Genome Medicine

REVIEW

Methods for the de-identification of electronic health records for genomic research

Khaled El Emam[1,2]*

Abstract

Electronic health records are increasingly being linked to DNA repositories and used as a source of clinical information for genomic research. Privacy legislation in many jurisdictions, and most research ethics boards, require that either personal health information is de-identified or that patient consent or authorization is sought before the data are disclosed for secondary purposes. Here, I discuss how de-identification has been applied in current genomic research projects. Recent metrics and methods that can be used to ensure that the risk of re-identification is low and that disclosures are compliant with privacy legislation and regulations (such as the Health Insurance Portability and Accountability Act Privacy Rule) are reviewed. Although these methods can protect against the known approaches for re-identification, residual risks and specific challenges for genomic research are also discussed.

Electronic health records and the need for de-identification

Electronic health records (EHRs) are increasingly being used as a source of clinically relevant patient data for research [1,2], including genome-wide association studies [3]. Often, research ethics boards will not allow data custodians to disclose identifiable health information without patient consent. However, obtaining consent can be challenging and there have been major concerns about the negative impact of obtaining patient consent on the ability to conduct research [4]. Such concerns are reinforced by the compelling evidence that requiring explicit consent for participation in different forms of health research can have a negative impact on the process and outcomes of the research itself [5-7]. For example, recruitment rates decline significantly when individuals are asked to consent; those who consent tend to be different from those who decline consent on a number of important demographic and socio-economic variables, hence potentially introducing bias in the results [8]; and consent requirements increase the cost of, and time for, conducting the research. Furthermore, often it is not practical to obtain individual patient consent because of the very large populations involved, the lack of a relationship between the researchers and the patients, and the time elapsed between data collection and the research study.

One approach to facilitate the disclosure of information for the purposes of genomic research, and to alleviate some of the problems documented above, is to de-identify data before disclosure to researchers or at the earliest opportunity afterwards [9,10]. Many research ethics boards will waive the consent requirement if the first 'use' of the data is to de-identify it [11,12].

The i2b2 project (informatics for integration of biology and the bedside) has developed tools for clinical investigators to integrate medical records and clinical research. A query tool in i2b2 allows the computation of cohort sizes in a privacy protective way, and a data export tool allows the extraction of de-identified individual-level data [13,14]. Also, the eMerge network, which consists of five sites in the United States, is an example of integrated EHR and genetic databases [3]. The BioVU system at Vanderbilt University, a member of the eMerge network, links a biobank of discarded blood samples with EHR data, and information is disclosed for research purposes after de-identification [3,15].

Here, I provide a description and critical analysis of de-identification methods that have been used in genomic research projects, such as i2b2 and eMerge. This is augmented with an overview of contemporary standards, best practices and recent de-identification methodologies.

De-identification: definitions and concepts

A database integrating clinical information from an EHR with a DNA repository is referred to here as a translational research information system (TRIS) for brevity [16]. It is assumed that the data custodian is extracting a particular set of variables on patients from a TRIS and

*Correspondence: kelemam@uottawa.ca
[2]Pediatrics, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, K1H 8L1, Canada
Full list of author information is available at the end of the article

disclosing that to a data recipient for research purposes, and that the data custodian will be performing the de-identification before the disclosure or at the earliest opportunity after disclosure. The concern for the data custodian is the risk that an adversary will try to re-identify the disclosed data.

Identity versus attribute disclosure

There are two kinds of re-identification that are of concern. The first is when an adversary can assign an identity to a record in the data disclosed from the TRIS. For example, the adversary would be able to determine that record number 7 belongs to a patient named 'Alice Smith'. This is called identity disclosure. The second type of disclosure is when an adversary learns something new about a patient in the disclosed data without knowing which specific record belongs to that patient. For example, if all 20-year-old female patients in the disclosed data who live in Ontario had a specific diagnosis, then an adversary does not need to know which record belongs to Alice Smith; if she is 20 years old and lives in Ontario then the adversary will discover something new about her: the diagnosis. This is called attribute disclosure.

All the publicly known examples of re-identification of personal information have involved identity disclosure [17-26]. Therefore, the focus is on identity disclosure because it is the type that is known to have occurred in practice.

Types of variable

The data in an EHR will include clinical information, and possibly socio-economic status information that may be collected from patients or linked in from external sources (such as the census). EHR information can be divided into four categories. The distinctions among these categories are important because they have an impact on the probability of re-identification and on suitable de-identification methods.

Directly identifying information

One or more direct identifiers can be used to uniquely identify an individual, either by themselves or in combination with other readily available information. For example, there are more than 200 people named 'John Smith' in Ontario, and therefore the name by itself would not be directly identifying, but in combination with the address it would be directly identifying information. Examples of directly identifying information include email address, health insurance card number, credit card number, and social insurance number.

Indirectly identifying relational information

Relational information can be used to probabilistically identify an individual. General examples include sex, geographic indicators (such as postal codes, census geography, or information about proximity to known or unique landmarks), and event dates (such as birth, admission, discharge, procedure, death, specimen collection, or visit/encounter).

Indirectly identifying transactional information

This is similar to relational information in that it can be used to probabilistically identify an individual. However, transactional information may have many instances per individual and per visit. For example, diagnosis codes and drugs dispensed would be considered transactional information.

Sensitive information

This is information that is rarely useful for re-identification purposes - for example, laboratory results.

For any piece of information, its classification into one of the above categories will be context dependant.

Relational and transactional information are referred to as quasi-identifiers. The quasi-identifiers represent the background knowledge about individuals in the TRIS that can be used by an adversary for re-identification. Without this background knowledge identity disclosure cannot occur. For example, if an adversary knows an individual's date of birth and postal code, then s/he can re-identify matching records in the disclosed data. If the adversary does not have such background knowledge about a person, then a date of birth and postal code in a database would not reveal the person's identity. Furthermore, because physical attributes and certain diagnoses can be inferred from DNA analysis (for example, gender, blood type, approximate skin pigmentation, a diagnosis of cystic fibrosis or Huntington's chorea), the DNA sequence data of patients known to an adversary can be used for phenotype prediction and subsequent re-identification of clinical records [27-29]. If an adversary has an identified DNA sequence of a target individual, this can be used to match and re-identify a sequence in the repository. Without an identified DNA sequence or reference sample as background knowledge, such an approach for re-identification would not work [16]. The manner and ease with which an adversary can obtain such background knowledge will determine the plausible methods of re-identification for a particular dataset.

Text versus structured data

Another way to consider the data in a TRIS is in terms of representation: structured versus free-form text. Some data elements in EHRs are in a structured format, which means that they have a pre-defined data type and semantics (for example, a date of birth or a postal code). There will also be plenty of free-form text in the form of, for example, discharge summaries, pathology reports,

and consultation letters. Any realistic de-identification process has to deal with both types of data. The BioVU and i2b2 projects have developed and adapted tools for the de-identification of free-form text [15,30].

De-identification standards

In the US, the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule provides three standards for the disclosure of health information without seeking patient authorization: the Safe Harbor standard (henceforth Safe Harbor), the Limited Dataset, and the statistical standard. Safe Harbor is a precise standard for the de-identification of personal health information when disclosed for secondary purposes. It stipulates the removal of 18 variables from a dataset as summarized in Box 1. The Limited Dataset stipulates the removal of only 16 variables, but also requires that the data recipient sign a data sharing agreement with the data custodian. The statistical standard requires an expert to certify that 'the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information.' Out of these three standards, the certainty and simplicity of Safe Harbor has made it attractive for data custodians.

Safe Harbor is also relevant beyond the US. For example, health research organizations and commercial organizations in Canada choose to use the Safe Harbor criteria to de-identify datasets [31,32], Canadian sites conducting research funded by US agencies need to comply with HIPAA [33], and international guidelines for the public disclosure of clinical trials data have relied on Safe Harbor definitions [34].

However, Safe Harbor has a number of important disadvantages. There is evidence that it can result in the excessive removal of information useful for research [35]. At the same time it does not provide sufficient protection for many types of data, as illustrated below.

First, it does not explicitly consider genetic data as part of the 18 fields to remove or generalize. There is evidence that a sequence of 30 to 80 independent single nucleotide polymorphisms (SNPs) could uniquely identify a single person [36]. There is also a risk of re-identification from pooled data, where it is possible to determine whether an individual is in a pool of several thousand SNPs using summary statistics on the proportion of individuals in the case or control group and the corresponding SNP value [37,38].

Second, Safe Harbor does not consider longitudinal data. Longitudinal data contain information about multiple visits or episodes of care. For example, let us consider the state inpatient database for California for the year 2007, which contains information on 2,098,578 patients. A Safe Harbor compliant dataset consisting only of the quasi-identifiers gender, year of birth, and year of admission has less than 0.03% of the records with a high probability of re-identification. A high probability of re-identification is defined as over 0.2. However, with two more longitudinal variables added, length of stay and time since last visit for each visit, then 16.57% of the records have a high probability of re-identification (unpublished observations). Thus, the second dataset also meets the Safe Harbor definition but has a markedly

Box 1. The 18 elements in the HIPAA Privacy Rule Safe Harbor standard that must be excluded/removed from a dataset

The following identifiers of the individual or of relatives, employers, or household members of the individual, are removed:

1. Names;
2. All geographic subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code if, according to the current publicly available data from the Bureau of the Census:

a) The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and

b) The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000.

3. All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;
4. Telephone numbers;
5. Fax numbers;
6. Electronic mail addresses;
7. Social security numbers;
8. Medical record numbers;
9. Health plan beneficiary numbers;
10. Account numbers;
11. Certificate/license numbers;
12. Vehicle identifiers and serial numbers, including license plate numbers;
13. Device identifiers and serial numbers;
14. Web Universal Resource Locators (URLs);
15. Internet Protocol (IP) address numbers;
16. Biometric identifiers, including finger and voice prints;
17. Full face photographic images and any comparable images; and
18. Any other unique identifying number, characteristic, or code.

Adapted from [87]

higher percentage of the population at risk of re-identification. Therefore, Safe Harbor does not ensure that the data are adequately de-identified. Longitudinal information, such as length of stay and time since last visit, may be known by neighbors, co-workers, relatives, and ex-spouses, and even the public for famous people.

Third, Safe Harbor does not deal with transactional data. For example, it has been shown that a series of diagnosis codes (International Statistical Classification of Diseases and Related Health Problems) for patients makes a large percentage of individuals uniquely identifiable [39]. An adversary who is employed by the healthcare provider could have access to the diagnosis codes and patient identity, which can be used to re-identify records disclosed from the TRIS.

Fourth, Safe Harbor does not take into account the sampling fraction - it is well established that sub-sampling can reduce the probability of re-identification [40-46]. For example, consider a cohort of 63,796 births in Ontario over 2004 to 2009 and three quasi-identifiers: maternal postal code, date of birth of baby, and mother's age. Approximately 96% of the records were unique on these three quasi-identifiers, making them highly identifiable. For research purposes, this dataset was de-identified to ensure that 5% or less of the records could be correctly re-identified by reducing the precision of the postal code to the first three characters, and the date of birth to year of birth. However, a cohort of 127,592 births de-identified in exactly the same way could have 10% of its records correctly re-identified. In this case the variables were exactly the same in the two cohorts but, because the sampling fraction varies, the percentage of records that can be re-identified doubles (from 5% to 10%, respectively).

Finally, other pieces of information that can re-identify individuals in free-form text and notes are not accounted for in Safe Harbor. The following example illustrates how I used this information to re-identify a patient. In a series of medical records that have been de-identified using the Safe Harbor standard, there was a record about a patient with a specific injury. The notes mentioned the profession of the patient's father and hinted at the location of his work. This particular profession lists its members publicly. It was therefore possible to identify all individuals within that profession in that region. Searches through social networking sites allowed the identification of a matching patient (having the same surname) with details of the specific injury during that specific period. The key pieces of information that made re-identification possible were the father's profession and region of work, and these are not part of the Safe Harbor items.

Therefore, universal de-identification heuristics that proscribe certain fields or prescribe specific generalizations of fields will not provide adequate protection in all situations and must be used with caution. Both the BioVU [15] and the i2b2 project [13] de-identify individual-level data according to the Safe Harbor standard, but also require a data sharing agreement with the data recipients as required by the Limited Dataset provision, and some sites implementing the i2b2 software use the Limited Dataset provision for de-identification [14].

Although the Limited Dataset provision provides a mechanism to disclose information without consent, it does not produce data that are de-identified. The challenge for data custodians is that the notices to patients for some repositories state that the data will be de-identified, so there is an obligation to perform de-identification before disclosure [15,47]. Where patients are approached in advance for consent to include their data in the repository, this is predicated on the understanding that any disclosures will be of de-identified data [3]. Under these circumstances, a more stringent standard than the Limited Dataset is required. Within the framework of HIPAA, one can then use the statistical standard for de-identification. This is consistent with privacy legislation and regulations in other jurisdictions, which tend not to be prescriptive and allow a more context-dependant interpretation of identifiability [26].

Managing re-identification risk

The statistical standard in the HIPAA Privacy Rule provides a means to disclose more detailed information for research purposes and still manage overall re-identification risk. Statistical methods can provide quantitative guarantees to patients and research ethics boards that the probability of re-identification is low.

A risk-based approach has been in use for a few years for the disclosure of large clinical and administrative datasets [48], and can be similarly used for the disclosure of information from a TRIS. The basic principles of a risk-based approach for de-identification are that (a) a re-identification probability threshold should be set and (b) the data should be de-identified until the actual re-identification probability is below that threshold.

Because measurement is necessary for setting thresholds, the supplementary material (Additional file 1) consists of a detailed review of re-identification probability metrics for evaluating identity disclosure. Below is a description of how to set a threshold and an overview of de-identification methods that can be used.

Setting a threshold

There are two general approaches to setting a threshold: (a) based on precedent and (b) based on an assessment of the risks from the disclosure of data.

Precedents for thresholds

Historically, data custodians have used the 'cell size of five' rule to de-identify data [49-58]. In the context of a

probability of re-identifying an individual, this is equivalent to a probability of 0.2. Some custodians use a cell size of 3 [59-62], which is equivalent to a probability of 0.33 of re-identifying a single individual. Such thresholds are suitable when the data recipient is trusted.

It has been estimated that the Safe Harbor standard results in 0.04% of the population being at high risk for re-identification [63,64]. Another re-identification attack study evaluated the proportion of Safe Harbor compliant medical records that can be re-identified and found that only 0.01% can be correctly re-identified [65]. In practice, setting such low thresholds can also result in significant distortion to the data [35], and is arguably more suitable when data are being publicly disclosed.

Risk-based thresholds

With this approach, the re-identification probability threshold is determined based on factors characterizing the data recipient and the data [48]. These factors have been suggested and have been in use informally by data custodians to inform their disclosure decisions for at least the last decade and a half [46,66], and they cover three dimensions [67], as follows.

First, mitigating controls: this is the set of security and privacy practices that the data recipient has in place. The practices used by custodians of large datasets and recommended by funding agencies and research ethics boards for managing sensitive health information have been reviewed elsewhere [68].

Second, invasion of privacy: this evaluates the extent to which a particular disclosure would be an invasion of privacy to the patients (a checklist is available in [67]). There are three considerations: (i) the sensitivity of the data: the greater the sensitivity of the data, the greater the invasion of privacy; (ii) the potential injury to patients from an inappropriate disclosure - the greater the potential for injury, the greater the invasion of privacy; and (iii) the appropriateness of consent for disclosing the data - the less appropriate the consent, the greater the potential invasion of privacy.

Third, motives and capacity: this considers the motives and the capacity of the data recipient to re-identify the data, considering issues such as conflicts of interest, the potential for financial gain from a re-identification, and whether the data recipient has the skills and the necessary resources to re-identify the data (a checklist is available in [67]).

For example, if the mitigating controls are low, which means that the data recipient has poor security and privacy practices, then the re-identification threshold should be set at a lower level. This will result in more de-identification being applied. However, if the data recipient has very good security and privacy practices in place, then the threshold can be set higher.

De-identification methods

The i2b2 project tools allow investigators to query for patients and controls that meet specific inclusion/exclusion criteria [13,69]. This allows the investigator to determine the size of cohorts for a study. The queries return counts of unique patients that match the criteria. If few patients match the criteria, however, there is a high probability of re-identification. To protect against such identity disclosure, the query engine performs several functions. First, random noise from a Gaussian distribution is added to returned counts, and the standard deviation of the distribution is increased as true counts approach zero. Second, an audit trail is maintained and if users are running too many related queries they are blocked. Also, limits are imposed on multiple queries so that a user cannot compute the mean of the perturbed data.

The disclosure of individual-level data from a TRIS is also important, and various de-identification methods can be applied to such data. The de-identification methods that have the most acceptability among data recipients are masking, generalization, and suppression (see below). Other methods, such as the addition of random noise, distort the individual-level data in ways that are sometimes not intuitive and may result in incorrect results if these distortions affect the multivariate correlational structure in the data. This can be mitigated if the specific type of analysis that will be performed is known in advance and the distortions can account for that. Nevertheless, they tend to have low acceptance among health researchers and analysts [5], and certain types of random noise perturbation can be filtered out to recover the original data [70]; therefore, the effectiveness of noise addition can be questioned. Furthermore, perturbing the DNA sequences themselves may obscure relationships or even lead to false associations [71].

Methods that have been applied in practice are described below and are summarized in Table 1.

Masking

Masking refers to a set of manipulations of the directly identifying information in the data. In general, direct identifiers are removed/redacted from the dataset, replaced with random values, or replaced with a unique key (also called pseudonymization) [72]. This latter approach is used in the BioVU project to mask the medical record number using a hash function [15].

Patient names are usually redacted or replaced with false names selected randomly from name lists [73]. Numbers, such as medical record numbers, social security numbers, and telephone numbers, are either redacted or replaced with randomly generated but valid numbers [74]. Locations, such as the names of facilities,

Table 1. Summary of de-identification methods for individual-level data

De-identification method	Techniques	Details
Masking (applied to direct identifiers)	Suppression/redaction	Direct identifiers are removed from the data or replaced with tags
	Random replacement/randomization	Direct identifiers are replaced with randomly chosen values (for example, for names and medical record numbers)
	Pseudonymization	Unique numbers that are not reversible replace direct identifiers
Generalization (applied to quasi-identifiers)	Hierarchy-based generalization	Generalization is based on a predefined hierarchy describing how precision on quasi-identifiers is reduced
	Cluster-based generalization	Individual transactions are empirically grouped or based on predefined utility policies
Suppression (applied to records flagged for suppression)	Casewise deletion	The full record is deleted
	Quasi-identifier deletion	Only the quasi-identifiers are deleted
	Local cell suppression	Optimization scheme is applied to the quasi-identifiers to suppress the fewest values but ensure a re-identification probability below the threshold

would also normally be redacted. Such data manipulations are relatively simple to perform for structured data. Text de-identification tools will also do this, such as the tool used in the BioVU project [15].

Generalization

Generalization reduces the precision in the data. As a simple example of increasing generalization, a patient's date of birth can be generalized to a month and year of birth, to a year of birth, or to a 5 year interval. Allowable generalizations can be specified *a priori* in the form of a generalization hierarchy, as in the age example above. Generalizations have been defined for SNP sequences [75] and clinical datasets [68]. Instead of hierarchies, generalizations can also be constructed empirically by combining or clustering sequences [76] and transactional data [77] into more general groups.

When a dataset is generalized the re-identification probability can be measured afterwards. Records that are considered high risk are then flagged for suppression. When there are many variables the number of possible ways that these variables can be generalized can be large. Generalization algorithms are therefore used to find the best method of generalization. The algorithms are often constrained by a value *MaxSup*, which is the maximum percentage of records in the dataset that can be suppressed. For example, if *MaxSup* is set to 5%, then the generalization algorithm will ignore all possible generalizations that will result in more than 5% of the records being flagged for suppression. This will also guarantee that no more than 5% of the records will have any suppression in them.

Generalization is an optimization problem whereby the algorithm tries to find the optimal generalization for each of the quasi-identifiers that will ensure that the probability of re-identification is at or below the required threshold, the percentage of records flagged for suppression is below *MaxSup*, and information loss is minimized.

Information loss is used to measure the amount of distortion to the data. A simple measure of information loss is how high up the hierarchy the chosen generalization level is. However, this creates difficulties of interpretation, and other more theoretically grounded metrics that take into account the difference in the level of precision between the original dataset and the generalized data have been suggested [5].

Suppression

Usually suppression is applied to the specific records that are flagged for suppression. Suppression means the removal of values from the data. There are three general approaches to suppression: casewise deletion, quasi-identifier removal, and local cell suppression.

Casewise deletion removes the whole patient or visit record from the dataset. This results in the most distortion to the data because the sensitive variables are also removed even though those do not contribute to an increase in the risk of identity disclosure.

Quasi-identifier removal removes only the values about the quasi-identifiers in the dataset. This has the advantage that all of the sensitive information is retained.

Local cell suppression is an improvement over quasi-identifier removal in that fewer values are suppressed. Local cell suppression applies an optimization algorithm to find the least number of values about the quasi-identifiers to suppress [78]. All of the sensitive variables are retained and in practice considerably fewer of the quasi-identifier values are suppressed than in casewise and quasi-identifier deletion.

Available tools

Recent reports have provided summaries of free and supported commercial tools for the de-identification of

structured clinical and administrative datasets [79,80]. Also, various text de-identification tools have recently been reviewed [81], although many of these tools are experimental and may not all be readily available. Tools for the de-identification of genomic data are mostly at the research stage and their general availability and level of support is unknown.

Conclusions

Genomic research is increasingly using clinically relevant data from electronic health records. Research ethics boards will often require patient consent when their information is used for secondary purposes, unless that information is de-identified. I have described above the methods and challenges of de-identifying data when disclosed for such research.

Combined genomic and clinical data can be quite complex, with free form textual or structured representations, as well as clinical data that are cross-sectional or longitudinal, and relational or transactional. I have described current de-identification practices in two genomic research projects, i2b2 and BioVU, as well as more recent best practices for managing the risk of re-identification.

It is easiest to use prescriptive de-identification heuristics such as those in the HIPAA Privacy Rule Safe Harbor standard. However, such a standard provides insufficient protection for the complex datasets referred to here and may result in the disclosure of data with a high probability of re-identification. Even when augmented with data sharing agreements, these agreements may be based on the inaccurate assumption that the data have a low probability of re-identification. Furthermore, notices to patients and consent forms often state that the data will be de-identified when disclosed. Disclosure practices that are based on the actual measurement of the probability of re-identification allow data custodians to better manage their legal obligations and commitments to patients.

Moving forward, several areas will require further research to minimize risks of re-identification of data used for genomic research. For example, improved methods for the de-identification of genome sequences or genomic data are needed. Sequence de-identification methods that rely on generalization that have been proposed thus far will likely result in significant distortions to large datasets [82]. There is also evidence that the simple suppression of the sequence for specific genes can be undone relatively accurately [83]. In addition, the re-identification risks to family members have not been considered here. Although various re-identification attacks have been highlighted [84-86], adequate familial de-identification methods have yet to be developed.

Additional files

Additional file 1. Measuring the probability of re-identification. This file describes metrics and decision rules for measuring and interpreting the probability of re-identification for identity disclosure.

Acknowledgements

The analyses performed on the California state inpatient database and the birth registry of Ontario were part of studies approved by the research ethics board of the Children's Hospital of Eastern Ontario Research Institute. Bradley Malin (Vanderbilt University) reviewed some parts of the draft manuscript, and Elizabeth Jonker (CHEO Research Institute) assisted with the formatting of the manuscript.

Abbreviations

EHR, electronic health record; HIPAA, Health Insurance Portability and Accountability Act; SNP, single nucleotide polymorphism; TRIS, translational research information system.

Competing interests

The author declares that he has no competing interests.

Author details

[1]Children's Hospital of Eastern Ontario Research Institute, 401 Smyth Road, Ottawa, Ontario K1J 8L1, Canada. [2]Pediatrics, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, K1H 8L1, Canada.

Published: 27 April 2011

References

1. Prokosch H, Ganslandt T: **Perspectives for medical informatics. Reusing the electronic medical record for clinical research.** *Methods Inf Med*, 2009 **48:**38-44.
2. Tannen R, Weiner M, Xie D: **Use of primary care electronic medical record database in drug efficacy research on cardiovascular outcomes: Comparison of database and randomized controlled trial findings.** *BMJ* 2009, **338:**b81.
3. McCarty C, Chisholm R, Chute C, Kullo I, Jarvik G, Larson E, Li R, Masys D, Ritchie M, Roden D, Struewing JP, Wolf WA: **The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies.** *BMC Med Genomics* 2011, **4:**13.
4. Ness R: **Influence of the HIPAA privacy rule on health research.** *JAMA* 2007, **298:**2164-2170.
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.** *J Am Med Inform Assoc* 2009, **16:**670-682.
6. Kho M, Duffett M, Willison D, Cook D, Brouwers M: **Written informed consent and selection bias in observational studies using medical records: systematic review.** *BMJ* 2009, **338:**b866.
7. El Emam K, Jonker E, Fineberg A: **The case for deidentifying personal health information.** Social Sciences Research Network 2011 [http://papers.ssrn.com/abstract=1744038]
8. Harris AL, AR; Teschke, KE: **Personal privacy and public health: potential impacts of privacy legislation on health research in Canada.** *Can J Public Health* 2008, **99:**293-296.
9. Kosseim P, Brady M: **Policy by procrastination: secondary use of electronic health records for health research purposes.** *McGill J Law Health* 2008, **2:**5-45.
10. Lowrance W: **Learning from experience: privacy and the secondary use of data in health research.** *J Health Serv Res Policy* 2003, **8 Suppl 1:**2-7.
11. Panel on Research Ethics: **Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (2nd Edition).** 2010 [http://www.pre.ethics.gc.ca/pdf/eng/tcps2/TCPS_2_FINAL_Web.pdf]
12. 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.** *J Med Ethics* 2008, **34:**308-314.
13. Murphy S, Weber G, Mendis M, Gainer V, Chueh H, Churchill S, Kohane I: **Serving the enterprise and beyond with informatics for integrating biology and the bedside.** *J Am Med Inform Assoc* 2010, **17:**124-130.

14. Deshmukh V, Meystre S, Mitchell J: **Evaluating the informatics for integrating biology and the bedside system for clinical research.** *BMC Med Res Methodol* 2009, **9:**70.
15. Roden D, Pulley J, Basford M, Bernard G, Clayton E, Balser J, Masys D: **Development of a large-scale de-identified DNA biobank to enable personalized medicine.** *Clin Pharmacol Ther* 2008, **84:**362-369.
16. Malin B, Karp D, Scheuermann R: **Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research.** *J Investig Med* 2010, **58:**11-18.
17. The Supreme Court of the State of Illinois: **Southern Illinoisan vs. The Illinois Department of Public Health.** Docket No. 98712. 2006 [http://www.state.il.us/court/opinions/supremecourt/2006/february/opinions/html/98712.htm]
18. Hansell S: **AOL removes search data on group of web users.** *New York Times* 8 August 2006 [http://www.nytimes.com/2006/08/08/business/media/08aol.html]
19. Barbaro M, Zeller JrT: **A face is exposed for AOL searcher No. 4417749.** *New York Times* 9 August 2006 [http://www.nytimes.com/2006/08/09/technology/09aol.html]
20. Zeller Jr T: **AOL moves to increase privacy on search queries.** *New York Times* 22 August 2006 [http://www.nytimes.com/2006/08/22/technology/22aol.html]
21. **Ochoa S, Rasmussen J, Robson C, Salib M: Reidentification of individuals in Chicago's homicide database: A technical and legal study. 2001** [http://groups.csail.mit.edu/mac/classes/6.805/student-papers/spring01-papers/reidentification.doc] Archived at [http://www.webcitation.org/5xyAv7j6M]
22. Narayanan A, Shmatikov V: **Robust de-anonymization of large sparse datasets.** In *Proceedings of the 2008 IEEE Symposium on Security and Privacy* 2008:111-125 [http://doi.ieeecomputersociety.org/10.1109/SP.2008.33]
23. Sweeney L: **Computational disclosure control: A primer on data privacy protection.** *PhD thesis.* Massachusetts Institute of Technology, Electrical Engineering and Computer Science department; 2001.
24. Appellate Court of Illinois - Fifth District: **The Southern Illinoisan v. Department of Public Health.** 2004 [http://law.justia.com/cases/illinois/court-of-appeals-fifth-appellate-district/2004/5020836.html]
25. Federal Court (Canada): **Mike Gordon vs. The Minister of Health: Affidavit of Bill Wilson.** Court File No. T-347-06. 2006.
26. El Emam K, Kosseim P: **Privacy interests in prescription records, part 2: patient privacy.** *IEEE Security Privacy* 2009, **7:**75-78.
27. Lowrance W, Collins F: **Ethics. Identifiability in genomic research.** *Science* 2007, **317:**600-602.
28. Malin B, Sweeney L: **Determining the identifiability of DNA database entries.** *Proc AMIA Symp* 2000 **2000**:537-541.
29. Wjst M: **Caught you: threats to confidentiality due to the public release of large-scale genetic data sets.** *BMC Med Ethics* 2010, **11:**21.
30. Uzuner O, Luo Y, Szolovits P: **Evaluating the state-of-the-art in automatic de-identification.** *J Am Med Inform Assoc* 2007, **14:**550-563.
31. El Emam K: **Data anonymization practices in clinical research: a descriptive study.** Health Canada, Access to Information and Privacy Division. 2006 [http://www.ehealthinformation.ca/documents/HealthCanadaAnonymizationReport.pdf]
32. Canadian Medical Association (CMA) Holdings Incorporated: *Deidentification/Anonymization Policy.* Ottawa: CMA Holdings; 2009.
33. UBC Clinical Research Ethics Board, Providence Health Care Research Ethics Board: *Interim Guidance to Clinical Researchers Regarding Compliance with the US Health Insurance Portability and Accountability Act (HIPAA).* Vancouver: University of British Columbia; 2003.
34. Hryanszkiewicz I, Norton M, Vickers A, Altman D: **Preparing raw clinical data for publications: tuidance for journal editors, authors, and peer reviewers.** *BMJ* 2010, **340:**c181.
35. Clause S, Triller D, Bornhorst C, Hamilton R, Cosler L: **Conforming to HIPAA regulations and compilation of research data.** *Am J Health Syst Pharm* 2004, **61:**1025-1031.
36. Lin Z, Owen A, Altman R: **Genomic research and human subject privacy.** *Science* 2004, **305:**183.
37. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson J, Stephan D, Nelson S, Craig D: **Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays.** *PLoS Genet* 2008, **4:**e1000167.
38. Jacobs K, Yeager M, Wacholder S, Craig D, Kraft P, Hunter D, Paschal J, Manolio T, Tucker M, Hoover R, Thomas GD, Chanock SJ, Chatterjee N: **A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies.** *Nat Genet* 2009, **41:**1253-1257.
39. Loukides G, Denny J, Malin B: **The disclosure of diagnosis codes can breach research participants' privacy.** *J Am Med Inform Assoc* 2010, **17:**322-327.
40. Willenborg L, de Waal T: *Statistical Disclosure Control in Practice.* New York: Springer-Verlag; 1996.
41. Willenborg L, de Waal T: *Elements of Statistical Disclosure Control.* New York: Springer-Verlag; 2001.
42. Skinner CJ: **On identification disclosure and prediction disclosure for microdata.** *Statistica Neerlandica* 1992, **46:**21-32.
43. Marsh C, Skinner C, Arber S, Penhale B, Openshaw S, Hobcraft J, Lievesley D, Walford N: **The case for samples of anonymized records from the 1991 census.** *J R Stat Soc A (Statistics in Society)* 1991, **154:**305-340.
44. Dale A, Elliot M: **Proposals for 2001 samples of anonymized records: an assessment of disclosure risk.** *J R Stat Soc A (Statistics in Society)* 2001, **164:**427-447.
45. Flora Felso JT, Wagner GG: **Disclosure limitation methods in use: results of a survey.** In *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies. Volume 1*. Edited by Doyle P, Lane J, Theeuwes J, Zayatz L. Washington, DC: Elsevier; 2003:17-38.
46. Jabine T: **Statistical disclosure limitation practices of United States statistical agencies.** *J Official Stat* 1993, **9:**427-454.
47. Pulley J, Brace M, Bernard G, Masys D: **Evaluation of the effectiveness of posters to provide information to patients about a DNA database and their opportunity to opt out.** *Cell Tissue Banking* 2007, **8:**233-241.
48. El Emam K: **Risk-based de-identification of health data.** *IEEE Security Privacy* 2010, **8:**64-67.
49. Subcommittee on Disclosure Limitation Methodology - Federal Committee on Statistical Methodology: **Working paper 22: Report on statistical disclosure control.** Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget. 1994 [http://www.fcsm.gov/working-papers/wp22.html]
50. Manitoba Center for Health Policy: **Manitoba Center for Health Policy Privacy code.** 2002 [http://umanitoba.ca/faculties/medicine/units/mchp/media_room/media/MCHP_privacy_code.pdf]
51. **Cancer Care Ontario: Cancer Care Ontario Data Use and Disclosure Policy. 2005,Updated 2008** [http://www.cancercare.on.ca/common/pages/UserFile.aspx?fileId=13234]
52. Health Quality Council: *Security and Confidentiality Policies and Procedures.* Saskatoon: Health Quality Council; 2004.
53. Health Quality Council: *Privacy code.* Saskatoon: Health Quality Council; 2004.
54. Statistics Canada: **Therapeutic abortion survey.** 2007 [http://www.statcan.ca/cgi-bin/imdb/p2SV.pl?Function=getSurvey&SDDS=3209&lang=en&db=IMDB&dbg=f&adm=8&dis=2#b9]. Archived at [http://www.webcitation.org/5VkcHLeQw]
55. **Office of the Information and Privacy Commissioner of British Columbia: Order No. 261-1998. 1998** [http://www.oipc.bc.ca/orders/1998/Order261.html]
56. Office of the Information and Privacy Commissioner of Ontario: **Order P-644.** 1994 [http://www.ipc.on.ca/images/Findings/Attached_PDF/P-644.pdf]. Archived at [http://www.webcitation.org/5inrVJyQp]
57. Alexander L, Jabine T: **Access to social security microdata files for research and statistical purposes.** *Social Security Bulletin* 1978, **41:**3-17.
58. Ministry of Health and Long Term care (Ontario): **Corporate Policy 3-1-21.** 1984 [Available on request]
59. Duncan G, Jabine T, de Wolf S: *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics.* Washington DC: National Academies Press; 1993.
60. de Waal A, Willenborg L: **A view on statistical disclosure control for microdata.** *Survey Methodol* 1996, **22:**95-103.
61. Office of the Privacy Commissioner of Quebec (CAI): **Chenard v. Ministere de l'agriculture, des pecheries et de l'alimentation (141).** CAI 141. 1997 [Available on request]
62. National Center for Education Statistics: *NCES Statistical Standards.* Washington DC: US Department of Education; 2003.
63. 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.** V.101907(15). 2007.
64. Sweeney L: **Data sharing under HIPAA: 12 years later.** Workshop on the HIPAA Privacy Rule's De-Identification Standard. 2010

[http://www.hhshipaaprivacy.com/]
65. Lafky D: **The Safe Harbor method of de-identification: an empirical test.** Fourth National HIPAA Summit West. 2010 [http://www.ehcca.com/presentations/HIPAAWest4/lafky_2.pdf]. Archived at [http://www.webcitation.org/5xA2HlOmj]
66. Jabine T: **Procedures for restricted data access.** *J Official Stat* 1993, **9:**537-589.
67. El Emam K, Brown A, AbdelMalik P, Neisa A, Walker M, Bottomley J, Roffey T: **A method for managing re-identification risk from small geographic areas in Canada.** *BMC Med Inform Decis Mak* 2010, **10:**18.
68. El Emam K, Dankar F, Vaillancourt R, Roffey T, Lysyk M: **Evaluating patient re-identification risk from hospital prescription records.** *Can J Hospital Pharmacy* 2009, **62:**307-319.
69. Murphy S, Chueh H: **A security architecture for query tools used to access large biomedical databases.** *Proc AMIA Symp* 2002:552-556 [http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2244204/pdf/procamiasymp00001-0593.pdf]
70. Kargupta H, Datta S, Wang Q, Sivakumar K: **Random data perturbation techniques and privacy preserving data mining.** *Knowledge Information Systems* 2005, **7:**387-414.
71. Malin B, Cassa C, Kantarcioglu M: **A survey of challenges and solutions for privacy in clinical genomics data mining.** In *Privacy-Preserving Knowledge Discovery.* Edited by Bonchi F, Ferrari E. New York: Chapman & Hall/CRC Press; 2011.
72. El Emam K, Fineberg A: **An overview of techniques for de-identifying personal health information.** Access to Information and Privacy Division of Health Canada. 2009 [http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1456490]
73. Tu K, Klein-Geltink J, Mitiku T, Mihai C, Martin J: **De-identification of primary care electronic medical records free-text data in Ontario, Canada.** *BMC Med Inform Decis Mak* 2010, **10:**35.
74. El Emam K, Jonker E, Sams S, Neri E, Neisa A, Gao T, Chowdhury S: **Pan-Canadian de-identification guidelines for personal health information.** Privacy Commissioner of Canada. 2007 [http://www.ehealthinformation.ca/documents/OPCReportv11.pdf]
75. Lin Z, Hewett M, Altman R: **Using binning to maintain confidentiality of medical data.** *Proc AMIA Symp* 2002:454-458 [http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2244360/pdf/procamiasymp00001-0495.pdf]
76. Malin B: **Protecting genomic sequence anonymity with generalization lattices.** *Methods Inf Med* 2005, **44:**687-692.
77. Loukides G, Gkoulalas-Divanis A, Malin B: **Anonymization of electronic medical records for validating genome-wide association studies.** *Proc Natl Acad Sci U S A* 2010, **107:**7898-7903.
78. Aggarwal G, Feder T, Kenthapadi K, Motwani R, Panigrahy R, Thomas D, Zhu A: **Anonymizing tables.** In *Proceedings of the 10th International Conference on Database Theory (ICDT05).* Springer; 2005:246-258.
79. Fraser R, Willison D: **Tools for De-Identification of Personal Health Information.** Canada Health Infoway. 2009 [http://www2.infoway-inforoute.ca/Documents/Tools_for_De-identification_EN_FINAL.pdf]. Archived at [http://www.webcitation.org/5xA2KBoMm]
80. Health System Use Technical Advisory Committee - Data De-Identification Working Group: **'Best Practice' Guidelines for Managing the Disclosure of De-Identified Health Information.** 2011 [http://www.ehealthinformation.ca/documents/Data%20De-identification%20Best%20Practice%20Guidelines.pdf]. Archived at [http://www.webcitation.org/5x9w6635d]
81. Meystre S, Friedlin F, South B, Shen S, Samore M: **Automatic de-identification of textual documents in the electronic health record: a review of recent research.** *BMC Med Res Methodol* 2010, **10:**70.
82. Aggarwal C: **On k-anonymity and the curse of dimensionality.** In *Proceedings of the 31st International Conference on Very Large Data Bases.* VLDB Endowment; 2005:901-909.
83. Nyhold D, Yu C, Visscher P: **On Jim Watson's APOE status: genetic information is hard to hide.** *Eur J Hum Genet* 2008, **17:**147-149.
84. Malin B: **Re-identification of familial database records.** *Proc AMIA Symp* 2006:524-528 [http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1839550/pdf/AMIA2006_0524.pdf]
85. Cassa C, Schmidt B, Kohane I, Mandl K: **My sister's keeper ? Genomic research and the identifiability of siblings.** *BMC Med Genomics* 2008, **1:**32.
86. Bieber F, Brenner C, Lazer D: **Finding criminals through DNA of their relatives.** *Science* 2006, **312:**1315-1316.
87. Pabrai U: *Getting Started with HIPAA.* Boston: Premier Press; 2003.

doi:10.1186/gm239
Cite this article as: El Emam K: Methods for the de-identification of electronic health records for genomic research. *Genome Medicine* 2011, **3:**25.

K. El Emam: Methods for the de-identification of electronic health records for genomic research.
Genome Medicine

Appendix: Measuring the Probability of Re-identification

Khaled El Emam[1,2]

[1]Children's Hospital of Eastern Ontario Research Institute, 401 Smyth Road, Ottawa, Ontario K1J 8L1, Canada.

[2]Pediatrics, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada.

The application of de-identification algorithms in practice requires the data custodian to be able to measure the probability of re-identification. Such measurement will inform the custodian whether the probability of re-identification is high or not. If the probability is high then de-identification methods need to be applied. This means that specific metrics for the measurement of the probability of re-identification are needed. In this appendix, we present a set of metrics and decision rules for measuring and interpreting the probability of re-identification for identity disclosure. We assume that we are measuring a disclosed dataset consisting of quasi-identifiers (columns) and records (rows). Each record pertains to a different individual.

Simple and Derived Metrics

When we measure re-identification risk for a dataset, we assign a probability of successful re-identification to each record in that dataset. For identity disclosure, the probability of re-identification means the probability of that record being assigned a correct identity. We will denote the probability of a record i being correctly re-identified by θ_i where $i = 1,\ldots,n$ and n is the total number of records in the dataset. Based on that simple metric, a number of derived metrics can be defined.

All the records that share the same values on a set of quasi-identifiers are called an *equivalence class*. For example, if the quasi-identifiers were age, gender, and date of admission, then all the records in a dataset about 17 year old males admitted on 1st January 2008 are an equivalence class. Equivalence class sizes for a data concept (such as age) potentially change during de-identification. For example, there may be 3 records for 17 year old males admitted on 1st January 2008. When the age is recoded to a five year interval, then there may be 8 records for males between 16 and 20 years old admitted on 1st January 2008.

Let J be the set of equivalence classes in the disclosed dataset, and $|J|$ be the number of equivalence classes in the dataset. In practice, all of the records in the same equivalence class will have the same probability value, θ_i. Therefore, we refer to the probability θ_j for an equivalence class where $j \in J$.

All of the derived metrics generalize this simple individual equivalence class metric to the whole dataset. For the sake of consistency we will scale all of the derived metrics to have a value between zero and one.

Derived metrics need to be converted to a binary value to reflect whether the probability is considered too high or not. The decision that needs to be made at the end of the measurement is a binary one after all. This is the *decision rule* for interpreting the metric value.

The decision rule converts measurements on a derived metric into a re-identification risk. Hence, if the decision rule determines that the risk is HIGH then de-identification methods would be necessary to bring it to LOW. If the decision rule determines that the risk is LOW, then no de-identification is necessary.

This is achieved by the use of thresholds. The thresholds are used to decide whether the derived metric is HIGH or LOW.

The first derived metric assesses the proportion of records that have a re-identification probability higher than a threshold. The threshold is denoted by τ:

$$R_a = \frac{1}{n}\sum_{j \in J} f_j \times I\left(\theta_j > \tau\right) \qquad \text{.....(1)}$$

where $I(\cdot)$ is the indicator function and f_j is the size of equivalence class j in the dataset. If the value of R_a is too high then the dataset is considered to have a risk of re-identification that is not acceptable.

We will denote the threshold for the maximum proportion of records that have a high probability of re-identification by α . Therefore, the decision rule for R_a is:

$$D_a = \begin{cases} HIGH & , R_a > \alpha \\ LOW & , R_a \leq \alpha \end{cases} \quad \text{…..……..(2)}$$

If the R_a value is higher than the threshold, the re-identification risk is considered HIGH. As will be noted, the D_a decision rule requires more than one threshold to be operationalized, τ and α .

Another kind of derived metric takes the worst case scenario and assumes that the equivalence class with the highest re-identification probability represents the whole dataset.

$$R_b = \max_{j \in J} \left(\theta_j \right) \quad \text{……(3)}$$

This is quite a stringent standard because even records with a low probability of re-identification are penalized as much as the records with a high probability of re-identification. However, there will be situations where this metric will be appropriate. The decision rule for R_b is:

$$D_b = \begin{cases} HIGH & , R_b > \tau \\ LOW & , R_b \leq \tau \end{cases} \quad \text{……..…..(4)}$$

Note that the threshold for this decision rule is the same as for equation (1) because that threshold has the same meaning: the maximum acceptable probability of a record being re-identified.

Another derived metric takes the average probability across all of the records in the dataset. It is essentially the expected value. This expected value is given by:

$$R_c = \frac{1}{n} \sum_{j \in J} f_j \theta_j \quad \text{……...(5)}$$

The decision rule for this type of derived metric is given by:

$$D_c = \begin{cases} HIGH & , R_c > \lambda \\ LOW & , R_c \leq \lambda \end{cases} \qquad \text{……...(6)}$$

where λ represents the maximum proportion of records that can be correctly re-identified.

The R_c metric gives the proportion of records that would be re-identified by an adversary (on average). On the surface this would seem to be similar to the R_a metric, which measures the proportion of records that have a high probability of re-identification. However, the records accounted for in R_a are not necessarily those that will be re-identified. For example, if we say that a dataset has 100 records, $\tau = 0.2$, $\alpha = 0$, $\lambda = 0.2$ and only two records have a probability of re-identification $\theta_j = 0.3$, with the rest having $\theta_j = 0.1$. When we use the R_a metric we are assuming that the adversary will try to re-identify a single record and that the adversary may select one of those two records with $\theta_j > \tau$ to re-identify. In such a case the risk would be unacceptable because two records would have a high probability of re-identification if they were selected (i.e., $D_a = HIGH$). On the other hand, when we use the R_c metric we are assuming that the adversary will try to re-identify *all* of the records in the dataset, e.g. by matching them against a registry, and that s/he will likely get ten correct matches. Because that is a smaller proportion than λ, it would be considered acceptable risk (i.e., $D_c = LOW$).

When discussing re-identification metrics it is therefore important to be clear about what kind of metric we are talking about and the decision rule that is being used. Equally important, we need to be clear about the thresholds being used because these are instrumental for interpreting the results. There will be a great difference in interpreting the results, say, using $\tau = 0.2$ versus using $\tau = 0.05$.

A summary of the derived metrics is provided in Table 1, a summary of the decision rules is provided in Table 2, and a summary of the thresholds and their interpretation is provided in Table 3.

Table 1: A summary of derived re-identification metrics.

Derived Risk Metric	Interpretation
$R_a = \frac{1}{n}\sum_{j \in J} f_j \times I\left(\theta_j > \tau\right)$	The proportion of records that have a re-identification probability higher than a threshold
$R_b = \max_{j \in J}\left(\theta_j\right)$	The maximum probability of re-identification in the dataset among all records
$R_c = \frac{1}{n}\sum_{j \in J} f_j \theta_j$	The proportion of records that can be correctly re-identified

Table 2: A summary of the decision rules.

Decision Rule	Interpretation of HIGH/LOW Risk
$D_a = \begin{cases} HIGH & , R_a > \alpha \\ LOW & , R_a \leq \alpha \end{cases}$	The proportion of records with a high probability of re-identification is not acceptable / acceptable
$D_b = \begin{cases} HIGH & , R_b > \tau \\ LOW & , R_b \leq \tau \end{cases}$	The highest probability of re-identification among all records is not acceptable / acceptable
$D_c = \begin{cases} HIGH & , R_c > \lambda \\ LOW & , R_c \leq \lambda \end{cases}$	The proportion of records that can be re-identified is not acceptable / acceptable

Table 3: A summary of thresholds.

Threshold	Interpretation
τ	The highest allowable probability of correctly re-identifying a single record
α	The proportion of records that have a high probability of re-identification which would be acceptable to the data custodian
λ	The proportion of records that can be correctly re-identified (on average) which would be acceptable to the data custodian

Simple Risk Metrics: Prosecutor and Journalist Risk

There are two simple re-identification metrics, or rather, there are two instantiations of θ_j. The main criterion that differentiates them is whether the adversary can determine whether a particular individual is in the disclosed dataset [1]. This individual is the one that is being re-identified, and we will call that individual the *target*. The target may be a specific individual that the adversary already has background information about. For example, this may be the adversary's neighbor, co-worker, ex-spouse, relative, or a famous person. Or the target may be an individual selected at random from a population list such as a voter registry. For example, if the adversary is a journalist who wishes to embarrass or expose a data custodian, then the journalist may select a target at random because the re-identification of any record will achieve the purpose.

If the adversary can determine whether the target is in the disclosed dataset then this is called *prosecutor risk*. If the adversary does not know or cannot determine whether the target is in the disclosed dataset, then this is called *journalist risk*.

How can the adversary determine if the target is in the disclosed dataset? If any of the three following conditions is true, then the data custodian needs to be concerned about prosecutor risk [1]:

- The disclosed dataset represents the whole population (e.g., a population registry) or has a large sampling fraction from that population. If the whole population is being disclosed then the adversary would have certainty that the target is included. Also, a large sampling fraction means that the target is very likely to be in the disclosed dataset.
- The disclosed dataset is not a population registry but is a sample from a population, *and* the identity of an individual(s) in the disclosed dataset can be easily determined by the adversary. For example, if the disclosed dataset is a sample drug-use survey of teenagers, then a parent would likely know that their teenage son participated because they had to consent for them to participate.
- The disclosed dataset is a sample and the individuals in it self-reveal that they are part of the sample. For example, consider the public release of a clinical trials dataset. Subjects in clinical trials do generally inform their family, friends, and even acquaintances that they are participating in a trial. An acquaintance may attempt to re-identify one of these self-revealing participants' records in the disclosed dataset. Individuals may also disclose information about themselves on their blogs and social networking site pages which may self-reveal that they are part of a study or a disease registry. However, it is not always the case that indi-

viduals know that their record is in a dataset. For example, for studies using existing data where consent has been waived or where individuals provide broad authorization for their data or tissue samples to be used in future research, the individuals may not know that their record is in a specific disclosed dataset, providing them no opportunity for self-revealing their inclusion.

If a dataset does not meet the above criteria, then the data custodian should be concerned about journalist risk and not prosecutor risk (i.e., it is either one or the other, not both). The distinction between the two types of risk is quite important because the way re-identification probability is measured or estimated differs between them.

Measuring Prosecutor Risk

Let us assume that the adversary is attempting to re-identify a specific target, Alice. The adversary also knows that Alice is in the disclosed dataset, and therefore the prosecutor risk criterion is met.

In the example shown in Figure 1, there is an original dataset that contains patient information and some prescription information (DIN – Drug Identification Number assigned by Health Canada). The directly identifying variable, the patient name, is suppressed. De-identification is applied by generalizing the year of birth. We now have a disclosed dataset. The adversary has some background information about Alice, namely, her year of birth and gender. The adversary also knows that Alice is in the disclosed dataset.

Because the adversary knows that Alice is in the disclosed dataset s/he can identify matching records using the year of birth and gender. There are two records that match using Alice's background information. Therefore, the matching equivalence class size is equal to two. Since the adversary does not know which one of these two records pertains to Alice, the adversary will select one at random. Therefore, her probability of re-identification is 0.5.

More generally, the probability of correct re-identification is given by:

$$_p\theta_j = 1/f_j \qquad \text{.....(7)}$$

where f_j is the size of the matching equivalence class in the disclosed dataset.

In practice, the data custodian will not know in advance that the adversary will be targeting Alice. The adversary may target any of the individuals in the disclosed dataset. Therefore, the custodian needs to compute the value of ${}_p\theta_j$ for all of the equivalence classes, as shown in Table 4.

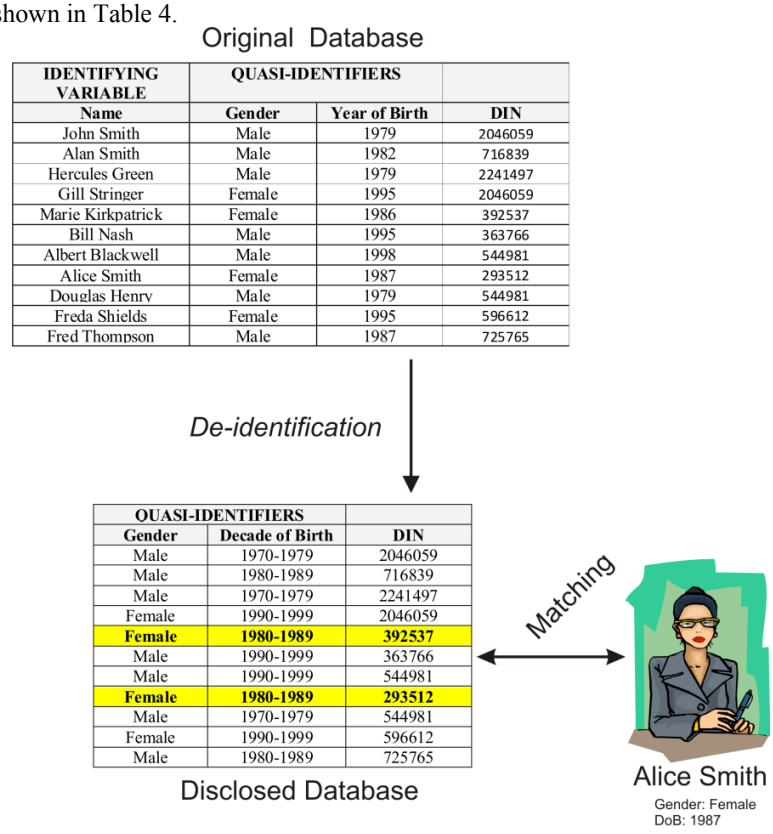

Figure 1: Illustration of prosecutor risk whereby the adversary is attempting to re-identify a record belonging to a specific target individual, Alice, about whom s/he has background information.

Table 4: The prosecutor risk of re-identification for every equivalence class in the example dataset.

Gender	Year of Birth	$_p\theta_j$
Male	1970-1979	0.33
Male	1980-1989	0.5
Male	1990-1999	0.5
Female	1990-1999	0.5
Female	1980-1989	0.5

The custodian must then choose a type of derived metric to compute from the individual $_p\theta_j$ values. The final equations and risk values based on the example in Figure 1 are shown in Table 5 assuming a $\tau = 0.33$.

As can be seen, the risk values can be dramatically different, which makes it important to decide carefully in advance which types of derived risk metrics to use for the assessment of prosecutor risk.

Table 5: The computation of all three types of derived prosecutor risk metrics for our example.

Derived Risk Metric	Equation	Risk Value
$_pR_a$	$\frac{1}{n}\sum_{j\in J} f_j \times I\left(\frac{1}{f_j} > \tau\right)$	0.73
$_pR_b$	$\max_{j\in J}\left(\frac{1}{f_j}\right) = \frac{1}{\min_{j\in J}\left(f_j\right)}$	0.5
$_pR_c$	$\frac{1}{N}\sum_{j\in J} f_j \times \frac{1}{f_j} = \frac{\lvert J\rvert}{N}$	0.45

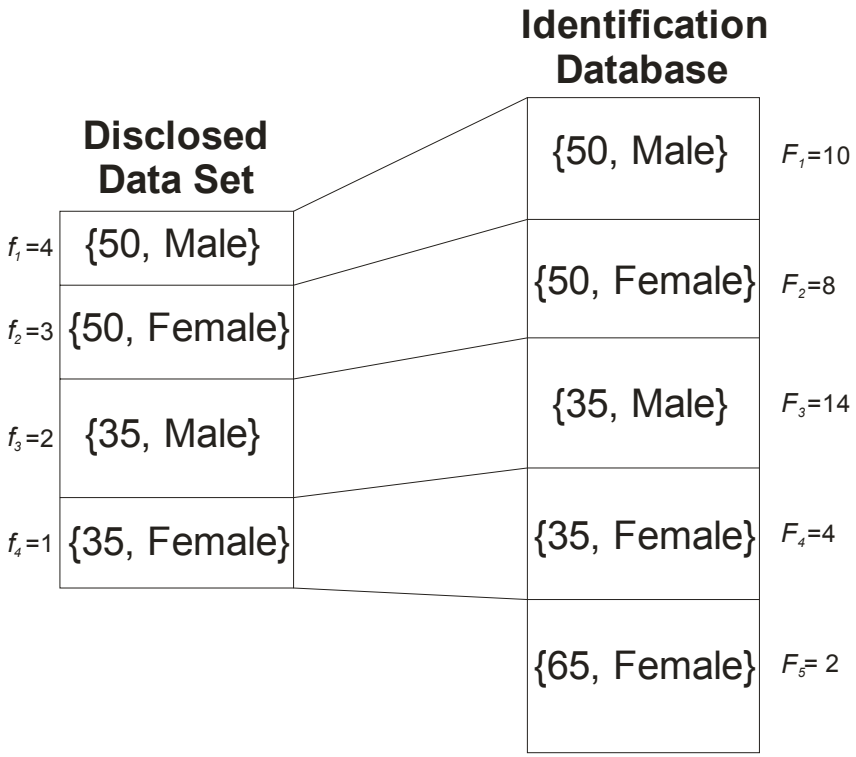

Figure 2: An example of a sample disclosed dataset drawn from an identification database.

Measuring Journalist Risk

For journalist risk to apply the dataset has to be a sample of some sort. The reasoning is that if the disclosed dataset is a population registry, then the adversary will likely know for sure that the target is in the population (because everyone is). Note that the reverse is not necessarily true: if a dataset is a sample that does not mean that journalist risk applies.

There are two general types of re-identification attacks that are under journalist risk: (a) the adversary is targeting a specific individual, and (b) the adversary is targeting any individual. With the former the adversary has background knowledge about a specific individual (e.g., a neighbor or a famous person). Whereas with the latter, the adversary does not care which individual is being targeted.

For journalist risk, we assume that the adversary will match the disclosed dataset with another *identification database.* For now we also assume that there is an identification database that is known to be a superset of the disclosed dataset. For example, the voter registration list in the US is often taken to represent the whole population [2] – although only approximately two thirds of eligible citizens actually register to vote [3]. For our illustrative purposes we will carry on with the common assumption that the voter list covers the whole population.

Under this scenario let K be the set of equivalence classes in the identification database, and $|K|$ is the number of equivalence classes in the identification database. We also have $J \subseteq K$ where $J = \{x | \forall x : x \in K \wedge f_x > 0\}$. Let the number of records in an equivalence class j in the identification database be denoted by F_j, where $F_j > 0$ for $j \in K$, and the total number of records in the identification database is given by $N = \sum_{j \in K} F_j$. The equivalence class size in the disclosed database is given by f_j, and the total number of records in this sample is n. We assume that the disclosed database is a simple random sample from the identification database.

Using the example in Figure 2, we will assume that the adversary wishes to re-identify a specific individual, his neighbor, who is a fifty year old male. The adversary knows that the neighbor is in the identification database. Then we need to consider the probability that the selected fifty year old male is in the disclosed dataset and the probability of a correct match with the background information given that the target is in the disclosed dataset. If we assume that this target individual is in equivalence class j, this gives us a probability $\frac{f_j}{F_j} \cdot \frac{1}{f_j} = \frac{1}{F_j}$. More generally, the probability of a correct match is ${}_J\theta_j = 1/F_j$ for any equivalence class where $f_j > 0$.

In the second type of attack, the adversary does not have background information about a specific individual, but any individual will do. In such a case the adversary

may pick a person at random from the identification database and match against the disclosed dataset. Alternatively, the adversary may select a person at random from the disclosed dataset and match with the identification database. The probability of re-identification under these attacks is also given by ${}_J\theta_j = 1/F_j$.

In practice the data custodian will often not have access to the identification database to compute the journalist risk. Therefore the value of ${}_J\theta_j$ must be estimated using only the information in the disclosed dataset. Various methods for doing this have been developed [1, 4].

The data custodian must then choose a type of derived risk to compute from the individual ${}_J\theta_j$ values. The final equations and risk values for the example in Figure 2 are shown in Table 6 assuming a $\tau = 0.33$.

Table 6: The computation of all three types of derived journalist risk metrics for our example under the scenario of the disclosed dataset being a proper subset of the identification database.

Derived Risk Metric	Equation	Risk Value
${}_J R_a$	$\frac{1}{n}\sum_{j \in J} f_j \times I\left(\frac{1}{F_j} > \tau\right)$	0.053
${}_J R_b$	$\max_{j \in J}\left(\frac{1}{F_j}\right) = \frac{1}{\min_{j \in J}\left(F_j\right)}$	0.5
${}_J R_c$	$\frac{1}{n}\sum_{j \in J} f_j \times \frac{1}{F_j}$	0.12

Summary

A summary of the derived metrics and notes on their interpretation are provided in Table 7. To operationalize some of these metrics with samples may require the use of estimators, which are discussed in more detail in the references.

The $_{\bullet}R_c$ type derived metrics have a special interpretation: the proportion of records in the disclosed dataset that would be correctly re-identified if the adversary tried to match it with an identification database. With that interpretation these derived metrics give us a measure of how many records would be re-identified and we therefore label them as *marketer risk* metrics [4, 5].

Table 7: A summary of the re-identification risk metrics.

Risk Type	Equation	Notes and Conditions
Prosecutor	$_pR_a = \frac{1}{n}\sum_{j\in J} f_j \times I\left(\frac{1}{f_j} > \tau\right)$ $_pR_b = \frac{1}{\min_{j\in J}\left(f_j\right)}$	Where f_j is the size of the equivalence class in the disclosed dataset. If the disclosed dataset is the same as the whole population then $f_j = F_j$, where F_j is the equivalence class size in the population.
Journalist	$_JR_a = \frac{1}{n}\sum_{j\in J} f_j \times I\left(\frac{1}{F_j} > \tau\right)$ $_JR_b = \frac{1}{\min_{j\in J}\left(F_j\right)}$	These metrics are suitable for the situation where the disclosed dataset is a proper subset of the identification database. The value of $1/F_j$ would have to be estimated by the data custodian unless the whole identification database is readily available, in which case F_j can just be counted.
Marketer	$_pR_c = \frac{\lvert J\rvert}{N}$	This metric is suitable if $n = N$ (i.e., the adversary has access to an identification database with records on exactly the same individuals as in the disclosed dataset).
	$_JR_c = \frac{1}{n}\sum_{j\in J}\frac{f_j}{F_j}$	This metric is suitable for the situation where $n < N$ and the disclosed dataset is a proper subset of the identification database $\left(J \subseteq K\right)$. The value of $1/F_j$ would have to be estimated by the data custodian unless the whole identification database is readily available, in which case F_j can just be counted.

References

1. El Emam K, Dankar F: **Protecting privacy using k-anonymity.** *Journal of the American Medical Informatics Association* 2008, **15:**627-637.
2. Benitez K, Malin B: **Evaluating re-identification risks with respect to the HIPAA privacy rule.** *Journal of the American Medical Informatics Association* 2010, **17:**169-177.
3. File T: **Voting and registration in the election of November 2006.** US Census Bureau. 2008; Available from: [http://www.census.gov/prod/2008pubs/p20-557.pdf]
4. Dankar F, El Emam K: **A method for evaluating marketer re-identification risk.** In *Proceedings of the 3rd International Workshop on Privacy and Anonymity in the Information Society (held in conjunction with the 13th International Conference on Extending Database Technology)*; *March 22-26; Lusanne, Switzerland*. Association for Computing Machinery; 2010: a28.
5. El Emam K: **Risk-Based De-Identification of Health Data.** *IEEE Security and Privacy* 2010, **8:**64-67.