Evaluating the Risk of Re-identification of Patients from Hospital Prescription Records

Khaled El Emam, Fida K Dankar, Régis Vaillancourt, Tyson Roffey, and Mary Lysyk

ABSTRACT

Background: Pharmacies often provide prescription records to private research firms, on the assumption that these records are de-identified (i.e., identifying information has been removed). However, concerns have been expressed about the potential that patients can be re-identified from such records. Recently, a large private research firm requested prescription records from the Children's Hospital of Eastern Ontario (CHEO), as part of a larger effort to develop a database of hospital prescription records across Canada.

Objective: To evaluate the ability to re-identify patients from CHEO'S prescription records and to determine ways to appropriately de-identify the data if the risk was too high.

Methods: The risk of re-identification was assessed for 18 months' worth of prescription data. De-identification algorithms were developed to reduce the risk to an acceptable level while maintaining the quality of the data.

Results: The probability of patients being re-identified from the original variables and data set requested by the private research firm was deemed quite high. A new de-identified record layout was developed, which had an acceptable level of re-identification risk. The new approach involved replacing the admission and discharge dates with the quarter and year of admission and the length of stay in days, reporting the patient's age in weeks, and including only the first character of the patient's postal code. Additional requirements were included in the data-sharing agreement with the private research firm (e.g., audit requirements and a protocol for notification of a breach of privacy).

Conclusions: Without a formal analysis of the risk of re-identification, assurances of data anonymity may not be accurate. A formal risk analysis at one hospital produced a clinically relevant data set that also protects patient privacy and allows the hospital pharmacy to explicitly manage the risks of breach of patient privacy.

Key words: privavy, de-identification, re-identification risk, data anonymity, secondary use of data

RÉSUMÉ

Contexte : Les pharmacies fournissent souvent des dossiers d'ordonnance aux firmes de recherche indépendantes, en supposant qu'ils sont dépersonnalisés (c.-à-d., que l'information pouvant identifier les patients a été retirée). Cependant, des inquiétudes ont été soulevées quant à la possibilité que l'on puisse reconstituer l'identité des patients à partir de ces dossiers. Récemment, une importante firme de recherche indépendante a demandé au Centre hospitalier pour enfants de l'est de l'Ontario (CHEO) d'obtenir les dossiers d'ordonnance, dans le cadre d'un projet plus vaste visant à développer une base de données pancanadienne des dossiers d'ordonnance hospitaliers.

Objectif : Évaluer la possibilité de reconstituer l'identité des patients à partir des dossiers d'ordonnance du CHEO afin de déterminer les moyens appropriés de dépersonnaliser les données si le risque de reconstitution est trop élevé.

Méthodes : Le risque de reconstitution de l'identité a été évalué à partir de données sur les ordonnances couvrant une période de 18 mois. Des algorithmes de dépersonnalisation ont été conçus pour réduire le risque à un niveau acceptable, tout en maintenant la qualité des données.

Résultats : La probabilité de reconstitution de l'identité des patients à partir des variables et des données originales demandées par la firme de recherche indépendante a été jugée assez élevée. Une nouvelle méthode de dépersonnalisation des dossiers comportant un niveau de risque de reconstitution de l'identité acceptable a été développée. La nouvelle méthode impliquait le remplacement des dates d'admission et de sortie par le trimestre et l'année d'admission et la durée du séjour en jours, l'expression de l'âge du patient en semaines, et l'insertion uniquement du premier caractère du code postal du patient. D'autres exigences ont été incluses dans l'entente de transmission de données avec la firme de recherche indépendante (p. ex., des exigences de vérification et un protocole de déclaration de violation de la vie privée).

Conclusion : En l'absence d'analyse structurée du risque de reconstitution de l'identité, il est difficile d'assurer la dépersonnalisation des données. Une analyse structurée du risque effectuée dans un hôpital a généré un ensemble de données pertinent sur le plan clinique qui protège également la confidentialité des renseignements personnels des patients et permet à la pharmacie de l'hôpital de gérer explicitement les risques de violation de la vie privée.

Mots clés : vie privée, dépersonnalisation, risque de reconstitution de l'identité, anonymat des données, utilisation secondaire des données

[Traduction par l'éditeur]

Can J Hosp Pharm 2009;62(4):307–319

INTRODUCTION

Many retail and hospital pharmacies across Canada disclose prescription data (referred to in this article as "prescription records") to private research firms. These firms use the records to produce reports on prescribing patterns and drug utilization[1] and to perform economic studies.[2] The reports are then sold primarily to the pharmaceutical industry and government agencies.

Each prescription record contains information about the prescriber and the patient, as well as the drug dispensed. It is clear that the prescribers can be identified in the prescription record. However, it has been argued that the patient information disclosed in such records is also sufficient to identify patients,[3-5] which jeopardizes the confidentiality of Canadians' health information.[3]

The Ontario Personal Health Information Privacy Act (PHIPA) defines identifying information as "information that identifies an individual or for which it is reasonably foreseeable in the circumstances that it could be utilized, either alone or with other information, to identify an individual".[6] If the prescription records cannot directly or indirectly identify patients, then there would be no legislative requirements or constraints on their disclosure, since they would not be considered personal health information, and there would be no legislative requirement to obtain patient consent to disclose the records to such a third party.

In mid-2008, a private research firm requested the pediatric prescription records of the Children's Hospital of Eastern Ontario (CHEO), located in Ottawa, Ontario. This was part of a larger effort to develop a hospital prescription record database across Canada.[7] At the time the firm contacted CHEO, 100 hospitals had already agreed to be part of the program,[7] and 18% of the hospital beds in Ontario were already represented.[2] In return for participation in the database, the private research firm agreed to provide benchmarking capability to the participating hospitals and to allow them to use the date to conduct province-wide and national studies on drug utilization and effectiveness.

The prescription records requested from the hospital pharmacies were believed to be de-identified. The reasoning was that no directly identifying information about the patients, such as name and address, was being collected.[7] However, data elements other than the name and address can be used to re-identify an individual. There are well-documented examples in which individuals have been re-identified from ostensibly anonymous data, with no names and no street address information.[8-16] Although children may be less susceptible to re-identification than adults, the risk to this age group can still be high. Consequently, CHEO was aware that it could not assume that the data collected by the private research firm could not, under certain plausible scenarios, be used to re-identify CHEO patients.

The hospital's administration therefore requested that a study be conducted to ensure that any and all patient information be adequately de-identified before being disclosed to the private research firm. At the same time, the hospital recognized that the risk of re-identifying patients had to be balanced against the need to perform meaningful analysis and gain the benefits of benchmarking.

Eighteen months' worth of prescription records were used to evaluate whether the risk of re-identification was sufficiently low to justify the claim that the requested records were de-identified. The analysis was used to develop recommendations on the appropriate disclosure of prescription records to the private research firm.

METHODS

Type of Re-identification Risk

To fulfill the request from the research firm, the hospital pharmacy would have to disclose prescription records to an external party. Because of concerns about patient privacy, the pharmacy had to ensure that patient information in the disclosed database was appropriately de-identified. Various degrees of de-identification may be applied in such a situation: too much de-identification may diminish the clinical utility of the data, but too little de-identification may lead to a breach of privacy.

A risk analysis was performed to determine the level of de-identification to be applied. A meaningful risk analysis requires an understanding of the nature of plausible re-identification scenarios.

An individual or entity that attempts to re-identify the records in a database, either accidentally or deliberately, is called an intruder. It is assumed that the intruder will somehow gain access to the disclosed prescription database. This access may be legitimate (e.g., if the intruder is employed by the private research firm or works at the hospital). Alternatively, the intruder may find a prescription database that has been lost (e.g., database on an unencrypted laptop or memory stick that was forgotten at an airport), the intruder may deliberately steal the database (with about three-quarters of data loss incidents being caused by sources other than the data custodian[17,18]), or disclosure may be compelled in a criminal or civil court case.[19,20] An intruder may inadvertently re-identify a patient (e.g., an epidemiologist may spontaneously recognize a particular record while analyzing the database) or may deliberately attempt re-identification (e.g., in a court case where the intruder is an expert witness demonstrating that individuals in a database can be re-identified[14,15]).

For CHEO, 2 kinds of re-identification were of concern. The first type of re-identification, called identity disclosure, occurs if the intruder is able to assign a particular identity to any record in the prescription database; for example, the intruder determines that record number 7 in the prescription database belongs to patient Alice Smith. The second type of re-identification, called attribute disclosure, occurs when an intruder learns something new about a patient in the database without knowing which specific record belongs to that patient. For example, if all 20-year-old female patients in the disclosed database who live in a particular area had a prescription for an antidepressant, then if the intruder knows that Alice Smith is 20 years old and lives in that particular area, he or she will learn that Alice Smith was taking an antidepressant, even if the particular record belonging to Alice Smith is unknown.

The focus of the analysis reported in this article was on assessing and preventing identity disclosure, i.e., ensuring that an intruder would not be able to determine the identity associated with any record in the prescription database.

To re-identify a specific patient, the intruder must have some background information about that person, which the intruder then uses to look for the person's specific record in the database. The risk of re-identification by this means is termed "prosecutor re-identification risk".[21] Variables representing a patient's background information that is already known to the intruder are called quasi-identifiers. Examples of these quasi-identifiers are age, sex, postal code, ethnicity, race, profession, and main language spoken. An intruder who is a neighbour of the specific patient would know such details through his or her personal association with the patient. Alternatively, the background information of a famous person who is represented in the database would be available to the intruder through the public domain.

An intruder might also have background information about many patients and might attempt to re-identify any one of them, rather than targeting one specific person. In this situation, the re-identified patient is assumed to have been randomly selected. The risk of re-identification by this means is called "journalist re-identification risk".[21] In this case, the intruder needs an external database, known as an identification database,[22] against which to compare the prescription database. In effect, the identification database contains background information about many patients. Such a database can be constructed from public registries.[22] For patients who are youth (generally 18 years of age or younger), there are few publicly available and easily accessible government databases (federal, provincial, or municipal) containing pertinent quasi-identifiers, since they do not own property, borrow money, have telephones in their own names, or vote.[22] However, the membership of sports teams is often publicly available (e.g., an Internet search using the term "youth roster birth" will generate lists of sports teams, along with dates of birth), and many of these lists contain detailed demographic information about the team members. Furthermore, youth increasingly reveal basic demographic information about themselves on blogs and social networking websites, such as Facebook.[23,24] Therefore, the Internet has made it easier to construct identification databases about youth from public sources.

If the whole population of patients is considered, then prosecutor risk and journalist risk are quantitatively the same.[21] The analysis presented in this paper involved the complete population of patient visits to CHEO during which prescriptions were dispensed; therefore, the analysis focused on prosecutor risk.

Risk Assessment Methodology

A flow chart for the risk assessment methodology used in this study is shown in Figure 1, and the activities are described in more detail below.

Determine Quasi-identifiers

Some of the key fields requested by the private research firm are listed in Table 1. The quasi-identifiers used in the risk assessment were selected from this list of variables.

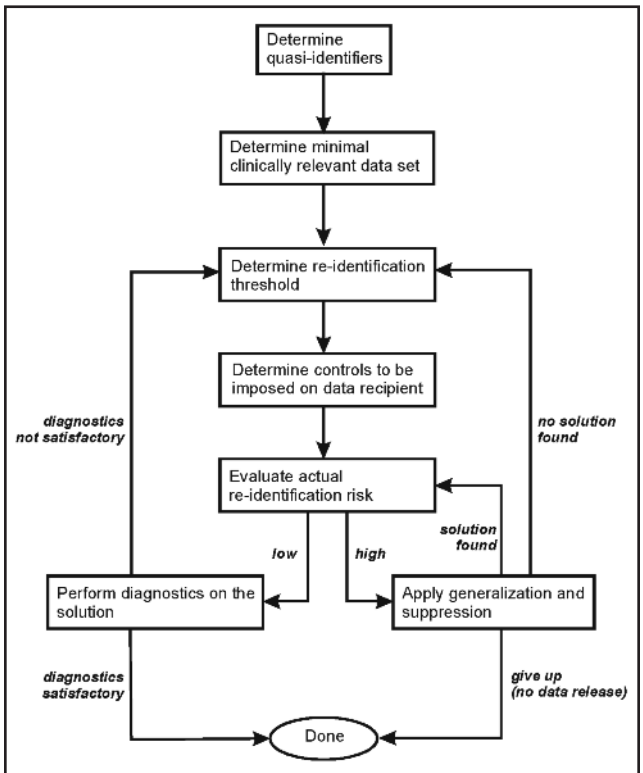

Figure 1. Overview of the risk assessment methodology used in this study. If no additional controls can be imposed and no solution is found, then the data custodian will give up, and the data will not be released.

Table 1. Some of the Original Data Fields Requested from Hospital Pharmacies by the Private Research Firm

Data Field	Description
Diagnostic data	
Diagnosis code	Primary condition that caused the patient to be registered (not the same as the ".Primary Problem." field in NACRS)
Diagnosis code version	ICD version of the diagnosis code (ICD-8, ICD-9, ICD-9CM, ICD-10)
Summary stage level	Summary cancer stage level, based on clinical and pathological stage values (for cancer patients only)
Data regarding hospital stay	
Patient's age	Patient's age (in years)
Patient's sex	Patient's sex
Patient's FSA	Patient's FSA
Admission date	Date on which the patient visited or was admitted to the facility
Discharge date	Date on which patient was discharged
Service cost centre	Medical or surgical service or clinic for cost assignments
Data regarding drug therapy	
Drug description	Text-based description of the product, as labelled for use
Drug code	Health Canada Drug Identification Number (DIN), if available
Dose transaction date	Date of this drug delivery event (equal to scheduled start date)
Dose administered	Actual dose administered during this visit
Measurement unit	Units of measurement for numeric dose given
Instructions for administration	Doses per day, if applicable; PRN, if not scheduled; dose, route, number of tablets; other specific instructions
Schedule	bid, tid, etc. (default: qd)
Route	Route to administer this drug (e.g., intravenous)
Transaction items	Number of items dispensed (should be a multiple of transaction doses)
Transaction doses	Number of doses dispensed
Transaction cost	Total cost of dispensed items
Dose cost	Cost per dose as prescribed (optional or calculated from transaction cost/transaction doses)
Days supplied	Number of 24-h periods of drug use supplied (i.e., unit dose = 1); if not provided, this is calculated from schedule and doses
Regimen	For cancer therapies and other therapies where concomitant and coordinated use is determined by defined regimens
Scheduled stop date	Scheduled stop order date
Actual stop date	Actual date of discontinuation; may be calculated from doses supplied if not renewed or refilled
Therapeutic intent	Reason for use, explicit or implicit (optional)
Location	Facility ward or department where drug was administered
Service cost centre	Medical or surgical service or clinic for cost assignments; may be same as facility admitting service above
Prescriber group	Facility-specific grouping code, if different from cost centre (optional; for future use)

FSA = forward sortation area (a geographic region designated by Canada Post, in which all postal codes start with the same 3 characters), ICD = International Classification of Diseases (8th, 9th, or 10th revision), NACRS = National Ambulatory Care Reporting System, PRN = as required.

Under the re-identification scenarios considered here, an intruder would need to have background information about a patient to re-identify him or her. For example, if the neighbour is defined as the archetype intruder, the variables in Table 1 that the neighbour would know are age, sex, and forward sortation area (FSA, a geographic region in Canada where the first 3 characters of the postal code are the same for all residents). It is also relatively straightforward to get this kind of information from public registries and then to construct an identification database. A neighbour may also know roughly when the patient was admitted and/or discharged and the approximate duration of stay. The full set of quasi-identifiers that were included in the current analysis are listed in Table 2.

Determine Minimal Variable Set

A technique often used to de-identify a data set is generalization.[25] Generalization means making the variables less precise. This is illustrated in Figure 2 with a generalization hierarchy for the length-of-stay variable. The greatest precision occurs at the bottom of the hierarchy, where length of stay is represented in days. The next higher level of generalization results in length of stay represented in weeks. Beyond that, lower-precision repre-

Table 2. Quasi-identifiers Included in the Analysis

Quasi-identifier	Generalization Hierarchy	Maximum Acceptable Generalization
Sex	NA	Required to assess disease prevalence by sex. However, very few drug therapies are affected by the patient's sex. Therefore, it is not absolutely critical that this variable be included in the disclosed database. No generalization is possible since there are only 2 possible values.
Age	Years Months Weeks Days	A critical variable, but exact date of birth is not necessary. Resolution of age data must provide the capacity to differentiate among age groups (for example, by weeks, months, or years). In a pediatric setting, it is important to distinguish small age differences among the very young, as organ maturity changes rapidly in the early stages of life, before reaching a plateau. Weeks of age was considered the most acceptable level of generalization for children up to 1 year; for older children, age in years was considered acceptable.
Postal code	1 character 2 characters 3 characters	Some geographic information is needed to assess regional distribution of burden of disease. Region (indicated by first character of postal code) was deemed acceptable.
Admission and discharge dates	Year Quarter/year Month/year Day/month/year	Required to determine length of stay, which is needed to assess effectiveness of drug therapy or disease burden on hospital system. Specific dates also needed to evaluate changes in therapy over time and seasonal variation. The maximum acceptable generalization was the quarter and year of admission, to maintain seasonal information. Because prescription data would be supplied on a quarterly basis, it would be possible to infer the quarter and year of admission for short-stay patients even if this field were not provided. Length of stay can be computed directly and used as a separate variable (see below). Date of discharge would not be critical in this case.
Length of stay	Weeks Days	Maximum acceptable generalization would be days.

NA = not applicable.

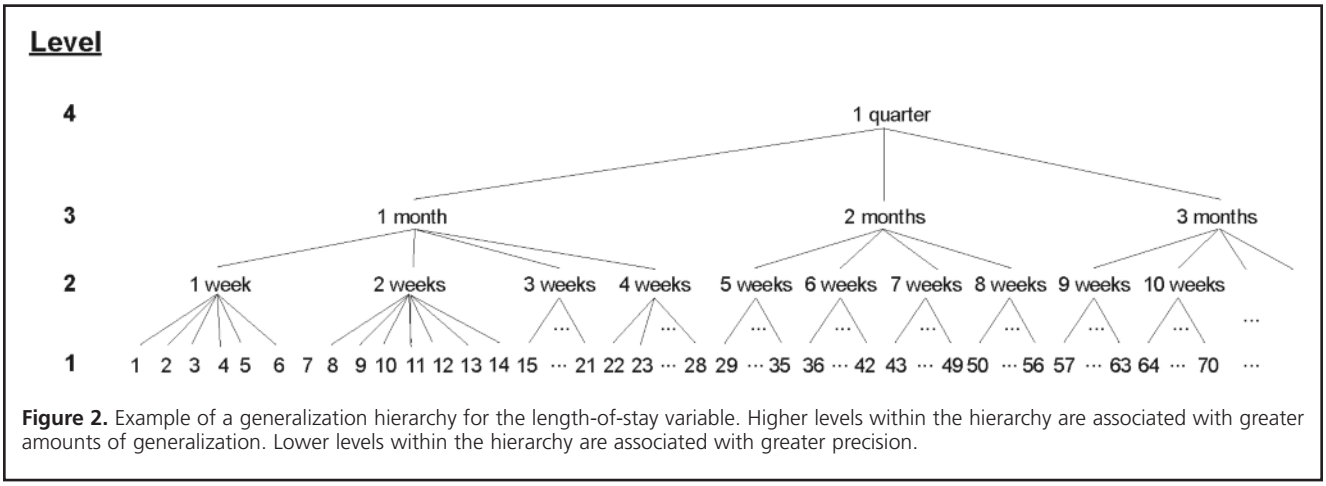

Figure 2. Example of a generalization hierarchy for the length-of-stay variable. Higher levels within the hierarchy are associated with greater amounts of generalization. Lower levels within the hierarchy are associated with greater precision.

sentations of length of stay would be in months and then quarters. Therefore, higher levels of the generalization hierarchy are associated with lower precision of the length-of-stay variable.

As data become increasingly generalized at higher levels of the hierarchy, they become less useful for data analysis. For example, if the precision of the length-of-stay variable in Figure 2 is generalized to quarters rather than days, then any effect that occurs during short stays would not be detectable. Therefore, there is a tradeoff between the amount of generalization and the utility of the data for analysis.

A consensus was reached among the key stakeholders in this data-release decision (i.e., the hospital's director of pharmacy and chief privacy officer and the private research firm) whereby a maximum amount of acceptable generalization was determined for each of the quasi-identifiers, as presented in Table 2. According to this scheme, the generalization during the de-identification process was not to reduce the precision of the variables beyond what appears in Table 2.

Evaluate Actual Risk of Re-identification

The risk of re-identification was measured as the probability that an intruder would find the correct identity of a single record.

Consider the example in Figure 3, whereby the intruder knows that Alice Smith was admitted to hospital and therefore knows that her data are present in the disclosed prescription database. As background information, the intruder also knows that Alice was born in 1987. To re-identify Alice's record the intruder needs to find all females that were born in 1987. However, because age was generalized to decade of birth, the intruder needs to find all records for females born in the range 1980-1989. If there are *f* matching records, then the probability of correct re-identification of Alice's record is 1/*f*. In the example shown in Figure 3, there are 2 matches, and the probability of correct re-identification is therefore 0.5.

The records with the same combination of values for the 2 quasi-identifiers mentioned above are called an equivalence class. In this example, the equivalence class consisted of all females born in the decade 1980–1989, and the size of the equivalence class was 2.

It is not known a priori if the intruder will try to re-identify Alice Smith or any one of the other patients. For example, if the intruder had background information about John Smith, who was born in 1979, then the equivalence class {male, 1970–1979} has a size of 3, and the probability of re-identification would be 0.33. A worst-case assumption must therefore be made: that the patient the intruder is trying to re-identify is in the smallest equivalence class in the database. In the example in Figure 3, the smallest equivalence class has a size of 2. This can be expressed more formally, as follows.

The probability of re-identifying a patient in the smallest equivalence class in the database represents the overall risk for the database. Let there be $j = 1 \ldots J$ equivalence classes in the data, where J is the number of equivalence classes in the database, and let the number of records in an equivalence class be denoted by f_j. In the example in Figure 3, $J = 5$ since there are 5 equivalence classes in the whole disclosed database. The overall probability of re-identification under the prosecutor or journalist scenario is therefore reported as the minimum value of $1/f_j$ across all equivalence classes[21]:

$$1/\min_j (f_j)$$

Determine Threshold for Re-identification

The re-identification probability must now be interpreted, by determining whether or not it is acceptable. This is done by comparing the actual probability of re-identification for the disclosed prescription database with a threshold value. If the actual probability is above the threshold, then the risk of re-identification is unacceptable.

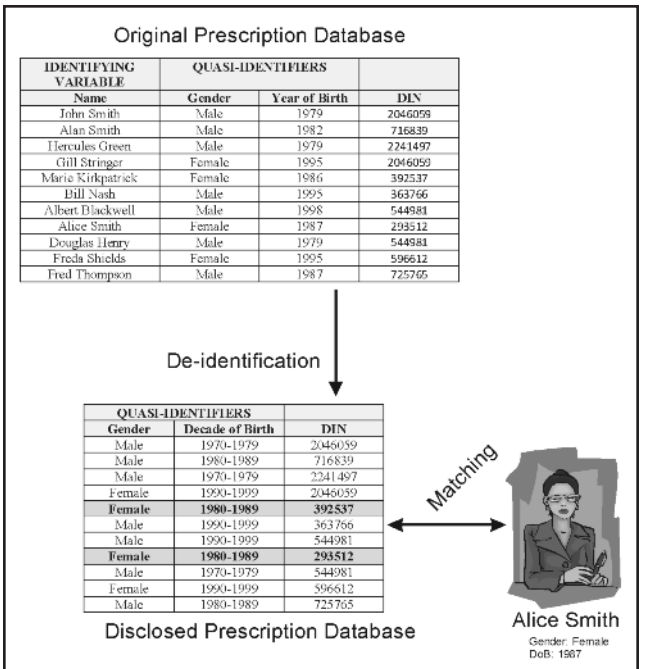

Figure 3. An example of how an intruder might attempt to re-identify Alice Smith in the disclosed prescription database. The original database had 11 records. After de-identification, the full name has been removed from the disclosed records, and the date of birth has been generalized to decade of birth. The intruder already has some background information about Alice, namely, her sex and her year of birth. Comparison of these details against the disclosed database reveals 2 possible records. Therefore, the probability of correct re-identification of Alice's record is 0.5.

In practice, quantitative thresholds are defined in terms of the smallest equivalence class sizes. When sensitive health data are disclosed, it is sometimes recommended that the smallest equivalence class have a size of at least 3,[26,27] which implies a threshold probability of 0.33. More often, a minimal equivalence class size of 5 is used,[28-34] which implies a threshold probability of 0.2.

Apply Generalization and Suppression

De-identification of hierarchical variables is performed by means of optimization algorithms.[35] Such algorithms determine the optimal amount of generalization to be performed on the data set. As part of that process, certain records are flagged for suppression. The process of suppression involves replacing some of the quasi-identifier values in these flagged records with a null value.

For the current study, a software program was developed to iteratively perform the generalization, suppression, and evaluation processes until an optimal solution was found.[36] The software implemented an algorithm based on the one described by Samarati,[37] but improved upon it by guaranteeing that a globally optimal solution would be found.

Evaluate De-identification Solutions

Once generalization and suppression have been performed, it is necessary to evaluate the quality of the de-identified data. An optimal de-identification solution is one that maintains the re-identification probability below the threshold while minimizing information loss.[38]

The Samarati algorithm[37] that was applied as the basis for this analysis used a precision information-loss metric[39] to measure how far the generalization had proceeded up the generalization hierarchy (an example of which is shown in Figure 2). Another popular information-loss metric used in the computational disclosure control literature is the discernability metric[40-46]:

$$\sum_{j=1}^{J} (f_j)^2$$

However, most end-users will find both of these metrics difficult to interpret. For example, a data analyst cannot intuitively say whether the value of the precision or discernability metric is too high. Therefore, a more intuitively meaningful information-loss metric was used in the current analysis, namely, the amount of data suppression in the disclosed prescription database.

Suppression can be done at the level of a variable, a record, or a cell. Suppression of a variable removes all data for that variable from the disclosed database. This is arguably an extreme form of generalization in which the variable is generalized to a single value. Suppression of a record removes the whole record from the disclosed prescription database. Suppression of a cell takes 2 possible forms: replacing the actual values for all of the quasi-identifiers with null values or selecting the optimal cells for replacement with null values. For this study, the latter approach was used, according to an optimization algorithm.[47] This is the best approximation to a globally optimal cell suppression available in the literature. The study team developed software to implement this suppression algorithm.

With cell suppression, other variables in the record (i.e., those that are not quasi-identifiers) remain untouched. For example, if the quasi-identifiers consist of age and sex, then the values for these 2 quasi-identifiers may be suppressed but the remaining variables, such as drug and diagnosis, would remain in the record. Cell suppression masks the information that the intruder needs to re-identify a record, but retains other clinical information that is useful for subsequent analysis.

In the current analysis, a maximum of 15% suppression on a single variable was deemed acceptable, as model-based imputation techniques could be used to estimate the missing values in subsequent data analysis.[48] Therefore, an optimal solution had to limit the percentage of records with suppression to no more than 15%. If that condition cannot be met, then it would be necessary to increase the value of the risk threshold. Increasing the risk threshold must be balanced by imposing additional security and privacy controls on the data recipient.

Impose Controls on the Data Recipient

Because of its common use in practice, the initial probability threshold was set at 0.2. If it is not possible to obtain a good de-identification solution with that threshold, then the threshold is increased to 0.33. However, the higher probability threshold must be balanced with greater security and privacy practices by the data recipient. Appendix 1 lists the practices that need to be in place at the higher threshold.

Perform Diagnostics

Because the measure of re-identification risk considers the worst-case scenario in the data set, it is often the case that a large percentage of the records in the disclosed database have a re-identification risk well below the threshold. A useful diagnostic is the cumulative distribution of risk values. For example, if the final probability threshold were increased from a baseline of 0.2 to 0.33, examination of the distribution might reveal that 95% of the records in the disclosed database have a re-identification risk below 0.2 and only 5% have a re-identification risk between 0.2 and 0.33. Therefore, even though the threshold seems quite high, the vast majority of the records still have a probability of re-identification below the baseline value of 0.2. In practice, this pattern is common. This diagnostic is most useful when the threshold is set above the baseline value of 0.2.

Data Set Analyzed

This study used data from CHEO. Records for all prescriptions dispensed from the CHEO pharmacy from the beginning of January 2007 to the end of June 2008 (18 months) were obtained following receipt of institutional ethics approval. In total, there were 94 100 records representing 10 364 patient visits and 6970 unique patients. The unit of analysis was the patient visit.

RESULTS

A variety of de-identification solutions were considered (Table 3). The table shows the level of detail for each variable, as well as the percentage of records with cell suppression needed to ensure that the re-identification risk would be below the threshold.

The first row in Table 3 (solution 1) shows the results for the fields that the private research firm initially requested. With the baseline risk threshold of 0.2, all of the records in the

Table 3. Results for Suppression of Records for Different Levels of Aggregation, Assuming an External Intruder*

Solution No.	Granularity of Variable to be Included in Disclosed Database						Risk Level for Scenario; % of Records with Cell Supression	
	Admission Date	Discharge Date	Length of Stay	Postal Code†	Age	Sex	Baseline-Risk Scenario‡	Higher-Risk Scenario§
Original request								
1	Day/month/year	Day/month/year	NA	FSA	Days	M or F	100	100
More generalized variables								
2	Day/month/year	Day/month/year	NA	Region	Days	M or F	100	100
3	Month/year	Month/year	NA	FSA	Days	M or F	98.8	90.4
4	Month/year	Month/year	NA	Region	Days	M or F	40.5	29.2
5	Quarter/year	Quarter/year	NA	FSA	Days	M or F	81.4	64.7
Optimal algorithms								
6	Quarter /year	Quarter/year	NA	Region	Days	M or F	NA	13.8
7	Quarter /year	NA	Days	Region	Weeks	M or F	NA	14.9

NA = not applicable.
*Suppression of a cell occurs only if the re-identification risk is higher than the threshold. Rate of cell suppression was calculated over the full 18-month period of the data set.
†Region = first character of postal code; FSA = forward sortation area (first 3 characters of the postal code).
‡Threshold 0.2. For solutions 6 and 7, no acceptable de-identification was possible at a risk threshold of 0.2, so no data are available for the baseline-risk scenario for these solutions.
§Threshold 0.33.

disclosed database would have some cell suppression. Therefore, the original data requested was clearly not de-identified, since all of the records were risky.

Solutions 2 through 5 illustrate how the extent of suppression decreases as more generalization is applied to the variables. For these solutions, the dates and postal code were generalized. The most generalized solution in this set (solution 4) would have the smallest amount of suppression. However, none of these solutions would have been acceptable because the level of suppression was still deemed too high.

Application of the optimal de-identification algorithms produced solutions 6 and 7. It was not possible to find a solution that was acceptable at a risk threshold of 0.2, and the risk threshold had to be raised. The last column in Table 3 shows the percentage of records suppressed with a 0.33 threshold.

At the higher risk threshold, solutions 6 and 7 maintained generalization below the maximum acceptable and ensured that cell suppression would be below the 15% threshold. These 2 solutions are differentiated by computation of the length-of-stay variable (solution 7) or retention of admission date (at a granularity of quarters) and discharge date (at a granularity of years) as separate variables (solution 6). Solution 7 was more useful, as it allowed for a more precise calculation of length of stay.

To manage overall risk at the higher threshold, the agreement with the private research firm stipulated implementation of the security and privacy practices described in Appendix 1, if they were not already in place.

Therefore, the following specific quasi-identifiers could be included in the disclosed prescription database:

- sex
- length of stay (in days)
- quarter and year of admission
- patient's region of residence (indicated by the first letter of the postal code)
- patient's age in weeks

In addition to these 5 quasi-identifiers, drug and diagnosis information was included.

According to the optimal suppression algorithm, 678 records (6.5% of all 10 364 records) had only a single quasi-identifier suppressed, 601 records (5.8%) had 2 quasi-identifiers suppressed, and 228 (2.2%) had 3 quasi-identifiers suppressed. Only a few records had 4 quasi-identifiers suppressed, and no records had all 5 quasi-identifiers suppressed. The optimal suppression approach was a big improvement over the more simplistic approach of removing all quasi-identifiers. In the optimal solution, in which 14.9% of the records had cell suppression (solution 7 in Table 3), only 35% of the quasi-identifier cells in the flagged records were actually suppressed.

The extent of suppression varied by quasi-identifier (Table 4). The age variable had the most suppression. Therefore, any analysis using age would have 11.3% of the total data set suppressed. The least affected variable was sex, with only 1.1% of all records having that variable suppressed.

The 0.33 value for risk of re-identification seems high. However, this represents the worst-case scenario. In practice, many records in the disclosed database would have a smaller risk. In the current case, just under 95% of the records that would be disclosed had a re-identification risk at or below 0.2,

Table 4. Extent of Optimal Suppression for Each Quasi-identifier

Quasi-identifier	No. (%) Records with Quasi-identifier Suppressed (n = 10 364)	
Sex	117	(1.1)
Age	1177	(11.4)
Region	475	(4.6)
Admission date	548	(5.3)
Length of stay	398	(3.8)

which is the commonly used baseline risk, and 80% of the records had a re-identification risk below 0.1. The remaining records had a risk higher than the 0.2 threshold. This can be illustrated by an empirical cumulative distribution plot (Figure 4), which shows that most of the records had a low probability of re-identification.

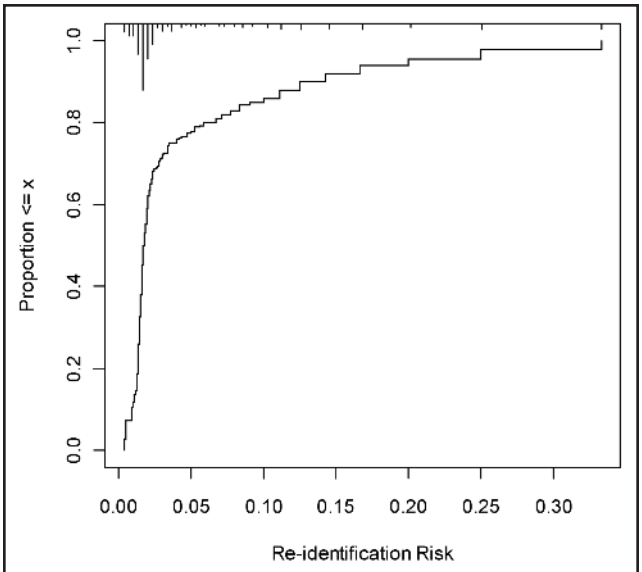

Figure 4. Ascending cumulative distribution plot showing the proportion of records with an actual re-identification risk below the *x* axis values. The proportion of records with a risk of 0.33 or less is 100%. However, as many as 95% of the records have a re-identification risk at or below 0.2. The histogram along the top axis shows the density of records with specific levels of re-identification risk. In this case, the majority of records are concentrated at the low end of re-identification risk.

DISCUSSION

The assessment reported here gave the hospital's director of pharmacy and chief privacy officer an objective basis for assessing the risk of re-identification when disclosing pharmacy data to the private research firm. It was clear from this analysis that the data fields originally requested by the company would not have resulted in a de-identified data set, as was initially assumed. The conclusions from the assessment emphasized the importance of this independent analysis.

This analysis of re-identification risk using 18 months of prescription records from CHEO allowed concrete recommendations about managing this risk. Methods of risk management included generalizing the data, suppressing some cell values within the data, and stipulating additional security and privacy practices in the data-sharing agreement with the private research firm. This kind of analysis makes the tradeoffs between privacy and utility explicit and ensures that underlying assumptions are brought to the surface.

Implementation Considerations

The analysis described here would typically be performed in conjunction with the privacy office of a hospital or other health care institution. The level of analytical sophistication that privacy officers can offer is increasing, and there is growing emphasis on the privacy issues associated with secondary use of hospital data by external organizations and researchers.

If such expertise does not exist within the hospital, we recommend that it be developed. One reason for doing so is the dependence of the results of a re-identification risk assessment on the data distribution. If the distribution of patient demographics, their admission rates, and/or their length of stay changes significantly over time, the risk assessment may no longer be accurate and a new risk assessment would be required. If the expertise to perform such assessments has been developed, the hospital can undertake re-analysis whenever it is needed.

De-identifying data in the manner described here and finding an optimal solution that balances generalization with information loss necessitates a variety of software tools, specifically tools to implement de-identification algorithms. For the study reported here, the software tools were developed within the hospital.

Limitations

The results presented here were obtained using a data set from a medium-sized pediatric hospital. Larger pediatric hospitals would have a larger number of admissions during a period of the same duration. In such cases, the risk assessment may allow additional precision for the same variables. Separate analyses would be needed for adult hospitals.

Rare and visible diseases present an additional risk of re-identification. Some work has already been done on identifying the rare and visible diseases that carry such extra risk.[49] For example, it is possible to eliminate records with disease codes (e.g., from the International Classification of Diseases, 9th revision) that match those on the list of rare and visible diseases. However, this was not done in the analysis described here, as the list has not yet been formally published. As such, some residual re-identification risks may remain, even after addressing the risks identified here.

In the assessment of the risk of breaching privacy reported here, rarely prescribed drugs were not considered. For an intruder to use information about rare drugs for re-identification, the intruder would need to have background information on the drugs that a specific patient was taking; under the scenarios considered for this analysis, this did not seem likely. Alternatively, if a drug is known as a treatment for a rare and/or visible disease, then the fact that a patient has taken the drug would indicate that he or she has that disease. To the authors' knowledge, no work has been done to identify the drugs commonly prescribed for the rare and visible diseases mentioned above, taking into account off-label use.

References

1. Kallukaran P, Kagan J. Data mining at IMS Health: how we turned a mountain of data into a few information-rich molehills [presentation]. In: *Proceedings of 24th Annual SAS Users Group International Conference*; 1999 Apr 11–14; Miami Beach (FL). Paper 127.
2. Foley J. An overview of the research value of administrative data from private sector drug plans [presentation]. 9th World Conference on Clinical Pharmacology and Therapeutics; 2008 Jul 27–Aug 1; Québec (QC).
3. Zoutman D, Ford B, Bassili A. The confidentiality of patient and physician information on pharmacy prescription records. *CMAJ* 2004;170(5): 815-816.
4. Zoutman D, Ford B, Bassili A. Privacy of pharmacy prescription records [letter: author response]. *CMAJ* 2004;171(7):712.
5. Porter C. De-identified data and third party data mining: the risk of re-identification of personal information. *Shidler J Law Comm Technol* 2008;5(1):Article 3.
6. Perun H, Orr M, Dimitriadis F. *Guide to the Ontario Personal Health Information Protection Act.* Toronto (ON): Irwin Law; 2005.
7. *MedMap drug utilization program: program overview 2008*. Ottawa (ON): Brogan Inc; 2008.
8. Hansell S. AOL removes search data on group of web users. *New York Times* 2006 Aug 8:C4.
9. Barbaro M, Zeller T Jr. A face is exposed for AOL searcher no. 4417749. *New York Times* 2006 Aug 9:A1.
10. Zeller T Jr. AOL moves to increase privacy on search queries. *New York Times* 2006 Aug 22.
11. Ochoa S, Rasmussen J, Robson C, Salib M. *Reidentification of individuals in Chicago's homicide database: a technical and legal study*. Cambridge (MA): Massachusetts Institute of Technology; 2008.
12. Narayanan A, Shmatikov V. *Robust de-anonymization of large datasets (how to break anonymity of the Netflix prize dataset)*. Austin (TX): University of Texas at Austin; 2008.
13. Sweeney L. *Computational disclosure control: a primer on data privacy protection*. Cambridge (MA): Massachusetts Institute of Technology; 2001.
14. *Southern Illinoisan* v. *Department of Public Health* (Ill. App. Dist. 5, 2004).
15. *Southern Illinoisan* v. *Illinois Department of Public Health* (Ill. 2006).
16. *Mike Gordon* v. *The Minister of Health* (2006), 2008 FC 258. Affidavit of Bill Wilson.
17. *Data loss database—2008 yearly report.* Glen Allen (VA): Open Security Foundation; 2008.
18. Verizon Business Risk Team. *2008 data breach investigations report*. Verizon; 2008.
19. Yolles B, Connors J, Grufferman S. Obtaining access to data from government-sponsored medical research. *N Engl J Med* 1986; 315(26):1669-1672.
20. Cecil J, Boruch R. Compelled disclosure of research data: an early warning and suggestions for psychologists. *Law Hum Behav* 1988; 12(2):181-189.
21. El Emam K, Dankar F. Protecting privacy using k-anonymity. *J Am Med Inform Assoc* 2008;15(5):627-637.
22. El Emam K, Jabbouri S, Sams S, Drouet Y, Power M. Evaluating common de-identification heuristics for personal health information. *J Med Internet Res* 2006;8(4):e28.
23. Gross R, Acquisti A. Information revelation and privacy in online social networks (the Facebook case) [presentation]. In: De Capitani di Vimercati S, Dingledine R, editors. P*roceedings of the Workshop on Privacy in the Electronic Society*; 2005 Nov 7; Alexandria (VA). New York (NY): The Association for Computing Machinery; 2005. p 71-80.
24. Dwyer C, Hiltz S. Trust and privacy concern within social networking sites: a comparison of Facebook and MySpace [presentation]. In: *Proceedings of the Americas Conference on Information Systems*; 2007 Aug 9–12; Keystone (CO). Paper 339.
25. Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. *Int J Uncertain Fuzziness Knowl Based Syst* 2002;10(5):571-588.
26. Duncan G, Jabine T, de Wolf S. *Private lives and public policies: confidentiality and accessibility of government statistics*. Washington (DC): National Academies Press; 1993.
27. de Waal A, Willenborg L. A view on statistical disclosure control for microdata. *Surv Methodol* 1996;22(1):95-103.
28. *Cancer Care Ontario data use and disclosure policy*. Toronto (ON): Cancer Care Ontario; 2005.
29. *Security and confidentiality policies and procedures*. Saskatoon (SK): Health Quality Council; 2004.
30. *Privacy code*. Saskatoon (SK): Health Quality Council; 2004.
31. *Privacy code*. Winnipeg (MB): Manitoba Centre for Health Policy; 2002.
32. Federal Committee on Statistical Methodology, Subcommittee on Disclosure Limitation Methodology. *Report on statistical disclosure control*. Working Paper 22. Washington (DC): Office of Management and Budget; 1994.
33. *Therapeutic abortion survey*. Ottawa (ON): Statistics Canada; 2007 [cited 2009 Jun 26]. Available from: http://www.statcan.gc.ca/cgibin/imdb/p2SV.pl?Function=getSurvey&SDDS=3209&lang=en&db=imdb&adm=8&dis=2
34. El Emam K. Heuristics for de-identifying health data. *IEEE Secur Priv* 2008 Jul-Aug;:72-75.
35. Ciriani V, De Capitani di Vimercati S, Samarati P. k-Anonymity. In: Yu T, Jajodia S, editors. *Secure data management in decentralized systems*. New York (NY): Springer; 2007.
36. El Emam K, Dankar F, Issa RJ, Jonker E, Amyot D, Cogo E, et al. A globally optimal k-anonymity method for the de-identification of health data. *J Am Med Inform Assoc*. Forthcoming in 2009.
37. Samarati P. Protecting respondents' identities in microdata release. *IEEE Trans Knowl Data Eng* 2001;13(6):1010-1027.
38. Trottini M. Assessing disclosure risk and data utility: a multiple objectives decision problem. UNECE/Eurostat Work Session on Statistical Data Confidentiality; 2003. Working Paper 19.
39. Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. *Int J Uncertain Fuzziness Knowl Based Syst* 2002; 10(5):18.
40. Bayardo R, Agrawal R. Data privacy through optimal k-anonymization. In: *Proceedings of 21st International Conference on Data Engineering*; 2005 Apr 5-8; Tokyo (Japan). Los Alamitos (CA): IEEE Computer Society; 2005. p 217-228.
41. LeFevre K, DeWitt D, Ramakrishnan R. Mondrian multidimensional k-anonymity. In: Barga RS, Zhou X, editors. *Proceedings of 22nd International Conference on Data Engineering*; 2006 Apr 3–8; Atlanta (GA). Los Alamitos (CA): IEEE Computer Society; 2006. p 25.
42. Hore B, Jammalamadaka R, Mehrotra S. Flexible anonymization for privacy preserving data publishing: a systematic search based approach. In: *Proceedings of SIAM International Conference on Data Mining*; 2007 Apr 26-28; Minneapolis (MN). Philadelphia (PA): Society for Industrial and Applied Mathematics: 2007. p 497-502.
43. Xu J, Wang W, Pei J, Wang X, Shi B, Fu A. Utility-based anonymization for privacy preservation with less information loss. *ACM Spec Int Group Knowl Discov Data Explor Newsl* 2006;8(2):21-30.
44. Nergiz M, Clifton C. Thoughts on k-anonymization. *Data Knowl Eng* 2007;63(3):622-643.
45. Polettini S. *A note on the individual risk of disclosure*. Rome (Italy): Istituto nazionale di statistica; 2003.

46. Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. L-diversity: privacy beyond k-anonymity. *ACM Trans Knowl Disc Data* 2007;1(1):Article 3.
47. Aggarwal G, Feder T, Kenthapadi K, Motwani R, Panigrahy R, Thomas D, et al. Anonymizing tables. Lecture Notes in Computer Science 3363. In: *Proceedings of the 10th International Conference on Database Theory*; 2005 Jan 5–7; Edinburgh (UK). New York (NY): Springer; 2005. p 246-258.
48. Little R, Rubin D. *Statistical analysis with missing data*. New York (NY): John Wiley & Sons; 1987.
49. Eguale T, Bartlett G, Tamblyn R. Rare visible disorders/diseases as individually identifiable health information. In: *Proceedings of the American Medical Informatics Association Symposium*; 2005 Oct 22–26: Washington (DC). p 947.
50. *Guidance document: taking privacy into account before making contracting decisions*. Ottawa (ON): Treasury Board of Canada Secretariat; 2006.
51. *Privacy impact assessment guidelines: a framework to manage privacy risks*. Ottawa (ON): Treasury Board of Canada Secretariat; 2002.
52. *Canadian Institute for Health Information privacy tool kit*. Ottawa (ON): Canadian Institute for Health Information; 2003.
53. *Privacy and confidentiality of health information at CIHI: principles and policies for the protection of personal health information 2007*. Ottawa (ON): Canadian Institute for Health Information; 2007.
54. *CIHR best practices for protecting privacy in health research*. Ottawa (ON): Canadian Institutes of Health Research; 2005.
55. *COACH guidelines for the protection of health information*. Toronto (ON): Canadian Organization for Advancement of Computers in Health; 2006.
56. Collins P, Slaughter P, Roos N, Weisbaum KM, Hirtle M, Williams JI, et al. *Privacy best practices for secondary data use: harmonizing research and privacy*. Toronto (ON): Institute for Clinical Evaluative Sciences; 2006.

Khaled El Emam, BEng, PhD, is with the CHEO Research Institute, Ottawa, Ontario.

Fida K Dankar, BSc, MSc, PhD, is with the CHEO Research Institute, Ottawa, Ontario.

Régis Vaillancourt, BPharm, PharmD, is with the Children's Hospital of Eastern Ontario, Ottawa, Ontario. He is also an Associate Editor with the *CJHP*.

Tyson Roffey, BA, is with the Children's Hospital of Eastern Ontario, Ottawa, Ontario.

Mary Lysyk, BA, MSc, is with the University of Ottawa, Ottawa, Ontario.

Address correspondence to:
Dr Khaled El Emam
Children's Hospital of Eastern Ontario Research Institute
401 Smyth Road
Ottawa ON K1H 8L1

e-mail: kelemam@uottawa.ca

Acknowledgements
This work was partially funded by the Ontario Centres of Excellence. We thank Bradley Malin of Vanderbilt University for his comments on an earlier version of this paper.

Appendix 1. Checklist for Security and Privacy Practices

Generalization and suppression are applied to a data set to ensure that the risk of re-identification is below the threshold. If the extent of generalization is deemed clinically unacceptable or the extent of suppression is deemed too high, then the threshold can be raised. However, if the threshold is raised, security and privacy practices must be put in place to make the higher threshold acceptable. The checklist in this appendix specifies these practices.

For example, for a particular database, assume that the probability threshold was set to the common value of 0.2. When the data were de-identified using the automated algorithm, the resultant data set had too much suppression. Therefore, the threshold had to be raised. However, doing that requires more security and privacy practices to compensate for the higher probability of re-identification with the higher threshold. For example, the recipient of the database may be required to accept surprise audits from the hospital, to implement extensive standard operating procedures related to information security, and to ensure that practices are compliant with these procedures. The risk threshold is then raised to 0.33. At that risk level, the de-identified data are clinically useful.

The checklist on page 318 (Table A1) is a summary of items directly specified in the following policies, guidelines, or application forms: Government of Canada guidelines,[50,51] Canadian Institute for Health Information guidelines,[52,53] Canadian Institutes of Health Research guidelines,[54] Canadian Organization for Advancement of Computers in Health guidelines,[55] secondary-use guidelines published by the Institute for Clinical Evaluative Science,[56] and guidelines from the research ethics boards of 13 large academic health centres (Toronto Academic Health Sciences Network Human Subjects Research Application [2006], Council of Research Ethics Boards Common REB Application Form—Ottawa [2007], the University of British Columbia Office of Research Services—Investigator and Study Team Human Ethics Application [2008], Queen's University Health Sciences Research Ethics Board [2008], University of Manitoba—Bannatyne Campus Research Ethics Boards [Biomedical and Health] [2007], Health Research Ethics Board of the University of Alberta [2008], Quebec Multicentre Project Review Application [2008], Memorial University of Newfoundland Faculty of Medicine Human Investigation Committee [2005], Dalhousie University Health Sciences Research Ethics Board [2008], McGill University Faculty of Medicine Institutional Review Board [2007], University of Western Ontario Health Sciences Research Ethics Board [2007], University of Saskatchewan Biomedical Research Ethics Board [2008], and St Joseph's Healthcare Hamilton/Hamilton Health Sciences/McMaster University Faculty of Health Sciences Research [2006]).

Table A1. Checklist of Practices That Must Be in Place at a Higher Threshold for Re-identification Risk, as Detailed in Policies, Guidelines, and Application Forms of Various Bodies

Practice	CIHI*	CIHR	COACH†	ICES/ SDU	% of REBs
Controlling access, disclosure, retention, and disposition of personal data	√	√	√	√	100
Requestor allows only "authorized" staff to access and use data on a "need-to-know" basis (i.e., when required to perform their duties)	√	√	√	√	85
Data-sharing agreement between collaborators and subcontractors has been or will be implemented	√	√	√	√	23
Nondisclosure or confidentiality agreement (pledge of confidentiality) is in place for all staff, including external collaborators and contractors	√	√	√	√	39
Requestor will only publish or disclose aggregated data that do not allow identification of individuals	√	√	√	√	100
Long-term retention of personal data will be subject to periodic audits and oversight by independent bodies	√	√	√		23
Data will be disposed of after a specified retention period	√	√	√	√	62
Information will not be processed, stored, or maintained outside of Canada, and parties outside of Canada will not have access to the data					8
Data will not be disclosed or shared with third parties	√	√	√	√	46
Safeguarding personal data	√	√	√	√	**100**
Assessment of threat and risk vulnerability has been conducted for information systems, and assessment has been conveyed to data custodian	√	√	√	√	NS
Organizational governance framework for privacy, confidentiality, and security is in place at requestor site	√		√	√	NS
Organizational policies for data storage, management, and access are in place at requestor site	√	√	√	√	15
Privacy and security policies and procedures are monitored and enforced	√	√	√	√	15
Mandatory and ongoing privacy, confidentiality, and security training is conducted for all individuals and/or team members, including those at external collaborating or subcontracting sites	√	√	√	√	31
Appropriate sanctions are in place for breach of privacy, confidentiality, or security, including dismissal and/or loss of institutional privileges, and these have been clearly stipulated in signed pledge of confidentiality	√	√	√	√	15
Privacy officers and/or data stewardship committees have been appointed at requestor site	√	√	√	√	15
Breach-of-privacy protocol is in place, including immediate written notification to data custodian	√	√	√	√	8
Internal and external privacy reviews and audits have been implemented	√	√	√	√	8
Authentication measures (such as computer password protection and unique log-on identification) have been implemented to ensure that only authorized personnel can access computer systems	√	√	√	√	NS
Authentication measures (such as computer password protection and unique log-on identification) have been implemented to ensure that only authorized personnel can access data	√	√	√	√	69
Special protection has been installed for remote electronic access to data	NS	√	√	√	23
Virus-checking programs have been implemented	√	√	√	√	23
Detailed monitoring system for audit trail has been instituted to document the person, time, and nature of data access, with flags for aberrant use and "abort" algorithms to end questionable or inappropriate access	√	√	√	√	15
If electronic transmission of data is required, an encrypted protocol will be used	√		√	√	54
Computers and files that hold the disclosed information are housed in secure settings in rooms protected by such methods as combination lock doors or smart card door entry, with paper files stored in locked storage cabinets	√	√	√	√	85
Staff have been provided with photo identification or coded card swipe	√	NS	√	√	NS
Visitors are screened and supervised	√	NS	√	√	NS
Alarm systems are in place	NS	NS	√	NS	NS
Number of locations in which personal information is stored has been minimized and specified in advance	NS	√	NS	NS	NS
Architectural space precludes public access to areas where sensitive data are held	NS	√	NS	NS	NS
Routine surveillance of premises is conducted	NS	√	√	√	NS
Physical security measures are in place to protect data from hazards such as floods or fire	√	√	√	√	8

Table A1. Checklist of Practices That Must Be in Place at a Higher Threshold for Re-identification Risk, as Detailed in Policies, Guidelines, and Application Forms of Various Bodies (continued)

Practice	CIHI*	CIHR	COACH†	ICES/ SDU	% of REBs
Ensuring accountability and transparency in the management of personal data	√	√	√	√	100
Contact information and title of senior individuals who will be accountable for privacy, confidentiality, and security of data have been provided to data custodian, and requestor notifies custodian of any changes to this information	√	√	√	√	‡
Contact information and title of senior individual who will be accountable for employees and contractors has been provided to data custodian, and requestor will notify custodian of any changes to this information	√	√	√	√	‡
Organizational transparency and public notification plan is in place at the requestor site and is open about collection or disclosure of information, research objectives, and privacy policy and practices	√	√	√	√	23
Requestor site has procedures in place to receive and respond to public complaints or inquiries about its privacy policies and practices related to the handling of personal data, and complaint procedures are easily accessible and simple to use	√	√	√	√	8
Independent authority (e.g., research ethics board) has approved the proposal for secondary use of data	√	√	√	√	100
Internal and/or external audit and monitoring mechanisms are in place as appropriate	√	√	√	√	39
Independent advisory or data stewardship committee serves in data oversight capacity (e.g., advisory committee for defining scope and strategic priorities of research studies)		√	√	√	31

CIHI = Canadian Institute for Health Information, CIHR = Canadian Institutes of Health Research, COACH = Canadian Organization for Advancement of Computers in Health, ICES/SDU = Institute for Clinical Evaluative Sciences/secondary data use, REB = research ethics board, NS = not specified (used when the source material had only an open-ended question and was not specific about what the data recipient should do; for example, the instruction "Please describe the administrative, technical, and physical safeguards that will be used to protect the confidentiality and security of data" would be designated as "NS" in this table).
*For CIHI, items reflect those pertaining to the Institute's privacy program, not data requestors.
†The COACH guidelines are for health information custodians.
‡All REBs require signing authority of the principal investigator, who is accountable for use of the data; however, accountability for privacy and confidentiality of data are rarely specified.