisions, which would include anonymous information that ***is*** anonymous (or de-identified) according to the specific Act's provisions. This is important to underscore because it has implications for what latitude a person who is in possession of the 'anonymous' information has in connection with its use (or disclosure). If it is anonymous, then statutory provisions relating to identifiable information will generally be inapplicable and the requirements of these statutes (consent, authorizations, approvals etc.) will also be inapplicable.

Herein lies the difficulty. Most of the statutes do not define what is meant by anonymous or de-identified information, rather this must be determined by reference to the definition of information that is covered by the Act and what the definition infers about what is excluded. This is made more difficult because definitions vary.

In some cases, a general definition is contained in the Act, for example, in the case of *PIPEDA* the definition is, "information about an identifiable individual, but does not include the name, title or business address or telephone number of an employee of an organization." By way of contrast, the definition in the Quebec legislation is "information concerning a natural person which allows the person to be identified."

Other legislation will contain a general definition that is similar to the definition found in *PIPEDA* (***about*** an identifiable individual), and will then go on to list the types of information included within the definition (see appendix A for the types of information that is typically listed here). These lists contain qualitatively different types of information; for example, sensitive information that if linked to identity would be problematic, information that can be used to uniquely identify (such as assigned numbers or biometrics such as fingerprints or genetic material), and information that may be both sensitive and identifying, for example, ethnicity.

Yet other legislation makes express provision to exclude anonymous or de-identified information and provides a description of what this means. Here too the standard varies. For example, the Saskatchewan health-sector legislation specifically excludes de-identified personal health information and defines this to mean "personal health information from which any information that may reasonably be expected to identify an individual has been removed."[6] By way of contrast, Alberta's health-sector specific legislation defines non-identifying information to mean "that the identity of the individual who is the subject of the information cannot be readily ascertained from the information", and Ontario's legislation states that identifying information means "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."[7]

[6] Note that this also includes when used in combination with other data.

[7] De-identification is also defined in the Act to mean: " to remove any information that identifies the individual or for which it is reasonably foreseeable in the circumstances that it could be utilized, either alone or with other information, to identify the individual, and "de-identification" has a corresponding meaning."

1.1.2 Additional Reasons to Know the Legal Standard

Although knowing whether or not specific information has been properly anonymized so as to remove it from the ambit of a statute is an important consideration, there are other aspects of privacy legislation that also relate to anonymization or de-identification that are relevant. For example, frequently when information is provided for research purposes in identifiable form there is a legislative requirement to remove identifiers at the earliest opportunity, but little guidance given as to what constitutes an identifier. Another example is the requirement in Alberta's health-sector legislation pertaining to research that prohibits the publication of "health information in a form that could reasonably enable the identity of an individual who is the subject of the information to be readily ascertained." Alberta's statute also carries a general expectation that collection, use and disclosure will occur with the "highest degree of anonymity that is possible in the circumstances."

1.1.3 The Authority to Anonymize

Although it appears to be commonly assumed that the act of anonymizing information is unproblematic, that is, that it can occur without legislative constraint – this is by no means clear. Most privacy legislation does not address this point directly and it could be argued that the act of anonymization constitutes a use of information that is governed by privacy laws. Whether such an argument would succeed is a question. When a similar question was raised before the U.K. Court of Appeal, the act of anonymizing data was seen to be unproblematic. There are also decisions by the federal and Alberta privacy commissioners that could be seen to imply that the act is unproblematic with respect to patient prescription data that has been de-identified. It is interesting to note that recent legislation pays more attention to this point and expressly includes the act of anonymizing data as a permitted use under the legislation, which further raises the question about the status of the act of anonymizing when the Act is silent.

1.2 Guidelines and Other Regimes

The Canadian Institutes of Health Research (CIHR) in its *Best Practices for Protecting Privacy in Health Research (September 2005)*, as a general principle and along similar lines to privacy legislation and the advice of others, counsels data limitation as a first principle (e.g. aggregated data or non-identified data is preferred to identifying data). The document provides a rank order of data identifiability according to the capacity to re-identify as follows:

I. **Directly identifiable**: The data contains direct identifiers of an individual (e.g. name, address, health number).

II. Coded:

a. **Single coded**: A participant's data are assigned a random code. Direct identifiers are removed from the dataset and held separately. The key

linking the code back to direct identifiers is available only to a limited number (e.g. senior members) of the research team.

b. **Double or multiple coded**: Two or more codes are assigned to the same participant's data held in different datasets (e.g. health administrative data, clinical data, genetic samples and data). The key connecting the codes back to participants' direct identifiers is held by a third party (such as the data holder) and is not available to the researchers.

III. Not directly identifiable and not coded: Direct identifiers were never collected or have been deleted, and there is no code linking the data back to the individual's identity.

IV. Non-identifiable: Any element or combination of elements that allows direct or indirect identification of an individual was never collected or has been removed, although some elements may indirectly identify a group or region. There is no code linking the data back to the individual's identity.[8]

The CIHR document distinguishes between direct and indirect identification as follows:

> Identifiable personal information may contain a direct link to a specific individual (e.g. name and street address, personal health number, etc.) or any element or a combination of elements that allows indirect identification of an individual (e.g. if birth date combined with postal code and other personal information on the record such as ethnicity could lead to the identification of an individual).[9]

These terms are further defined in the glossary as follows:

> **Direct identifiers**. These are variables such as name and address, health insurance number, etc., that provide an explicit link to a respondent. (Statistics Canada)
>
> **Indirect identifiers**. These are variables such as date of birth, sex, marital status, area of residence, occupation, type of business, etc. that, in combination, could be used to identify an individual. (Adapted from Statistics Canada)[10]

The U.S., in its 1996 *Health Insurance Portability and Accountability Act* (*HIPAA*), has taken a somewhat different approach. Like Canadian legislation it governs identifiable information; however, it goes a further step and list 18 elements and considers that if one or more of these elements is contained in the data then the information is identifiable for the purposes of *HIPAA*. The 18 elements are contained in Appendix B.

---

[8] Canadian Institutes of Health Research, *Best Practices for Protecting Privacy in Health Research* (Ottawa: Public Works and Government Services, September 2005) p.33.
[9] *Ibid*. p.19.
[10] *Ibid*. p.111.

Finally, in the U.K. the data commissioner has taken a fairly strong position on the issue of what will count as anonymized data for the purposes of excluding it from the *Data Protection Act*:

> The Commissioner considers anonymisation of personal data difficult to achieve because the data controller may retain the original data set from which the personal identifiers have been stripped to create the "anonymised" data. The fact that the data controller is in possession of this data set which, if linked to the data which have been stripped of all personal identifiers, will enable a living individual to be identified, means that all the data, including the data stripped of personal identifiers, remain personal data in the hands of the data controller and cannot be said to have been anonymised. The fact that the data controller may have no intention of linking these two data sets is immaterial. [11]

[11] U.K. Information Commissioner *Data Protection Act 1998: Legal Guidance* (London: Information Commissioner, 2002). A lengthier excerpt of this guidance is contained in Appendix C.

QUESTIONS

Question 1: The Standard for Anonymization

When data is anonymized, what standard ought to be aimed for or applied?

This question is ***not*** about which variables to remove or which methods to deploy to remove them (these questions are dealt with under different topics below). This question concerns the standard to be aimed for or applied to determine whether or not data can be defined (or described) as anonymous. The question assumes (based on the work of Sweeny, Emam and Williston *et al*) that uniform methods are not being used, which means that in a given context there is a risk that data could be unintentionally re-identified. However, it also assumes that there is not a common understanding of or agreement about what 'counts' as anonymous information.

The question also engages the use of language generally and the fact that terms are not sharply defined and sometimes used interchangeably. This is made even more complex by the differing uses and definitions found in legislation and policy documents. In particular, words such as, anonymous, de-identified, and non-identified may imply different things to different people, which might account for some of the variations in practices.

- Do we know what people who are anonymizing (de-identifying etc.) information are trying to accomplish?
- Do we know what standard those who are anonymizing (de-identifying etc.) data are deploying (even if they are unable to meet the standard that they set)?
- Do we know whether the standards (what is aimed for) that are being used correspond to legal and policy definitions, including those that would remove the information from the ambit of the legislative regime?

These questions also concern different definitions. Anonymization could mean:

- that data is anonymization for all occasions, with no key back to the original (identifying) data set since this too has been anonymized (the UK Data Commissioner Model);
- that data is anonymized when specified data elements have been removed (US Model) – this model implies that the original data set is intact;
- that data is anonymized prior to disclosure for discrete purposes (for example, the person who releases the data in anonymized form continues to hold the key to re-linking).

These questions also engage more pragmatic considerations, which include whether or not data is sufficiently anonymized to exclude it from data protection legislation, including risks associated with incorrectly assuming that data is not re-linkable, and general issues of public confidence based on the public's understanding of these terms.

Question 2: The Ability to Re-Identify and Knowledge Gaps Concerning Variables

How easy is it to re-identify data in Canada and what can be done to fill the knowledge gap of those currently responsible for anonymization?

This question relates both to the sources of available data to enable re-identification, and to knowledge on the part of those who are anonymizing data as to the specific risks associated with variable contained in their data.

Souces of data

Dr. El Emam's work is limited to externally available sources in Ontario. These would be insufficient grounds to assert that data-linkage of the type identified by Dr. Sweeney is not an issue in Canada. Options for further work include:

- Extend the study of external sources to other Canadian jurisdictions;
- Extend the study to include the possibility of re-linking across private data bases where data sharing is assumed to be on an anonymous basis;
- Investigate further sources of data, for example, what information is available commercially through data-brokers (in Canada and the U.S.);
- Extend the study to explore other variables that may pose equal problems to the ones found by Dr. Sweeney using date of birth, gender and a partial ZIP code.

Knowledge gap

Dr. El Emam and Dr. Willison have also identified glaring knowledge gaps in those who are charged with anonymizing data for a variety of purposes, including data-linkage. What practical measures can be taken immediately to raise general awareness in the community about the risks associated with variables and should this include a list of variables that are particularly problematic. If so, what would they be?

Question 3: The Use of Statistical and Scientific Methods, and IT applications

Are statistical and scientific methods, and IT applications available to assist in eliminating problematic variables? If so, what are the impediments to their use? Are there practical measures that can be taken to overcome identified impediments?

Dr. El Emam identified that there is minimal use of statistical or other methods to assist in the identification and elimination of problematic variables. This is perhaps, not surprising since the use of these methods is complex. In addition, there are applications available that can assist in eliminating variables (e.g. Datafly in the U.S.)[12]; however, these applications come with an associated cost. Moreover, these applications tend to be developed for the U.S. market, which inclines them to the HIPPA standard, which may or may not be suitable for the Canadian context. Nevertheless, there are mechanisms and applications available to properly identify problematic variables; consequently, the further question becomes whether it is irresponsible to continue to rely on intuition and hearsay as a method of anonymization.

To investigate how Canadian practices could be improved, which implies that there are practical and accessible options, would it be worthwhile to develop a benchmark problem and investigate to develop options?

[12] Carnegie Mellon, Data Privacy Lab., online: < http://privacy.cs.cmu.edu/datafly/>

Question 4: The Use of Other Mechanisms to Prevent Re-Identifying Data Linkage

What other mechanisms are available to prevent re-identifying data linkage and how can these mechanisms be implemented?

Increasingly, a significant degree of reliance is placed on data-sharing agreements and REBs. Often legislation either requires the use of data-sharing agreements in the research context and if it doesn't, it may either be specifically recommended in the legislation or through the offices of Privacy Commissioners or government agencies charged with administering privacy legislation. In addition, through the combination of the Tri-Council policy environment and increasingly through legislation, REBs are playing a significant role. Are these mechanisms adequate to ensure that anonymization is occurring properly so that re-identifying data linkage is minimized? Should, for example, the use of audits be increased?

It is interesting to note how institutions in the U.S. are approaching these issues and the seriousness and sophistication of their approach. See for example, the Human Investigation Committee of Yale University School of Medicine (http://www.med.yale.edu/hic/index.html), which includes rich resources for researchers and others as well as significant procedural safeguards.

APPENDIX A
LIST OF FACTORS LISTED IN LEGISLATION AS INCLUDED IN THE DEFINTION OF PERSONAL INFORMATION

"personal information" means recorded information about an identifiable individual, including

(i) the individual's name, address or telephone number,

(ii) the individual's race, national or ethnic origin, colour, or religious or political beliefs or associations,

(iii) the individual's age, sex, sexual orientation, marital status or family status,

(iv) an identifying number, symbol or other particular assigned to the individual,

(v) the individual's fingerprints, blood type or inheritable characteristics,

(vi) information about the individual's health-care history, including a physical or mental disability,

(vii) information about the individual's educational, financial, criminal or employment history,

(viii) anyone else's opinions about the individual, and

(ix) the individual's personal views or opinions, except if they are about someone else.[13]

[13] Freedom of Information and Protection of Privacy Act, S.N.S. 1993, c. 5.

APPENDIX B
HIPAA'S 18 DATA ELEMENTS

1. Names

2. All geographic subdivisions smaller than a State, including:
   - street address
   - city
   - county
   - precinct
   - zip codes and their equivalent geocodes, except for the initial three digits of a zip code if, according to the current publicly-available data from the Bureau of the Census: (1) the geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people, and (2) the initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000.

3. Telephone numbers
4. Fax numbers
5. E-mail addresses
6. Social Security numbers
7. Medical record numbers
8. Health plan beneficiary numbers
9. Account numbers

10. All elements of dates (except year) for dates related to an individual, including:
    - birth date
    - admission date
    - discharge date
    - date of death
    - all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older

11. Certificate/license numbers
12 Vehicle identifiers and serial numbers, including license plate numbers
13. Device identifiers and serial numbers
14. Web Universal Resource Locators (URLs)
15. Internet Protocol (IP) address numbers
16. Biometric identifiers, including finger and voice prints
17. Full face photographic images and any comparable images
18. Any other unique identifying numbers, characteristics, or codes

APPENDIX C

EXTRACT FROM LEGAL GUIDANCE PROVIDED BY THE U.K. DATA COMMISSIONER

"The Commissioner recognises that the aim of anonymisation is to provide better data protection. However, true anonymisation may be difficult to achieve in practice. Nevertheless, the Commissioner would encourage that, where possible, information relating to a data subject, which is not necessary for the particular processing being undertaken, should be stripped from the personal data being processed. This may not amount to anonymisation but is in line with the requirements of the Data Protection Principles.

The Commissioner considers anonymisation of personal data difficult to achieve because the data controller may retain the original data set from which the personal identifiers have been stripped to create the "anonymised" data. The fact that the data controller is in possession of this data set which, if linked to the data which have been stripped of all personal identifiers, will enable a living individual to be identified, means that all the data, including the data stripped of personal identifiers, remain personal data in the hands of the data controller and cannot be said to have been anonymised. The fact that the data controller may have no intention of linking these two data sets is immaterial.

A data controller who destroys the original data set retaining only the information which has been stripped of all personal identifiers and who assesses that it is not likely that information will come into his possession to enable him to reconstitute the data, ceases to be a data controller in respect of the retained data.

Whether or not data which have been stripped of all personal identifiers are personal data in the hands of a person to whom they are disclosed, will depend upon that person being in possession of, or likely to come into the possession of, other information which would enable that person to identify a living individual.

It should be noted that the disclosure of personal data by a data controller amounts to processing under the Act.

For example:
The obtaining of clinical information linked to a National Health Service number by a person having access to the National Health Service Central Register will amount to processing of personal data by that person because that person will have access to information enabling him to identify the individuals concerned.

It will be incumbent upon anyone processing data to take such technical and organisational measures as are necessary to ensure that the data cannot be reconstituted to become personal data and to be prepared to justify any decision they make with regard to the processing of the data.

For example:
In the case of data collected by the Office of National Statistics, where there is a disclosure of samples of anonymised data, it is conceivable that a combination of information in a particular geographic area may be unique to an individual or family who could therefore be identifiable from that information. In recognition of this fact, disclosures of information are done in such a way that any obvious identifiers are removed and the data presented so as to avoid particular individuals being distinguished.

If data have been stripped of all personal identifiers such that the data controller is no longer able to single out an individual and treat that individual differently, the data cease to be personal data. Whether this has been achieved may be open to challenge. Data controllers may therefore be required to justify the grounds for their view that the data are no longer personal data. “