



# An Overview of Techniques for De-identifying Personal Health Information

*14<sup>th</sup> August 2009*

**Khaled El Emam**  
*CHEO Research Institute &  
University of Ottawa*

**Anita Fineberg**  
*Anita Fineberg & Associates Inc.*

This report was funded by the Access to Information and Privacy Division of Health Canada.

## **Document Information**

**Document Title:** An Overview of Techniques for De-identifying Personal Health Information

**Original Document Date:** 26th January 2009

**Document Version:** Version 15

**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 Publications and Reports**

- K. El Emam, F. Dankar, R. Issa, E. Jonker, D. Amyot, E. Cogo, JP. Corriveau, M. Walker, S. Chowdhury, R. Vaillancourt, T. Roffey, J. Bottomley: "A Globally Optimal k-Anonymity Method for the De-identification of Health Data ." In *Journal of the American Medical Informatics Association*, 2009.
- K. El Emam, A. Brown, and P. AbdelMalik: "Evaluating predictors of geographic area population size cutoffs to manage re-identification risk." In *Journal of the American Medical Informatics Association*, 16(2):256-266, 2009.
- P. Kosseim and K. El Emam: "Privacy interests in prescription data. Part 1: Prescriber privacy." In *IEEE Security and Privacy*, January/February, 7(1):72-76, 2009.
- K. El Emam and P. Kosseim: "Privacy interests in prescription data. Part 2: Patient privacy." In *IEEE Security and Privacy*, March/April, 7(2):75-78, 2009.
- K. El Emam, F. Dankar, R. Vaillancourt, T. Roffey, and M. Lysyk: "Evaluating Patient Re-identification Risk from Hospital Prescription Records." In the *Canadian Journal of Hospital Pharmacy*, 62(4):307-319, 2009.
- K. El Emam: "Heuristics for de-identifying health data." In *IEEE Security and Privacy*, July/August, 6(4):58-61, 2008.
- K. El Emam, and F. Dankar: "Protecting privacy using k-anonymity." In the *Journal of the American Medical Informatics Association*, September/October, 15:627-637, 2008.
- K. El Emam, E. Neri, and E. Jonker: "An evaluation of personal health information remnants in second hand personal computer disk drives." In *Journal of Medical Internet Research*, 9(3):e24, 2007.
- K. El Emam, S. Jabbouri, S. Sams, Y. Drouet, M. Power: "Evaluating common de-identification heuristics for personal health information." In *Journal of Medical Internet Research*, 8(4):e28, November 2006.
- K. El Emam: "Overview of Factors Affecting the Risk of Re-Identification in Canada", Access to Information and Privacy, Health Canada, May 2006.
- K. El Emam: "Data Anonymization Practices in Clinical Research: A Descriptive Study", Access to Information and Privacy, Health Canada, May 2006.

*More information is available from*  
<http://www.ehealthinformation.ca/>

## Table of Contents

<b>1</b>	<b>INTRODUCTION .....</b>	<b>2</b>
1.1	WHEN TO DE-IDENTIFY PERSONAL HEALTH INFORMATION ?.....	2
1.2	THE NEED FOR DE-IDENTIFICATION .....	6
1.2.1	<i>Scenario A: Mandatory Disclosures</i> .....	6
1.2.2	<i>Scenario B: Uses by an Agent</i> .....	6
1.2.3	<i>Scenario C: Permitted Disclosures</i> .....	7
1.2.4	<i>Scenario D: De-identification vs. Consent</i> .....	7
1.3	DE-IDENTIFICATION TECHNIQUES .....	8
<b>2</b>	<b>DECIDING WHICH DE-IDENTIFICATION TECHNIQUE TO USE.....</b>	<b>9</b>
2.1	DEALING WITH THE IDENTIFYING VARIABLES.....	11
2.2	DEALING WITH QUASI-IDENTIFIERS .....	13
2.3	WHEN DOES PHI BECOME DE-IDENTIFIED ? .....	14
<b>3</b>	<b>OVERVIEW OF DE-IDENTIFICATION TECHNIQUES.....</b>	<b>17</b>
3.1	RANDOMIZATION .....	17
3.2	IRREVERSIBLE CODING.....	18
3.3	REVERSIBLE CODING.....	19
3.4	HEURISTICS .....	20
3.5	ANALYTICS .....	22
<b>4</b>	<b>FURTHER CONSIDERATIONS .....</b>	<b>26</b>
4.1	ATTRIBUTE DISCLOSURE.....	26
4.2	MEMBERSHIP IN A DATABASE .....	26
4.3	AUDIT TRAILS .....	26
4.4	RESIDENCE TRAILS.....	27
4.5	ENCOUNTER TRAILS .....	27
4.6	NUMBER OF RESIDENCES AND ENCOUNTERS .....	27
4.7	UNSTRUCTURED DATA .....	27
4.8	OTHER MEDIA .....	28
4.9	DATA QUALITY .....	28
<b>5</b>	<b>ACRONYMS.....</b>	<b>29</b>
<b>6</b>	<b>ACKNOWLEDGEMENTS .....</b>	<b>30</b>
<b>7</b>	<b>REFERENCES .....</b>	<b>31</b>

# 1 Introduction

---

There is increasing demand for the use and disclosure of personal health information (PHI) for secondary purposes. In the context of health information, “secondary purposes” is defined as any retrospective processing of existing data that is not part of providing care to the patient. For example, data used or disclosed for analysis, research, safety and quality measurement and improvement, public health, payment, provider certification and accreditation, and marketing are considered secondary purposes under this definition [1].

This report describes, at a high level, the techniques that can be used to de-identify PHI. In particular, our objective is to make clear when specific de-identification techniques are applicable and how to decide when to use each.

The scope of this report is micro-data or individual patient level data. We do not discuss techniques that are suitable for de-identifying magnitude or count tables, or other forms of aggregate statistics.

As a starting point it is first important to establish some basic data flows and determine when de-identification is necessary or recommended as a good practice.

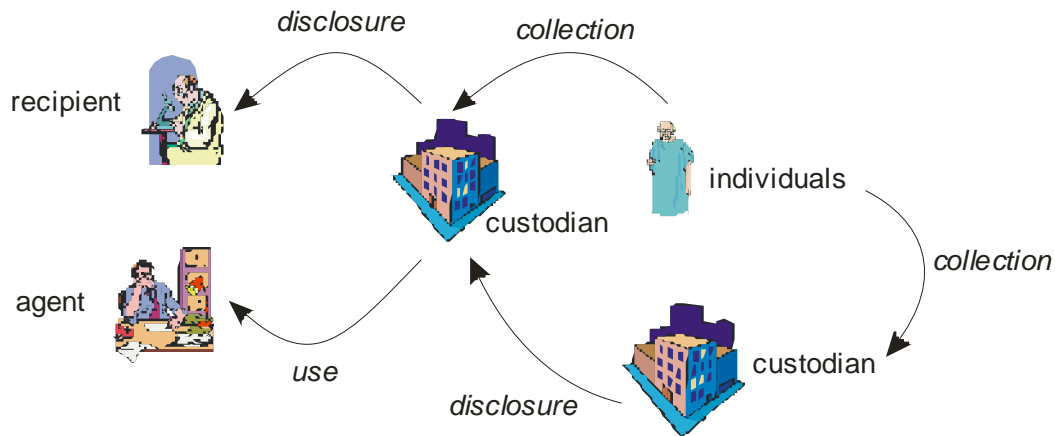
## 1.1 When to De-identify Personal Health Information ?

The data flows that are relevant for this report are illustrated in Figure 1. It should be noted that the terms we use are often Ontario-specific, but there are almost always equivalents in other jurisdictions.

PHI is about individuals. They may be patients or healthy clients. PHI may be collected from individuals directly or indirectly through reporters. An example of indirect collection is in the case of an adverse drug reaction, a hospital or a physician may report the adverse reaction rather than the patient herself.

This PHI goes to a custodian. An example of a custodian is a hospital or a disease registry. The custodian may have collected the information for a primary purpose, such as providing a service to a customer or providing care to a patient. The custodian may have also collected the information explicitly for a secondary purpose, such as constructing a prospective diabetes registry for subsequent research.

A custodian may disclose PHI to another custodian. For example, a hospital may disclose PHI to a public health agency. In such a case, the information is not coming directly from an individual, but indirectly through one (or possibly more) custodians.



**Figure 1:** Basic data flow during a disclosure or use of PHI for secondary purposes.

An agent of the custodian may *use* the information for a secondary purpose. An agent is broadly defined as a person who acts on behalf of the custodian in respect of the personal information for the purposes of the custodian. For example, a data analyst employed by a hospital to produce reports on resource utilization would be an agent. There is generally no legislative requirement to de-identify information that an agent uses and no requirement to obtain additional consent from the individuals/patients for such uses.

The custodian may also get a request to *disclose* the information to a recipient for some secondary purpose. The recipient can be an individual (e.g., a researcher), or an organization (e.g., a pharmaceutical company). The recipient can also be internal or external to the custodian. For example, a researcher may be based within a hospital or can be an external researcher at a university or a government department requesting the information from the hospital.

Some disclosures are mandatory and some are discretionary by the custodian. An example of a mandatory disclosure is reporting communicable diseases or reporting gunshot wounds in some jurisdictions. In these situations the disclosure of PHI to a particular recipient is required.

Otherwise, there are different types of recipients and purposes where disclosures of personal information are discretionary. However, the conditions for discretionary disclosure do vary. There are a set of permitted disclosures in privacy legislation where personal information may be disclosed without consent. Table 1 presents some examples of such permitted disclosures.

Purpose of Disclosure/Recipient	Mandatory (M)/ Discretionary (D)	Restrictions on Recipient	Legislation
For health or other programs: <ul style="list-style-type: none"> <li>• Eligibility to receive health care</li> <li>• Audits or accreditation of the custodian</li> <li>• QA Activities</li> <li>• Prescribed registries</li> <li>• Medical officer of Health under the <i>Health Protection and Promotion Act</i> (Ont.) for the purposes of that Act</li> <li>• Public health authority in Canada or another jurisdiction for a purpose substantially similar to that in the <i>Health Protection and Promotion Act</i> (Ont.)</li> </ul>	D *note that all of the disclosures under Nfld. PHIA ss.39, 41 and 43 are mandatory	Restrictions or requirements in the regulations	PHIPA, ss.39(1)(2); Nfld. PHIA, s.39(1)(a); HIA, s.35(1)(f); Nfld. PHIA, s.39(4)(b); HIA, s.35(1)(g); Sask. THIPA, s.27(4)(g); O. Reg. 39/04 s.13 (registries); Nfld. PHIA s.39(4)(d); Nfld. PHIA, s.39(4)(e); Nfld. PHIA, s.39(4)(f)
Related to risks/care and custody: <ul style="list-style-type: none"> <li>• Eliminating or reducing significant risk of serious bodily harm to a person or group</li> <li>• Penal/psychiatric institutions to assist in decision-making regarding care and/or placements</li> </ul>	D		PHIPA, s.40; HIA, s.35(1)(e); HIA, s.35(1)(m); Manitoba PHIA, s. 22(2)(b); HIPA, s.27(4)(a); Nfld. PHIA, s.40(1); Nfld. PHIA, s.40(2)
Compliance with subpoena, warrant, court order	D		PHIPA, s.41(1)(d)(i); HIA, s.35(1)(i); Manitoba PHIA s.22(2)(l); HIPA, s.27(2)(i); Nfld. PHIA, s.41(1)(b)
Disclosure for Proceedings	D	Requirements or restrictions in the regulations (none currently prescribed in PHIPA)	PHIPA s.41(1); HIA, s.35(1)(h); Manitoba PHIA, s.22(2)(k); THIPA, s.27(2)(i); Nfld. PHIA, s.41(2)(a)
Disclosure to (potential) successor of custodian	D	Potential successor must enter into an agreement with the custodian re confidentiality and security of the PHI (PHIPA)	PHIPA, ss.42(1)(2); HIA, s.35(1)(q); Manitoba PHIA, s.27(4)(c); Nfld. PHIA, s.39(1)(i),(j)
Disclosures related to other Acts	D		PHIPA, s.43; HIA, s.35(1)(p); Manitoba PHIA, s.22(2)(o); THIPA, s.27(2)(l); Nfld. PHIA, s. 43

Purpose of Disclosure/Recipient	Mandatory (M)/ Discretionary (D)	Restrictions on Recipient	Legislation
Disclosures for research	D	Researcher must provide custodian with a research plan (as prescribed), obtain REB approval (as prescribed) and enter into a research agreement (PHIPA; HIA). Researcher is permitted to disclose information as prescribed (PHIPA).	<p>PHIPA, s.44; O. Reg. 39/04, s.15 (REB requirements); s.16 (research plan requirements); s.17 (disclosure by researchers)</p> <p>HIA, ss. 48-54 (note that the researcher <b>may</b> submit a proposal to an REB; it is not mandatory)</p> <p>Manitoba PHIA, s.24 (project must be approved by a health information privacy committee (Reg.245/97, s.8.1 and 8.2) or REB; agreement required (Reg. 245/97, s.8.3)</p> <p>PHIPA, s.29 (research ethics committee approval is required; research agreement is required)</p> <p>Nfld. PHIA, s.44 (REB approval required under the <i>Health Research Ethics Authority Act</i>)</p> <p>Sask. THIPA, s.29</p>
Disclosure for planning and management of the health system: <ul style="list-style-type: none"> <li>• To prescribed entities for analysis or compilation of statistical information</li> </ul>	D	Prescribed entity must have practices and procedures in place to protect privacy and maintain confidentiality; reviewed by the Commissioner in Ontario	PHIPA, s.45; O. Reg. 39/04, s.18 (prescribed entities); Nfld. PHIA, s.39(1)(h)
For monitoring health care payments: <ul style="list-style-type: none"> <li>• To the Minister on his/her request</li> <li>• To verify or monitor claims for payment for care funded in whole or in part by the Ministry or LHIN</li> </ul>	D		PHIPA, s.46; Nfld. PHIA, s.39(1)(b)
For analysis of the health system: <ul style="list-style-type: none"> <li>• At the Minister's request (to a health data institute in PHIPA only)</li> <li>• Management, evaluation, monitoring, allocation of resources or planning</li> </ul>	D	Minister must submit proposal to the Commissioner for review and comment (PHIPA); Subject to prescribed restrictions (none enacted for PHIPA)	PHIPA ss.47, 48; HIA s.46

**Table 1:** Examples of permitted discretionary disclosures of personal health information.

Other discretionary disclosures that are not explicitly permitted in legislation require that either consent be obtained from the individuals/patients or that the information is de-identified. For example, the disclosure of PHI to a pharmaceutical company for marketing purposes requires that consent be obtained or that the information is deemed to be de-identified.

Therefore, to summarize, there are four scenarios to consider:

- A. It is mandatory to disclose PHI to a recipient (usually external to the custodian), and no consent is required.

- B. PHI is used by an agent without consent.
- C. It is permitted by legislation to disclose PHI to a recipient (either internal or external to the custodian) without consent under the discretion of the custodian.
- D. The custodian *must* de-identify the PHI *or* obtain consent before disclosing the data to the recipient.

The need for de-identification of the information under each of the above scenarios will vary. This is discussed further below.

## 1.2 The Need for De-identification

In three out of the above four scenarios where data is used or disclosed, a strong case can be made for de-identification. Below we consider each in turn.

### 1.2.1 Scenario A: Mandatory Disclosures

Disclosures under this scenario are outside our scope since they do not require any de-identification.

### 1.2.2 Scenario B: Uses by an Agent

While agents are permitted to access PHI, if PHI is not necessary to perform their functions then it may be better to de-identify that information as a risk mitigation exercise. There are two reasons why a custodian may wish to do so. The “general limiting principles” in privacy laws stipulate that PHI should be collected, used, or disclosed where no other information will serve the purpose [2]. For example, if de-identified information will serve the purpose of a use by an agent, then the application of these principles means that the information should be de-identified first. The second consideration is that number of data breaches in medical establishments and of medical records is quite high, increasing the potential liability and reputation risks to the custodians, as well as decreasing patient trust. For example, between January 2007 and June 2009 there were at least 143 data breaches from medical establishments (including organizations such as hospitals and health insurance companies) in the US and Canada affecting more than 6.3 million personal records, and 106 breaches involving 2.5 million medical records [3]. The consequences of data breaches are significantly reduced if the data is de-identified.

For instance, consider a hospital network that has developed a system to provide its patients web access to its electronic health records. The hospital has sub-contracted the work to perform quality control for this software to a testing company across town. The testing company needs realistic patient data to test the software, for example, to make sure that the software can handle large volumes of patient records, that it displays the correct information to each patient, and so on. The testing company would be considered an agent of the hospital, and therefore it can obtain identifiable patient records without consent for the purpose of software testing. Giving the testing company PHI potentially exposes the hospital to additional risk if there is an inadvertent disclosure of this data (e.g., a breach at the testing company’s site). It is always preferable from a

risk management perspective to minimize the number of people who have access to PHI, and making that information available to the whole test team should be avoided if possible. Therefore in cases where there is a legitimate use of the PHI, one should still consider using de-identification techniques on the test data even if this is not a legal or regulatory requirement.

### **1.2.3 Scenario C: Permitted Disclosures**

In some cases, even though the disclosure of identifiable health information is permitted by legislation, the custodian may consider de-identification anyway. This, of course, makes sense only if the purpose can be satisfied without having identifiable information. In practice, achieving many purposes does not require identifiable information. A good example of that is in the context of research.

A Research Ethics Board (REB) determines whether custodians can disclose personal health information to researchers, and whether that information needs to be de-identified. REBs have total discretion to make that decision.

In practice, most REBs will require that either consent from the patients be sought if the information needs to be identifiable or they will require that the disclosed information is adequately de-identified [4]. However, because of the discretionary nature of this type of disclosure, they may also allow identifiable information to be disclosed without consent under certain conditions.

For example, consider the situation where a researcher is collecting clinical information from electronic health records (EHRs) and wants to link it with data in a provincial administrative database. The linking will not work if the EHR data is de-identified. In that case the REB may allow identifiable information to be disclosed for the purpose of linking without requiring the consent of the patients. The REB may then require the de-identification of the linked data.

### **1.2.4 Scenario D: De-identification vs. Consent**

In this scenario the custodian does not have the option to disclose identifiable information without consent. However, there will be situations where obtaining consent is not possible or practical. For example, in a health research context, making contact with a patient to obtain consent may reveal the individual's condition to others against their wishes, the size of the population represented in the data may be too large to obtain consent from everyone, many patients may have relocated or died, there may be a lack of existing or continuing relationship with the patients to go back and obtain consent, there may be a risk of inflicting psychological, social or other harm by contacting individuals and/or their families in delicate circumstances, it may be difficult to contact individuals through advertisements and other public notices, and undue hardship may be caused by the additional financial, material, human, organizational or other resources required to obtain consent. In those instances, the disclosure of personal information would not be permissible and de-identification provides the only practical option for disclosure (assuming that

the purpose can be achieved with the de-identified information). There is no legislative requirement to obtain consent for de-identified information.

Even if obtaining consent was possible and practical, it may have a severe adverse consequence on the information's quality because individuals who consent tend to be different on many characteristics than those who do not consent (e.g., on age, gender, socioeconomic status, whether they live in rural or urban areas, religiosity, disease severity, and level of education) [5]. These differences can result in biased findings when the information is analyzed or used. In such circumstances a strong case can be made for not seeking consent and de-identifying the information instead (again, assuming that the de-identified information will achieve the purpose of the disclosure).

Consider an example where a hospital is disclosing prescription data to a commercial data broker. The hospital can make the case that it is not practical to obtain consent from the patients for this disclosure. The cost and delays of prospectively administering the additional consent forms for all admitted patients may be difficult to justify, and it would be quite time consuming to do so retroactively for historical data. Therefore, the hospital would have to de-identify the prescription data before disclosure.

### **1.3 De-identification Techniques**

There are many different techniques for de-identifying PHI. The objective of this report is to provide a high level overview of these de-identification techniques, and give some guidance on when each of these should be used. The de-identification techniques that we cover are:

- Randomization / masking / obfuscation
- Coding / pseudonymization
- Heuristics
- Analytics

The literature does not always use a consistent terminology to describe the above techniques. Therefore, the terms we use reflect common usage but a reader may on occasion come across different terms elsewhere that represent the same techniques.

An assumption that will be made at the outset is that the PHI is in structured format (e.g., fields in a database) rather than in unstructured text. Hence, we will talk about de-identifying data or datasets. The topic of de-identifying text will be addressed at the end of the report.

## 2 Deciding Which De-identification Technique to Use

The flow chart in Figure 2 shows the main decisions that need to be made when selecting a de-identification technique.

The first pair of questions to ask is whether there are any identifying variables and whether there are any quasi-identifiers in the dataset? An identifying variable would be, for example, a name, full address, telephone number, email address, health insurance number, and social insurance number. This type of information directly identifies an individual. A quasi-identifier means a variable that can indirectly identify an individual, such as a date (birth, death, admission, discharge, autopsy, specimen collection, or visit), postal code or other location information, profession, and diagnosis information. Examples of common quasi-identifiers are shown in Table 2.

In some instances the answers to the above two questions are obvious. For example, if the data consists of a mailing list of breast cancer patients and it contains names and addresses, then there are only identifying variables in this dataset.

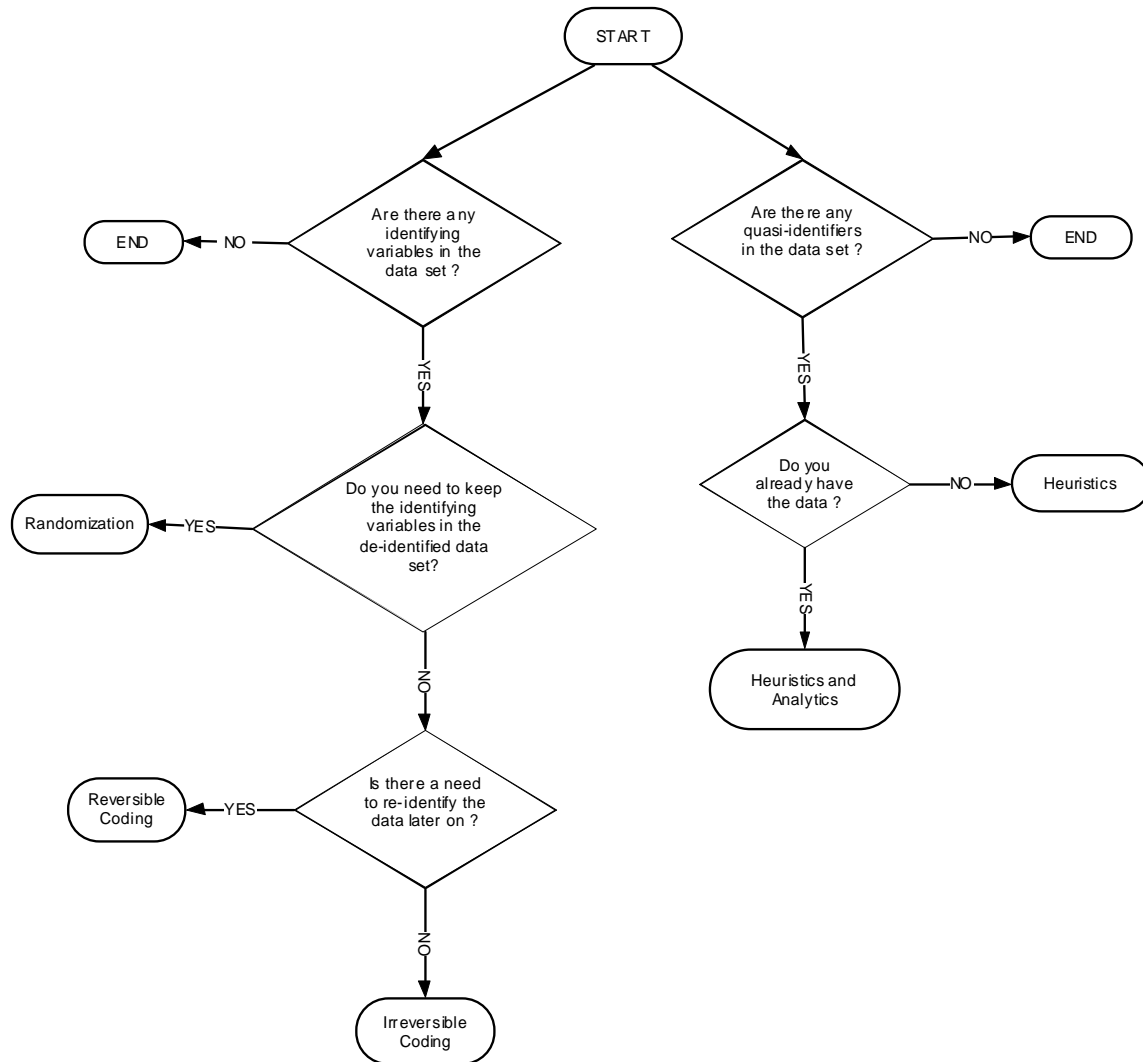
Sex	Total income
Date of Birth / Age	Visible minority status
Geocodes (such as postal codes, census geography, information about proximity to known or unique landmarks, other location information)	Activity difficulties/reductions
Home language	Profession
Ethnic Origin	Event dates (admission, discharge, procedure, death, specimen collection, visit)
Aboriginal Identity	Codes (diagnosis, procedure, and adverse event)
Total years of schooling	Religious denomination
Marital Status (Legal)	Country of birth
Criminal history	Birth plurality

**Table 2:** Common quasi-identifiers [6-8].

The line between identifying variables and quasi-identifiers is sometimes fuzzy. For example, sometimes postal codes are described as identifying variables and in other cases as quasi-identifiers. A general rule of thumb to resolve the ambiguity is that if a variable is important for subsequent analysis of the data then it should be treated as a quasi-identifier. The techniques for dealing with identifying variables cause considerable distortion to the data, and therefore, if data quality and data integrity are important, treat the variable as a quasi-identifier.

If there are identifying variables in the dataset then the two main types of relevant de-identification techniques are: randomization, and coding. If there are quasi-identifiers in the dataset then heuristics and analytics are called for.

In many datasets there will be both types of variables. Therefore in practice it will be necessary to apply more than one set of techniques. For example, an adverse drug reaction report may have identifying information about the patient and the reporter. It will also have basic demographics on the patient. The former are the identifying variables and most of the patient demographics would be quasi-identifiers.



**Figure 2:** A flow chart showing the key decisions that need to be made when selecting the type of de-identification technique to use.

## 2.1 Dealing with the Identifying Variables

An initial decision is whether the identifying variables need to be included in the disclosed dataset. There will be some situations where these variables need to be included in the disclosed dataset, although the values themselves will be distorted. For example, consider a situation where an order entry system is being implemented in a hospital and the system does not behave as expected. The hospital then calls the vendor and explains the problem. The vendor asks for some example data and transactions to recreate and troubleshoot the problem. However, because the example data would include sensitive personal health information, the hospital is reluctant to send data extracts to the vendor. In this particular example, the data extracts need to include all of the fields with the identifying information, but the actual values need not be true

patient values; the hospital only needs to provide realistic data values that recreate the problem. In those scenarios randomization techniques may be used.

If it is not necessary to retain the identifying variables in the disclosed dataset, then the next question is whether it will be necessary to re-identify the data at a later stage. For example, if participants in a clinical study are undergoing genetic tests beyond the standard of care, it may be necessary to inform the patient or their physician if these additional tests reveal a risk to the participants or their families (unless that requirement is waived by the REB). In such a case, the research investigators would need to be able to re-identify the patients with positive results. If the patients need to be re-identified, then reversible techniques must be used, such as reversible coding. Otherwise irreversible coding can be used.

If there is no need for the identifying variables in the dataset, then these can be removed completely (i.e., suppressed).

<b>General Examples of Re-identification</b>	
AOL search data [9-11]	AOL put anonymized Internet search data (including health-related searches) on its web site. New York Times reporters were able to re-identify an individual from her search records within a few days.
Chicago homicide database [12]	Students were able to re-identify a significant percentage of individuals in the Chicago homicide database by linking with the social security death index.
Netflix movie recommendations [13]	Individuals in an anonymized publicly available database of customer movie recommendations from Netflix are re-identified by linking their ratings with ratings in a publicly available Internet movie rating web site.
<b>Health-specific Examples of Re-identification</b>	
Re-identification of the medical record of the governor of Massachusetts [14]	Data from the Group Insurance Commission, which purchases health insurance for state employees, was matched against the voter list for Cambridge, re-identifying the governor's record.
Southern Illinoisan vs. The Department of Public Health [15, 16]	An expert witness was able to re-identify with certainty 18 out of 20 individuals in a neuroblastoma dataset from the Illinois cancer registry, and was able to suggest one of two alternative names for the remaining two individuals.
Canadian Adverse Event Database [17]	A national broadcaster aired a report on the death of a 26 year-old student taking a particular drug who was re-identified from the adverse drug reaction database released by Health Canada.

**Table 3:** Some examples of re-identification attempts in general datasets and of health datasets.

The former type of data can contain health information (as in the case of the individual re-identified in the AOL example), and life style and sexual orientation information (as in the case of one of the individuals re-identified in the Netflix example). This table is based on the one which appeared in [18]

## 2.2 Dealing With Quasi-identifiers

Dealing with directly identifying variables, as described in the section above, is insufficient to ensure that the data is truly de-identified. As illustrated in Table 3, there are real examples of datasets that had the identifying variables suppressed or coded, and the individuals were still re-identified. If there are quasi-identifiers in a dataset, then their de-identification is necessary.

The main question to ask about the quasi-identifiers is whether we are making de-identification decisions before data is collected or after the data is collected. If before, then a set of heuristics need to be used. The heuristics are like "rules of thumb" that can inform data collection activities to ensure that data is collected anonymously, or that the data is de-identified at the earliest opportunity after collection [19]. For example, assume that an REB is reviewing a research

protocol, the investigator has claimed that the data to be collected is de-identified, and is accordingly making a case for waiving consent. The REB can then apply the heuristics to make a judgment on whether the data to be collected is truly de-identified. Heuristics are approximations to the best decision regarding how to de-identify a dataset. But in the absence of actual data, they are a reasonable approach to de-identification.

If data have already been collected, then both the heuristics as well as analytics (statistical and computational methods) can be applied to the data. The analytics techniques involve the analysis of the dataset itself and then transforming it so that it is de-identified. Some of this will be illustrated in the next section.

### **2.3 When Does PHI Become De-identified ?**

We have already noted that dealing with the identifying variables (Section 2.1) does not make a dataset de-identified. It is important to also deal with the quasi-identifiers. There are degrees to which heuristics and analytics can be applied, meaning that there are degrees of identifiability. A relevant question then is to decide how much quasi-identifier de-identification to apply to ensure that the dataset is truly de-identified ?

Privacy legislation, and its various interpretations, makes a distinction between identifiable and de-identified information. These definitions, implicitly or explicitly, provide a threshold for when information ceases to be de-identified and becomes identifiable. We will first review some of these definitions and then make some recommendations about what would be a reasonable approach.

Under the EU Data Protection Directive, an “individual shall not be regarded as identifiable if the identification requires an unreasonable amount of time and manpower”, and the German Data Protection Act defines “rendering anonymous” as “the modification of personal data so that the information concerning personal or material circumstances can no longer or only with a disproportionate amount of time, expense and labour be attributed to an identified or identifiable individual” [20]. These two definitions refer to the amount of effort or resources required for re-identification as a criterion for deciding if a dataset is de-identified.

The Article 29 Data Protection Working Party notes that the term “identifiable” should account for “all means likely reasonably to be used either by the controller or by any other person to identify the said person” [21]. The reasonableness argument is also used often in Canadian health privacy legislation [22]. For example, Ontario’s PHIPA 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.

The Canadian Institutes of Health Research (CIHR) offered the follow interpretation of “information about an identifiable individual” to include only information that can [23]: i) identify, either directly or indirectly, a specific individual; or, ii) be manipulated by a reasonably foreseeable method to identify a specific individual; or, iii) be linked with other accessible information by a reasonably foreseeable method to identify a specific individual. CIHR also noted that “information about an identifiable individual” shall not include [23]: i) anonymized information which has been permanently stripped of all identifiers or aggregate information which has been grouped and averaged, such that the information has no reasonable potential for any organization to identify a specific individual; or ii) unlinked information that, to the actual knowledge of the disclosing organization, the receiving organization cannot link with other accessible information by any reasonably foreseeable method, to identify a specific individual. They note that whether or not a method is reasonably foreseeable shall be assessed with regard to the circumstances prevailing at the time of the proposed collection, use or disclosure of the information. Using the reasonableness standard “identifiable” would encompass only those technical possibilities that are realistically, practically and rationally foreseeable in the circumstances, while excluding those which are highly unlikely, immoderate or unfeasible to expect [24].

Another EU regulation on data sharing defines “anonymized microdata” as “individual statistical records which have been modified in order to minimize, in accordance with current best practice, the risk of identification of the statistical units to which they relate” [25]. One of the methods specified in the US HIPAA Privacy Rule is to get a qualified statistician to certify that a data set is de-identified<sup>1</sup>. The Secretary of Health and Human Services has approved two federal documents as sources of guidance to what is generally accepted statistical and scientific principles for de-identification [28]. In these cases the definition of identifiability refers to what are considered current best practices.

The Privacy Commissioner of Canada has proposed the “serious possibility” test to determine whether information is about an identifiable individual “Information will be about an identifiable individual where there is a serious possibility that an individual could be identified through the use of that information, alone or in combination with other available information” [29]. This is a more stringent test than a “reasonable expectation” test, which is sometimes proposed, in that if it is expressed probabilistically, the probability of re-identification threshold is higher than for a “serious possibility”.

---

<sup>1</sup> However, the statistician method is not used often in practice because it is perceived as not precise and too complex 26. Beach J. *Health care databases under HIPAA: Statistical approaches to de-identification of protected health information*. DIMACS Working Group on Privacy/Confidentiality of Health Data. 2003.. Furthermore, there have been concerns about the liability of the statisticians should the data be re-identified at a later point 27. American Public Health Association. *Statisticians and de-identifying protected health information for HIPAA*. 2005; Available from: [<http://www.apha.org/membergroups/newsletters/sectionnewsletters/statis/fall05/2121.htm>].

Another test that has been proposed to determine whether information is identifiable is to ask “Could this information ever be linked to the individual by the police for use as evidence ?” [30]. It is argued that anything that is easy for the police to do, is usually easy for hackers or inside operators to do.

The Supreme Court of the State of Illinois ruled that even if re-identification was shown empirically to be possible by an expert in the field, it was not reasonable to expect that non-experts, or even different experts in the art of re-identification, would be able achieve the same re-identification outcome [16]. By that interpretation, it must be demonstrated that non-experts *and* multiple experts can re-identify a data set before it can be considered personal information.

The most precise definition of de-identified information is provided in the Safe Harbor provision of the US HIPAA Privacy Rule [31]. This lists 18 specific data elements whose absence deem a data set de-identified. This precision has also resulted in the Safe Harbor list being used in Canada [32]. It has been estimated that following the Safe Harbor provision implies a 0.04% chance of re-identification in the US [33], but also that it results in significant information loss [34].

As can be seen, there are a number of varied approaches to deciding when a dataset is de-identified. We therefore make two practical recommendations.

The first recommendation is an obvious one. Depending on the jurisdiction of a custodian they have to follow a particular definition. For example, federal government departments in Canada will apply the “serious possibility” test to determine if information is de-identified. Covered entities in the US will follow the stipulations of HIPAA, and custodians in the EU will use the definitions in the Data Protection Directive. Legal counsel would then determine whether the test in that jurisdiction has been met.

Because there is quite a bit of subjectivity in most definitions, one can interpret them using a strong precedent. Many custodians, within and outside healthcare, for more than 20 years, have been using two thresholds to decide if a dataset is de-identified. These thresholds refer to minimal cell sizes, which mean the number of records that have the same values on the quasi-identifiers. One threshold that has been suggested and used is a minimal cell size of three in data sets that are disclosed [35-38]. Another more common value is a minimal cell size of five [39-48]. Because of the extensive use of these two thresholds over such an extended period of time, one can argue that these represent the risks that society has decided to accept when releasing sensitive personal information.

## 3 Overview of De-identification Techniques

---

In this section we will provide a general overview of each de-identification technique.

### 3.1 Randomization

This technique keeps the identifying variables in the dataset, but replaces the actual values with random values that look real. For example, we would replace the real names and addresses with fake names and addresses. The fake names and addresses would be taken from a large database of real Canadian/American names and addresses. This approach ensures that all of that information looks real (for example, if a randomization tool replaces a male name with a randomly selected name from its database, it will also be a male name) [6, 49].

A good randomization tool will select a name randomly with the same probability that it appears in the actual population of Canada/US. This ensures that very uncommon names do not appear disproportionately often in the de-identified data.

Randomization tools can also replace the values of health insurance numbers with fake numbers that will pass validation checks (e.g., the Luhn algorithm). Similarly, social insurance numbers, and credit card numbers can be randomized.

Various constraints can be placed on such value substitutions. For example, if a first name is substituted with a random one, a constraint can be imposed to ensure that the replacement name has the same number of characters or that it is of the same ethnicity as the original name. Also, for example, if a credit card number is substituted, it may be desirable to ensure that the replacement number is from the same financial institution or the same type of card.

More sophisticated randomization techniques will ensure the internal consistency within the data. For example, let there be two variables “Postal Code” and “Telephone Number”. If one of these is modified, say “Postal Code”, then a randomization system should also modify the area code for “Telephone Number” to make it is consistent with the new “Postal Code” value. That way, if someone examines the de-identified data it will look realistic throughout.

Another level of sophistication would ensure that the replacement names are likely in the geographic region. For example, if the “Postal Code” is changed to one with residents who are predominantly of Portuguese origin, the likelihood of having many non-Portuguese replacement names should be quite small.

For complex relational datasets, randomization needs to be performed carefully. For example, if there are two tables in a database that are related by a patient’s health insurance number, then we would not want to randomize one table in one way and in the second table to assign the same

patient a different random health number. Therefore, referential integrity across tables must be ensured where appropriate.

Randomization should be irreversible, but there are instances where this would not be the case. For example, if an audit trail of all substitutions is maintained, then it would be possible for any person with access to the audit trail to check later on what the original values were. Also, there are some randomization tools that will add, for example, a fixed constant to a date. This makes it quite easy to reverse engineer the original dates from the randomized ones. Therefore, audit trails should be protected or disabled.

There are a number of commercial tools available that implement randomization functions. Some are standalone in that they will work with a specific dataset that is imported. Others are enterprise level tools that can access complex relational datasets.

The main drawback with randomization techniques is that they do not provide any guarantees about the difficulty of reversing the replacements or substitutions that they perform: there is no real concept of re-identification risk. Therefore, the privacy analyst must make a judgment call as to whether randomizations ensure that the plausibility of re-identifying the data is sufficiently low.

### **3.2 Irreversible Coding**

One can irreversibly replace the identifying variables with a pseudonym [8, 50]. An irreversible pseudonym can be random or unique.

A random pseudonym will be different if it is generated multiple times for the same individual. For example, consider a data warehouse that is being asked to disclose datasets to researchers. A researcher may ask for a coded dataset *A* containing lab results on all patients admitted to a hospital in 2006. Then six months later the same researcher asks for a coded dataset *B* containing the demographics on all patients admitted in 2006. The REB would not have allowed the researcher to access the lab results plus demographics in the same dataset, but if the pseudonyms generated are the same, that researcher can link *A* and *B* to get the lab results and demographics on the same patients. The pseudonyms generated for dataset *A* would have to be different from those generated for dataset *B* to make such linking difficult to do. This example can be recast where you have different people in collusion asking for different datasets, or asking for the datasets at different points in time and then create cross-sectional or longitudinal records on the same patients. Therefore, when it is not desirable to allow the recipients to be able to access different coded datasets and link them, a random pseudonym must be generated anew for each individual each time the dataset is disclosed.

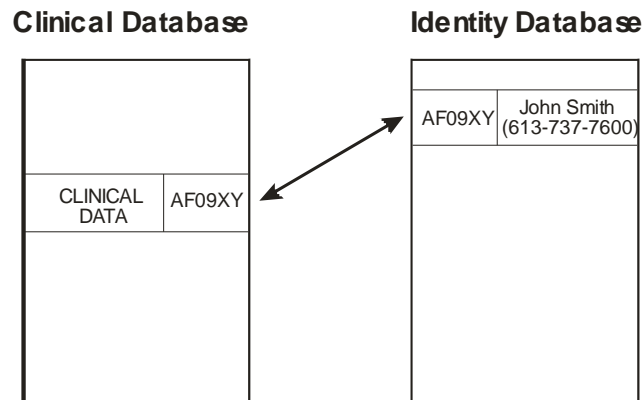
A unique pseudonym is the same when it is generated for the same patient. So patients in dataset *A* would have the same pseudonym if they appear in dataset *B*. There will also be occasions where one wishes to link multiple datasets. For example, multiple agencies may agree

to provide coded data to a researcher and permit the researcher to link the different datasets together. The agencies may use the same algorithm to generate the pseudonyms for the individuals so that the same individual in multiple datasets will have the same pseudonym. This way the researcher can perform anonymous linking of the datasets. Another example is where individuals need to be tracked longitudinally, and there will be multiple data disclosures over time. To facilitate the anonymous linking of the different datasets, it would be desirable to have the same pseudonym used for the same individuals over time.

### 3.3 Reversible Coding

When coding is reversible it allows individuals to be re-identified if necessary. This is also sometimes called reversible pseudonymization [8, 50]. Common reversible coding schemes that are used are single or double coding.

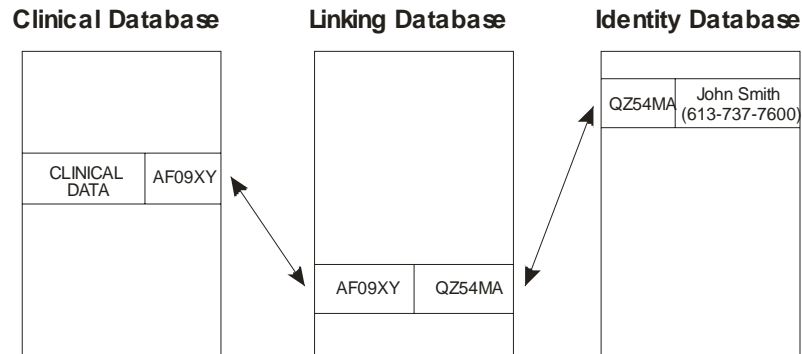
Single coded data means that identifiers are removed from the dataset and each record is assigned a new code (a pseudonym). Identifiers are kept in a different dataset with the pseudonym to allow linking back to the original data. This is illustrated in the clinical example of Figure 3. Here the value “AF09XY” is the pseudonym. The identity database would normally be kept separate from the clinical database with different access control permissions where only specifically authorized individuals would have access.



**Figure 3:** A clinical example showing single coding.

If there is ever a need to re-identify a patient in the clinical database, then the party holding the identity database is contacted and asked to reveal the identity of the individual with a particular pseudonym. The procedure for such re-identification would have to be decided upon in advance before data collection commences.

Double coded data means that the pseudonyms associated with the original data and the identity database are different, and the information linking them is kept in a separate linking database. The linking database is maintained in a secure location by an authorized party (also referred to as a trusted third party), for example. This is illustrated in the clinical example of Figure 4.



**Figure 4:** A clinical example showing double coding through a linking database.

Again, in this scenario if there is ever a need to re-identify a patient, both the authorized party and the party holding the identity database would be asked to facilitate the linking of the three tables. Double coding is useful where one wants to protect against collusion between the holder of the clinical database and the holder of the identity database.

Note that the same points above on unique versus random pseudonyms apply to reversible coding as well.

### 3.4 Heuristics

In general there are two types of heuristics that are in use: (a) those based on uniqueness and rareness in the population, and (b) those based on record linkage experiments using public registries.

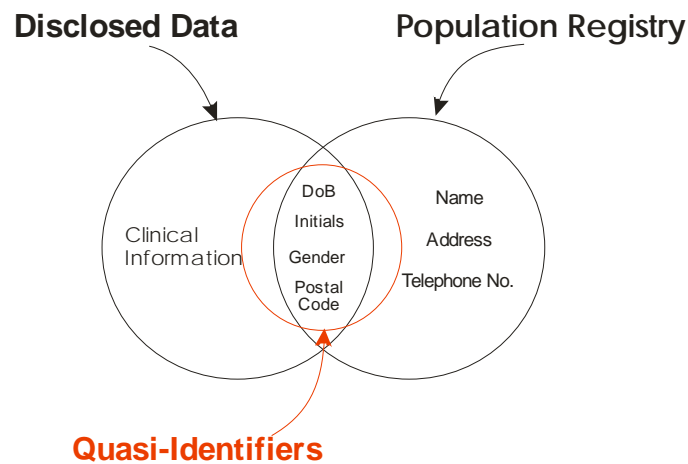
Members of the population who are unique on some characteristic are at a higher risk of re-identification. For example, if a dataset contains the two variables “Profession” and “City of Residence” and one person has the values “Mayor” and “Ottawa” respectively, then that person is unique. The uniqueness heuristics are intended to identify a priori what makes people unique and protect against that. Examples of uniqueness and rareness heuristics are:

- Geographic information shall not be released if there are less than 20,000 people living in the geographic area (this comes from US Health Insurance Portability and Accountability

Act – HIPAA – Privacy Rule, but similar heuristics are used by Statistics Canada and the Census Bureau) [7].

- If the quasi-identifier values represent less than 0.5% of the population, then these values represent a rare group and response values on the quasi-identifiers should be generalized or the observations in the dataset suppressed. This heuristic is used by Statistics Canada to top code age at 89+, for example.
- There are a number of rare (prevalence of, say, less than 10 in 10,000) and visible diseases and conditions [51]. If a diagnostic code representing one of these is included in the dataset, then that value (or record) should be suppressed.

One can attempt to match records in the released dataset with records from a *population registry*. The matching would have to be made on the basis of quasi-identifiers.



**Figure 5:** Illustration of how the disclosed data can be linked with a population registry.

Figure 5 shows how record linkage would occur. Because the dataset and the population registry have the quasi-identifiers in common (in this example they are the date of birth, initials, gender, and postal code), then one could match the records in both datasets. The disclosed dataset does not have any identifying information, but the population registry 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 the disclosed data and re-identify them.

Record linkage can be exact or approximate (e.g., two approximate methods are probabilistic and distance based). Approximate record linkage methods can be quite powerful as they can deal with data quality problems (e.g., due to data entry errors or missingness).

Studies have shown that population registries can be created relatively easily using publicly available registries [6]. In Canada these include the Private Property Security Registration, the Land Registry, the white pages telephone directory, and membership lists of professional associations (e.g., the College of Surgeons and Physicians of Ontario and the Law Society of Upper Canada). Record linkage heuristics are based on experiments done on re-identifying datasets using such population registries. For example, we know that if we have the date of birth, gender, and profession of an individual, the probability of re-identifying that individual through record linkage with public registries would be quite high.

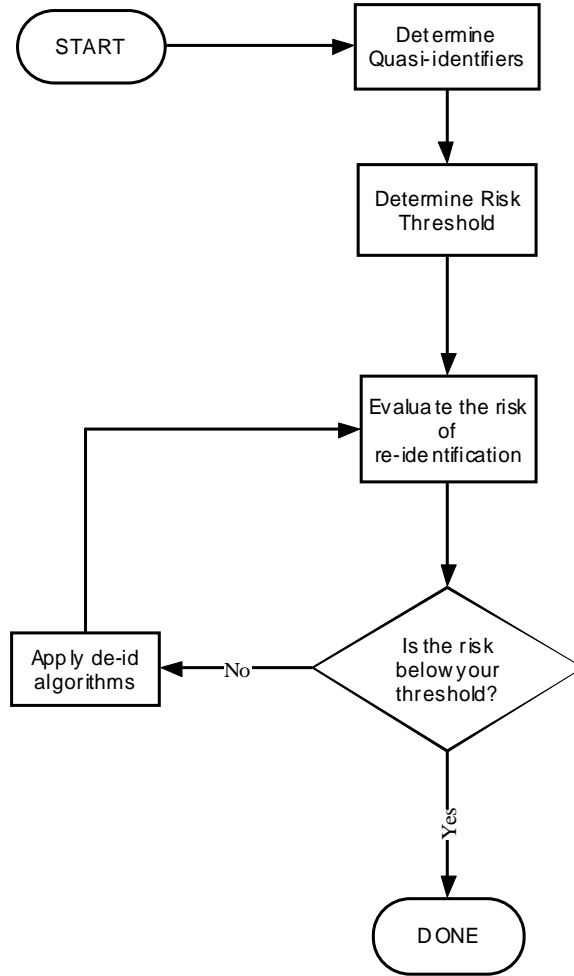
One important thing to note about heuristics is that they, ideally, should be evidence-based. Often rules-of-thumb about how to ensure that datasets are de-identified have been handed down over time without a clear continuing empirical basis for where they came from. When these traditional heuristics are evaluated, often they turn out to be too stringent or do not provide much meaningful protection. Therefore, it is always advisable to enquire about and assess the empirical basis for any heuristics that are being applied.

### **3.5 Analytics**

With *analytics* type de-identification techniques, it is assumed that the full data exists and the objective is to de-identify it. The two main principles of dataset de-identification are that (a) it needs to be risk based, and (b) it needs to take account of the utility/value of the data. The former means that the analyst must be able to evaluate the risk of re-identification, either quantitatively or qualitatively. Once the risk of re-identification is known, then decisions can be made about how and how much to de-identify the data. The latter implies that the amount of distortion to the dataset should be minimal because the more the data is distorted the less useful it will be for a recipient.

The general process for de-identifying datasets is shown in Figure 6. The first step is to determine the quasi-identifiers. The list in Table 2 is a good starting set to consider.

If the recipient is known to hold other databases, then the basic quasi-identifier set may be extended to include other variables known to be in those private databases. The criterion to use is: if the recipient were to link the private database with the disclosed data would there be significant matches ?



**Figure 6:** General steps to follow for de-identifying actual datasets.

It is also important to note that some quasi-identifiers can be inferred from other variables in the data. For example, age can be inferred from graduation year, date of death from autopsy date, baby’s date of birth from mother’s discharge date, and disease or condition from drug taken. Inferred quasi-identifiers should also be included in the de-identification analysis.

The next step is to determine the risk threshold. The risk threshold should be quantitative, and reflect the amount of re-identification risk that the custodian is willing to take. There are two situations to consider here.

The first situation to consider is when the risk threshold is standardized, as in the case of access to information requests. Here, the custodian has a fixed threshold to be used for all data disclosures (i.e., the threshold is not contingent) and would apply it to all access requests irrespective of who has made the request or the characteristics of the data. For example, in the context of the Canadian federal government, this would be a threshold based on an interpretation of the “serious possibility” test.

There will be other situations, however, where the custodian can adopt a context dependant threshold. Three general criteria that can be used to determine the risk threshold are [52, 53]: (a) the extent to which there would be an invasion-of-privacy if the personal information was inappropriately disclosed or processed by the recipient, (b) the motives and capacity of the recipient to re-identify the data if it was given to them in de-identified form, and (c) the extent to which the recipient will have good security and privacy practices in place. This is not necessarily a comprehensive set of criteria, but they reflect some of the practices that are in place today to manage re-identification risk.

A simple way to de-identify a dataset is to sub-sample. This means that the custodian only discloses a random sample of the data that is requested. This approach, however, provides very limited protection against certain re-identification techniques and should not be used as the sole approach.

More sophisticated algorithms have been developed to de-identify a dataset to ensure that the actual risk is below the threshold. Most such algorithms also ensure that the amount of distortion to the data is minimal. These algorithms are implemented as software tools, and there are a number available commercially and possibly from academic research groups. The most common software tools implement the k-anonymity de-identification criterion. This criterion ensures that there are at least k records in the dataset that have the same values on the quasi-identifiers for every combination of values. For example, if the quasi-identifiers are age and gender, then it would ensure that, say, there are at least k records with “50, male” values. An example of a 3-anonymized dataset is shown in Figure 7.

Admission Date	Gender	Age
01/01/2008	M	18
01/01/2008	M	17
01/01/2008	M	18
01/01/2008	M	13
01/01/2008	M	19
02/01/2008	F	18
02/01/2008	F	22
02/01/2008	F	23
02/01/2008	F	21
01/01/2008	M	22

**(a)**

Admission Date	Gender	Age
01/01/2008	M	15-19
01/01/2008	M	15-19
01/01/2008	M	15-19
<del>01/01/2008</del>	<del>M</del>	<del>10-14</del>
01/01/2008	M	15-19
<del>02/01/2008</del>	<del>F</del>	<del>15-19</del>
02/01/2008	F	20-24
02/01/2008	F	20-24
02/01/2008	F	20-24
01/01/2008	M	20-24

**(b)**

**Figure 7:** This example illustrates 3-anonymity on a three quasi-identifier dataset. The original dataset (a) is converted into a 3-anonymous dataset by generalizing the age into a range. The crossed-out records have to be suppressed in the (b) dataset to ensure 3-anonymity.

Tools that implement the k-anonymity criterion often use generalization and suppression as the specific mechanisms to achieve k-anonymity. Generalization means reducing the precision of a variable, and suppression means replacing an actual value in the data by a missing value. Data recipients tend to be sensitive to excessive suppression because it limits what can be done with the data. Therefore suppression should be minimized or other techniques used to impute the missing values with estimates. But again, in practice data custodians are often concerned about imputed values because they are not “true” values.

## 4 Further Considerations

---

In this section we will address some practical issues and questions that come up during the application of de-identification techniques. They are not intended to be comprehensive, but only the ones we found to cause uncertainty or confusion.

### 4.1 Attribute Disclosure

The de-identification techniques described above pertain to *identity disclosure*. This means the risk being mitigated is that of identifying the individual(s) associated with one or more records in the dataset. However, it is also important to consider *attribute disclosure*. This kind of disclosure occurs when one can discover something (an attribute) about individuals without identifying their specific record in the dataset. For example: if all females between 50 and 55 in a dataset have breast cancer as a diagnosis, then it does not matter which record belongs to any particular 50-55 year old female since the breast cancer diagnosis will be true for everyone. As another example, if there is a survey of drug use by teenagers, and all respondents in a particular postal code admitted to drug use, then a parent living in that postal code can infer that their child was a drug user if it is known that their child responded to the survey.

There has been research work to develop algorithms that would protect against attribute disclosure, but there are limited options for software tools currently available. This, of course, may change on the future.

### 4.2 Membership in a Database

In some cases even membership in a database may be personal information. For example, if a database of individuals receiving HIV treatment is disclosed, then mere membership of that database is sensitive information even if the records do not contain any clinical data. One solution in such cases is to combine multiple databases together so that it would not be known which database a record originated from.

### 4.3 Audit Trails

Many systems maintain extensive audit trails. For example, electronic systems used in clinical trials must record the user identity for every insertion or modification of a record in an electronic case report form. Similarly, systems used in providing health care services maintain access audit trails to detect privacy breaches.

In many cases the audit trails will contain identity as well as the sensitive health information. Therefore, if any dataset containing audit trails will be disclosed, the audit trails themselves may also need to be de-identified.

#### **4.4 Residence Trails**

Longitudinal databases that record residence information about individuals present some challenges. This is because the movement of individuals can make them uniquely identifiable. For example, over a ten year period, a person who has lived in five locations may have a unique trail of postal codes because out of the whole population he or she is the only one who has moved to these locations at these points in time.

Therefore, whenever longitudinal data will be disclosed, consideration should be given to residence trails that may be unique in the dataset.

#### **4.5 Encounter Trails**

The dates that patients visit their physician or get a laboratory test represent an encounter profile that can make patients unique. This is similar to residence trails but the risk pertains to encounter dates. For example, over a 5 year period the visits of a particular person to their family doctor have a particular pattern that is very likely to be uniquely identifiable. The more visits the more unique that pattern is likely to be. This risk is likely to increase in prominence as more electronic medical record data is disclosed for secondary purposes.

#### **4.6 Number of Residences and Encounters**

Individuals who have moved quite often and therefore their residence trails are larger than anyone else, or who have had a relatively large number of encounters may be at a higher risk of re-identification because they stand out. So not only do the actual values represent a risk, but the number of instances needs to also be considered.

The number of instances can also extend to other types of information. For example, patients who have a relatively large number of prescribed medications or a relatively large number of diagnoses would stand out.

#### **4.7 Unstructured Data**

The assumption thus far has been that the data is structured. However, in many cases unstructured data, such as text in clinical notes and comments, will also be disclosed. Text may contain directly identifying information as well as sensitive (clinical) information. Text is more difficult to de-identify because tools need to be specific to the type of text (e.g., de-identification of discharge summaries will be different from progress notes), but there are tools available that can provide some automated support.

In general, if disclosures will include text, consideration should be given to how that text will be de-identified. If no attempt is made to de-identify text then that information should be removed from the dataset.

## **4.8 Other Media**

In some cases electronic files containing other media, such as audio and video, will be disclosed. Again, these represent special challenges for de-identification, and the most appropriate approach for addressing privacy concerns in such circumstances will be case dependent.

## **4.9 Data Quality**

All of the techniques discussed this far assume that the data is of high quality. This will not necessarily be the case. Prospective registries where data is being collected for analysis and research tend to have higher quality data. Data used for providing care and coming directly from EMRs tends to have a significant amount of quality problems (e.g., misspellings, data entered in the wrong fields, different ways of referring to the same thing, and data entry errors). Therefore, when planning a de-identification effort, thought should be put into the need for data cleansing beforehand.

## 5 Acronyms

---

CIHR	Canadian Institutes of Health Research
EHR	Electronic Health Record
HIA	Health Information Act
HPPA	Health Protection and Promotion Act
PHI	Personal Health Information
PHIA	Personal Health Information Act
PHIPA	Personal Health Information Protection Act
PI	Personal Information
REB	Research Ethics Board
RHAA	Regional Health Authority Act
THIPA	The Health Information Protection Act

## **6 Acknowledgements**

---

We wish to thank Mary Lysyk and Pat Milliken for their extensive comments and feedback on an earlier version of this report. The contents of this report are the sole responsibility of the authors and do not represent positions of Health Canada.

## 7 References

---

1. Safran C, Bloomrosen M, Hammond E, Labkoff S, S K-F, Tang P, Detmer D. *Toward a national framework for the secondary use of health data: An American Medical Informatics Association white paper*. Journal of the American Medical Informatics Association, 2007; 14:1-9.
2. Perun H, Orr M, Dimitriadis F. *Guide to the Ontario Personal Health Information Protection Act*. 2005: Irwin Law.
3. El Emam K, King M. *The data breach analyzer*. 2009; Available from: [\[http://www.ehealthinformation.ca/dataloss\]](http://www.ehealthinformation.ca/dataloss).
4. Willison D, Emerson C, Szala-Meneok K, Gibson E, Schwartz L, Weisbaum K. *Access to medical records for research purposes: Varying perceptions across Research Ethics Boards*. Journal of Medical Ethics, 2008; 34:308-314.
5. El Emam K, Dankar F, Issa R, Jonker E, Amyot D, Cogo E, Corriveau J-P, Walker M, Chowdhury S, Vaillancourt R, Roffey T, Bottomley J. *A Globally Optimal k-Anonymity Method for the De-identification of Health Data* Journal of the American Medical Informatics Association, 2009.
6. El Emam K, Jonker E, Sams S, Neri E, Neisa A, Gao T, Chowdhury S. *Pan-Canadian De-Identification Guidelines for Personal Health Information (report prepared for the Office of the Privacy Commissioner of Canada)*. 2007; Available from: [\[http://www.ehealthinformation.ca/documents/OPCReportv11.pdf\]](http://www.ehealthinformation.ca/documents/OPCReportv11.pdf). Archived at: [\[http://www.webcitation.org/5Ow1Nko5C\]](http://www.webcitation.org/5Ow1Nko5C).
7. El Emam K, Brown A, Abdelmalik P. *Evaluating Predictors of Geographic Area Population Size Cutoffs to Manage Re-identification Risk*. Journal of the American Medical Informatics Association, 2009; 16(2):256-266.
8. ISO/TS 25237. *Health Informatics: Pseudonymization*. 2008.
9. Hansell S. *AOL Removes Search Data on Group of Web Users* in *New York Times*. 2006: 8 August.
10. Barbaro M, Zeller Jr. T. *A Face Is Exposed for AOL Searcher No. 4417749* in *New York Times*. 2006: 9 August.
11. Zeller Jr. T. *AOL Moves to Increase Privacy on Search Queries*, in *New York Times*. 2006: August 22.
12. Ochoa S, Rasmussen J, Robson C, Salib M. *Reidentification of individuals in Chicago's homicide database: A technical and legal study*. 2001; Massachusetts Institute of Technology.
13. Narayanan A, Shmatikov V. *Robust de-anonymization of large datasets (how to break anonymity of the Netflix prize dataset)*. 2008; University of Texas at Austin.
14. Sweeney L. *Computational disclosure control: A primer on data privacy protection*. 2001, Massachusetts Institute of Technology.
15. Appellate Court of Illinois - Fifth District. *The Southern Illinoisan v. Department of Public Health*. 2004.
16. The Supreme Court of the State of Illinois. *Southern Illinoisan vs. The Illinois Department of Public Health*. 2006.

17. Federal Court (Canada). *Mike Gordon vs. The Minister of Health: Affidavit of Bill Wilson*. 2006.
18. El Emam K, Dankar F. *Protecting privacy using k-anonymity*. Journal of the American Medical Informatics Association, 2008; 15:627-637.
19. El Emam K. *Heuristics for de-identifying health data*. IEEE Security and Privacy, 2008:72-75.
20. *Federal Data Protection Act (Germany)*. 2006.
21. Article 29 Data Protection Working Party. *Opinion 4/2007 on the concept of personal data: Adopted on 20th June. 2007*; Available from: [[http://ec.europa.eu/justice\\_home/fsj/privacy/docs/wpdocs/2007/wp136\\_en.pdf](http://ec.europa.eu/justice_home/fsj/privacy/docs/wpdocs/2007/wp136_en.pdf)]. Archived at: [<http://www.webcitation.org/5Q2YBu0CR>].
22. El Emam K, Kosseim P. *Privacy Interests in Prescription Records, Part 2: Patient Privacy*. IEEE Security and Privacy, 2009; 7(2):75-78.
23. Canadian Institutes of Health Research. *Recommendations for the Interpretation and Application of the Personal Information Protection and Electronic Documents Act (S.C.2000, c.5) in the Health Research Context Canadian Institutes of Health Research*. 2001; Available from: [[http://www.cihr-irsc.gc.ca/e/documents/recommendations\\_e.pdf](http://www.cihr-irsc.gc.ca/e/documents/recommendations_e.pdf)].
24. Canadian Institutes of Health Research. *Background Legal Research and Analysis in Support of CIHR's Recommendations with Respect to the Personal Information Protection and Electronic Documents Act (PIPEDA) (S.C. 2000, c. 5)*. 2001; Available from: [[http://www.cihr-irsc.gc.ca/e/documents/legal\\_analysis\\_e.pdf](http://www.cihr-irsc.gc.ca/e/documents/legal_analysis_e.pdf)].
25. *Commission regulation (EC) No 831/2002 of 17 May 2002 on implementing council regulation (EC) No 322/97 on community statistics, concerning access to confidential data for scientific purposes*. Official Journal of the European Communities, 2002.
26. Beach J. *Health care databases under HIPAA: Statistical approaches to de-identification of protected health information*. DIMACS Working Group on Privacy/Confidentiality of Health Data. 2003.
27. American Public Health Association. *Statisticians and de-identifying protected health information for HIPAA*. 2005; Available from: [<http://www.apha.org/membersgroups/newsletters/sectionnewsletters/statis/fall05/2121.htm>].
28. Brownlee C, Waleski B. *Privacy Law*. 2006: Law Journal Press.
29. *Mike Gordin and the Minister of Health and the Privacy Commissioner of Canada: Memorandum of Fact and Law of the Privacy Commissioner of Canada*. 2007; Federal Court.
30. Long M, Perrin S, Brands S, Dixon L, Fisher F, Gellman R. *Privacy enhancing tools and practices for an electronic health record (EHR) environment: Phase 2 of a research report for Health Canada's Office of Health and the Information Highway*. 2003; Health Canada.
31. Pabrai U. *Getting Started with HIPAA*. 2003: Premier Press.
32. El Emam K. *Data Anonymization Practices in Clinical Research: A Descriptive Study*. 2006; Health Canada, Access to Information and Privacy Division.
33. National Committee on Vital and Health Statistics. *Report to the Secretary of the US Department of Health and Human Services on Enhanced Protections for Uses of Health Data: A Stewardship Framework for "Secondary Uses" of Electronically Collected and Transmitted Health Data*. 2007.

34. Clause S, Triller D, Bornhorst C, Hamilton R, Cosler L. *Conforming to HIPAA regulations and compilation of research data*. American Journal of Health-System Pharmacy, 2004; 61(10):1025-1031.
35. Duncan G, Jabine T, de Wolf S. *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*. 1993: National Academies Press.
36. de Waal A, Willenborg L. *A view on statistical disclosure control for microdata*. Survey Methodology, 1996; 22(1):95-103.
37. Office of the Privacy Commissioner of Quebec (CAI). *Chenard v. Ministere de l'agriculture, des pecheries et de l'alimentation (141)*. 1997.
38. National Center for Education Statistics. *NCES Statistical Standards*. 2003; US Department of Education.
39. *Cancer Care Ontario Data Use and Disclosure Policy*. 2005; Cancer Care Ontario.
40. *Security and confidentiality policies and procedures*. 2004; Health Quality Council.
41. *Privacy code*. 2004; Health Quality Council.
42. *Privacy code*. 2002; Manitoba Center for Health Policy.
43. Subcommittee on Disclosure Limitation Methodology - Federal Committee on Statistical Methodology. *Working paper 22: Report on statistical disclosure control*. 1994; Office of Management and Budget.
44. Statistics Canada. *Therapeutic abortion survey*. 2007; Available from: <http://www.statcan.ca/cgi-bin/imdb/p2SV.pl?Function=getSurvey&SDDS=3209&lang=en&db=IMDB&dbq=f&adm=8&dis=2#b9>. Archived at: <http://www.webcitation.org/5VkcHLeQw>.
45. Office of the Information and Privacy Commissioner of British Columbia. *Order No. 261-1998*. 1998.
46. Office of the Information and Privacy Commissioner of Ontario. *Order P-644*. 1994; Available from: [http://www.ipc.on.ca/images/Findings/Attached\\_PDF/P-644.pdf](http://www.ipc.on.ca/images/Findings/Attached_PDF/P-644.pdf).
47. Alexander L, Jabine T. *Access to social security microdata files for research and statistical purposes*. Social Security Bulletin, 1978; 41(8):3-17.
48. Ministry of Health and Long Term care (Ontario). *Corporate Policy 3-1-21*. 1984.
49. El Emam K, Sams S. *Anonymization case study 1: Randomizing names and addresses*. 2007; Available from: <http://www.ehealthinformation.ca/documents/PACaseStudy-1.pdf>. Archived at: <http://www.webcitation.org/5OT8Y1eKp>.
50. Numeir R, Lemay A, Lina J-M. *Pseudonymization of radiology data for research purposes*. Journal of Digital Imaging, 2007; 20(3):284-295.
51. Eguale T, Bartlett G, Tamblyn R. *Rare visible disorders / diseases as individually identifiable health information*. Proceedings of the American Medical Informatics Association Symposium. 2005.
52. El Emam K. *De-identifying health data for secondary use: A framework*. 2008; Available from: <http://www.ehealthinformation.ca/documents/SecondaryUseFW.pdf>.
53. El Emam K, Dankar F, Vaillancourt R, Roffey T, Lysyk M. *Evaluating patient re-identification risk from hospital prescription records*. Canadian Journal of Hospital Pharmacy, 2009; 62(4):307-319.