Journal of Biomedical Informatics 63 (2016) 174–183

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A unified framework for evaluating the risk of re-identification of text de-identification tools

Martin Scaiano [a,b], Grant Middleton [b], Luk Arbuckle [d], Varada Kolhatkar [a,b], Liam Peyton [a], Moira Dowling [e], Debbie S. Gipson [f], Khaled El Emam [a,b,c,d,*]

[a] *School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, Canada*
[b] *Privacy Analytics Inc., Ottawa, Canada*
[c] *Department of Pediatrics, University of Ottawa, Ottawa, Canada*
[d] *Children's Hospital of Eastern Ontario Research Institute, Ottawa, Canada*
[e] *Michigan Institute for Data Science (MIDAS), University of Michigan Medical School, Office of Research, Ann Arbor, United States*
[f] *Department of Pediatrics, University of Michigan, Ann Arbor, United States*

ARTICLE INFO

Article history:
Received 10 December 2015
Revised 14 June 2016
Accepted 13 July 2016
Available online 15 July 2016

Keywords:
De-identification
Re-identification risk
Medical text
Evaluation framework
Natural language processing
Data sharing

ABSTRACT

Objectives: It has become regular practice to de-identify unstructured medical text for use in research using automatic methods, the goal of which is to remove patient identifying information to minimize re-identification risk. The metrics commonly used to determine if these systems are performing well do not accurately reflect the risk of a patient being re-identified. We therefore developed a framework for measuring the risk of re-identification associated with textual data releases.

Methods: We apply the proposed evaluation framework to a data set from the University of Michigan Medical School. Our risk assessment results are then compared with those that would be obtained using a typical contemporary micro-average evaluation of recall in order to illustrate the difference between the proposed evaluation framework and the current baseline method.

Results: We demonstrate how this framework compares against common measures of the re-identification risk associated with an automated text de-identification process. For the probability of re-identification using our evaluation framework we obtained a mean value for direct identifiers of 0.0074 and a mean value for quasi-identifiers of 0.0022. The 95% confidence interval for these estimates were below the relevant thresholds. The threshold for direct identifier risk was based on previously used approaches in the literature. The threshold for quasi-identifiers was determined based on the context of the data release following commonly used de-identification criteria for structured data.

Discussion: Our framework attempts to correct for poorly distributed evaluation corpora, accounts for the data release context, and avoids the often optimistic assumptions that are made using the more traditional evaluation approach. It therefore provides a more realistic estimate of the true probability of re-identification.

Conclusions: This framework should be used as a basis for computing re-identification risk in order to more realistically evaluate future text de-identification tools.

1. Introduction

There has been significant research on developing tools for the de-identification of free-form medical text [1,2]. The evaluation methods currently used to determine whether these tools are performing well enough are borrowed from the areas of entity extraction and information retrieval [3]. There has been some recognition that these evaluation approaches are not always the most appropriate for measuring the probability of re-identification nor are the benchmarks typically used to decide what is "good enough" directly relevant to the de-identification task [4]. Such concerns triggered the current work.

In this paper we critically examine the methods that are currently used to evaluate medical text de-identification tools [1,2], identify their weaknesses, and propose improvements. We then propose a unified framework for evaluation in terms of the probability of re-identification when medical text is de-identified using automated tools. Our framework builds on existing work, and its

[*] Corresponding author at: Children's Hospital of Eastern Ontario Research Institute, 401 Smyth Road, Ottawa, Ontario K1H 8L1, Canada.

E-mail address: kelemam@uottawa.ca (K. El Emam).

http://dx.doi.org/10.1016/j.jbi.2016.07.015

main contribution is that it brings multiple concepts together from the disclosure control literature, the information retrieval literature, and the risk modeling literature to provide a more detailed evaluation scheme for measuring re-identification risk.

The issues we identify in current evaluation methods can in some instances inflate the performance of de-identification tools by making them look better than they really are, and in other instances may also penalize them by making them seem much worse than they really are. This means that our proposed evaluation framework will not consistently give higher risk values or lower risk values than currently used methods, although we argue that it represents a more accurate modeling of the probability of re-identification because it better accounts for the distribution of identifiers in documents. We illustrate the differences between our framework and conventional evaluation approaches using theoretical and empirical examples. We then illustrate the application of this framework on a clinical data set, and compare the findings to what would be obtained using current evaluation methods.

2. Background

2.1. Evaluation approaches used in text de-identification

Most of the current text de-identification systems treat Personal Health Information (PHI) identification as a named entity recognition problem. Consequently, they evaluate the identification performance with metrics used in the named entity recognition and information retrieval literature [3]. In particular, they typically annotate different types of entities (or categories), such as *date*, *patient name*, and *ID*, and report performance primarily using three metrics: *precision*, *recall*, and *f-measure*. Let *tp* be the number of true positive annotations, *fp* be the number of false positive annotations, and *fn* be the number of false negative annotations. Then, recall *r* is given by

$$r = tp/(tp + fn), \tag{1}$$

and precision *p* is given by

$$p = tp/(tp + fp). \tag{2}$$

Recall and precision answer two questions about a de-identification tool, respectively: “Did we find all that we were looking for?” and “Did we only label what we were looking for?” The metric *f-measure* combines precision and recall, typically by taking the harmonic mean of the two. To get a sense of the overall performance of a system, the most commonly used metrics are *micro-average* and *macro-average* precision, recall, and f-measure. To compute micro-average, one creates a confusion matrix for all categories and then computes precision and recall from this table, giving equal weight to each PHI instance irrespective of its category. To compute macro-average, one computes precision and recall for each category separately and then averages them over all categories, giving equal weight to each category, to get an overall measure of performance.

In Appendix A we summarize evaluation metrics currently used in the text de-identification literature. This review indicates that micro-average recall is a primary metric for evaluating such tools. We also conclude that the number of clinical notes (i.e., number of patients) used in different studies range from 100 to 7193, and that the number of test documents used in different studies range from 220 to 514.

In the context of text de-identification, current evaluation approaches are limited in three ways. First, they report performance on all instances of an entity across all documents. However, none of them consider the number of PHI elements missed within a document, which is an important aspect in de-identification, as a document typically corresponds to a patient and any leaks within a document mean potentially revealing the identity of that patient. In other words, current evaluation approaches do not truly reflect the risk of a patient being re-identified. Second, they evaluate all types of entities with the same evaluation metric, giving equal weight to each entity type even though directly identifying entities, such as name and address, have a higher risk of re-identification compared to indirectly identifying entities, such as age and race. Finally, they do not account for the distribution of PHI across documents. For example, an entity type that is rare and appears in very few documents will have a higher sensitivity to the performance of an information extraction tool than a more prevalent entity type. We examine each of these issues below.

2.2. Basic concepts

The key assumptions that we make in developing our evaluation framework are detailed below. Some of these assumptions are already made in the literature implicitly, but it is important in our context to make them explicit.

2.2.1. One document = one patient

We assume that every document that is being analyzed pertains to an individual patient (i.e., there is a one-to-one mapping between documents and patients). This means that if a document pertains to multiple patients then that information is split into multiple documents. This assumption simplifies the presentation of our framework and its rationale.

In the case where a simple split is not possible, as in the case of clinical study reports from clinical trials, then we assume that all of the information pertaining to an individual trial participant can be extracted as a unit and treated as a separate virtual document for the purposes of evaluation.

This assumption also means that each patient only has one document in the corpus. For example, if the evaluation corpus consists of hospital discharge records, then each patient has a single discharge record.

2.2.2. Information leak = re-identification

Furthermore, we assume that if an annotation is not detected (i.e., “leaked”) then it can be used to re-identify a patient. So the probability of re-identifying a patient is conditional on a leak occurring. We have:

$$Pr(\text{reid}, \text{leak}) = Pr(\text{reid}|\text{leak}) \times Pr(\text{leak}) \tag{3}$$

The probability of a leak in a set of documents is directly related to recall, *r*, given by:

$$Pr(\text{leak}) = 1 - r \tag{4}$$

Based on our assumptions we can then say:

$$Pr(\text{reid}|\text{leak}) = 1 \tag{5}$$

We will examine further below *how much* information needs to be leaked to re-identify a patient. This simplifying assumption is conservative in that it will inflate the risk of re-identification.

2.2.3. Re-identification from correct information extraction

A corollary to the assumption above is that if an annotation is detected, or “caught”, then it is either redacted or re-synthesized, such that the probability of re-identifying a patient from that information is zero.

We can formulate this probability as:

$$Pr(\text{reid}, \text{catch}) = Pr(\text{reid}|\text{catch}) \times Pr(\text{catch}) \tag{6}$$

where $Pr(\text{catch}) = 1 - \text{Pr}(\text{leak})$, which is recall. Clearly the annotations that were leaked versus those that were caught are mutually exclusive. The overall probability of re-identification is therefore given by $Pr(\text{reid}, \text{catch}) + Pr(\text{reid}, \text{leak})$, or:

$$[Pr(\text{reid}|\text{catch}) \times [1 - Pr(\text{leak})]] + [Pr(\text{reid}|\text{leak}) \times Pr(\text{leak})] \tag{7}$$

Which, given the assumption that $\Pr(\text{reid}|\text{leak}) = 1$ in equation (5), simplifies to:

$$Pr(\text{reid}|\text{catch}) + [Pr(\text{leak}) \times (1 - Pr(\text{reid}|\text{catch}))] \tag{8}$$

The above equation represents the overall probability of re-identification from annotations that were detected during information extraction and modified, and those that were leaked. For now, we will assume that $Pr(\text{reid}|\text{catch}) = 0$, which will be valid in most cases where redaction or re-synthesis are used. For specific contexts in which generalization or other transformations are performed on the detected identifiers, such as for documents shared in the context of clinical trials transparency efforts, we drop this assumption and allow $Pr(\text{reid}|\text{catch}) > 0$. The relaxation of this assumption is discussed further in Appendix B.

2.2.4. Distinction between direct and indirect identifiers

As is commonly done in the disclosure control literature [5–7], we consider two types of PHI annotations in text: *direct identifiers* and *quasi-identifiers*. Direct identifiers are annotations such as first name, last name, telephone numbers, unique identifiers (for example, medical record numbers (MRNs) and social security numbers (SSNs)), and email addresses. Quasi-identifiers are annotations that can indirectly identify the patients, such as dates, ZIP codes, city, state, and facility names. Direct and quasi-identifiers are the types of features in health information that are typically targeted during the de-identification of health data [8]. In our analysis we will make a distinction between these two types of annotations because the manner in which they need to be evaluated will differ.

2.2.5. Focus on micro-average recall

Given that our focus is mostly on a unified framework for measuring re-identification risk, recall is most relevant. This does not mean that precision is not important as a metric to evaluate the performance of de-identification tools: only that in the context of the current paper it will not be the focus of our analysis.

Since we do not consider precision further in this paper, we also do not consider the f-measure since it combines recall and precision. We can also see in the literature review in Appendix A that the most commonly used metric for evaluating the risk of re-identification is micro-average recall. Micro-average recall is therefore used as the baseline measure of re-identification risk.

2.3. Critical appraisal of performance evaluation methods

We now consider the weaknesses in conventional approaches to performance evaluation and address these weaknesses. To illustrate some of these points, we use the 2006 i2b2 de-identification challenge data set [2]. The data from this challenge has become a standard for text de-identification evaluation.[1] The data has been manually de-identified "for the challenge by replacing authentic PHI with synthesized surrogates", however the surrogate PHI is not realistic. For our purposes, we used a rule-based de-identification tool described in [9] for our illustrations below.

2.3.1. All-or-nothing recall

Imagine there is an evaluation set of 100 clinical documents, and these documents have 250 different instances of the last name of a patient. Then micro-average recall would be computed across all of these 250 instances. If 230 of the instances were detected by the de-identification tool then the recall would be 0.92 (i.e., 230/250).

The micro-average does not account for the fact that there were 100 documents, and it does not account for how these names were distributed across these documents. This is important because for direct identifiers, the general assumption is that a single instance of a direct identifier is sufficient to determine the identity of the patient. Although one can come up with counter-examples to this assumption (for example, the name "James" would not directly identify a unique patient because it is so common), it is one assumption commonly made in the disclosure control community and errs on the conservative side. The implication of this assumption is that we will be conservative because any single leaked direct identifier is equated with a successful re-identification. If the true re-identification can only happen if two direct identifiers leak then we would be being overly protective.

If a single instance of a direct identifier in a document can reveal the identity of the patient, then all that is needed to reveal the identity of a patient is for a single direct identifier to leak (or not to be detected) in a document. If a document has 10 instances of a patient's last name and 9 of those instances are detected, from a re-identification risk perspective this is not a 90% recall but a 0% recall because there was at least one leak. This is the *all-or-nothing* recall.

To continue with our example, if the 230 names that were correctly detected were all the names in 80 documents, and the remaining 20 names that were not detected were in the other 20 documents (i.e., one name in each document), then the probability of determining the identity of the patient in these 20 documents is almost certain. The micro-average recall of 0.92 inflates the performance of the de-identification tool. The all-or-nothing recall in this case is 0.8, and the correct probability of re-identifying an individual in these documents is then 0.2 instead of 0.08.

Therefore, for direct identifiers it is important to use the all-or-nothing recall value rather than the micro-average recall value [9]. Consider Table 1, which illustrates the magnitude of the differences between micro-average recall and all-or-nothing recall on the i2b2 data set. The "DI" group contains all annotation types that would be classified as direct identifiers. Notice how the micro-average recall remains fairly constant when including many PHI types, while the all-or-nothing recall drops markedly. Adding more annotation types can only add more opportunities to leak values, which leads to monotonically decreasing all-or-nothing recall. However, micro-average, by definition of an average, need not decrease; adding an annotation type with a high recall could increase average, even though the documents previously containing leaks still contain leaks. Micro-average can be extremely misleading about the rate of re-identification for leaked direct identifiers.

2.3.2. Masking recall

During information extraction a particular type of annotation is detected. For example, if there is a "James" in the document then it is identified and then classified as a "First Name". If both of these steps (identification and classification) are true, then this is typically considered a true positive. However, from a de-identification/recall perspective it does not matter whether "James" is classified as a first name or a last name. All that matters is that it has been detected. Of course the classification as a "First Name" may matter from a precision perspective, but it does not matter from a recall perspective.

Consider Table 2 where the annotation provided manually by an expert does not match what a de-identification tool could determine. However, in a redacted document the net effect is the same – the name of the facility will be protected. All the identifying information is removed. Therefore, a more precise recall would consider

[1] https://www.i2b2.org/NLP/DataSets/Main.php – Last accessed on March 31, 2016.

Table 1
Comparison of micro-average recall and all-or-nothing recall on i2b2 data.

	ID	Names	DI
Micro-average recall	0.8396	0.7909	0.8049
All-or-nothing recall	0.5997	0.3618	0.1863

Table 2
Example of a mismatch between expert annotation and the classification by a de-identification tool.

Expert annotation	[ORG]Thompson's Ohio Children's Hospital[/ORG]
De-identification tool	[NAME]Thompson's[/NAME] [STATE]Ohio[/STATE] [ORG]Children's Hospital[/ORG]

the organization completely masked, even though it is masked by several annotations of different types.

Therefore, we define masking recall as the recall value calculated only based on whether a particular direct or quasi-identifier in the text has been detected or not [4] (also called PHI-level evaluation in [10]). Masking recall should use a token level evaluation: evaluate that each token is masked.

Consider Table 3, which shows the comparison of masking recall and conventional recall for different annotation types in the i2b2 data set. For all annotation types, masking recall is markedly higher than the conventional recall. The crucial question from the de-identification perspective is whether we missed a PHI or not. The masking recall more clearly answers this question, as it indicates the extent to which instances of an annotation type were identified as PHI. For example, among all IDs, 83.96% of them were identified as a PHI of some annotation type.

However, a token level evaluation would be problematic if the frequencies of tokens in the data set are not similar. For example, consider a data set with 1000 documents. All 1000 documents have a first name, and only 10 have a last name. The de-identification tool detected the first names in 999 of the 1000 documents and only 2 of the last names. If we pool both names as suggested above the recall would be 1001/1010 = 0.99. This, however completely hides the very low recall on last names because of the extreme imbalance in the frequency of occurrence of each name. Therefore, the concept of masking recall is only appropriate if the frequencies of all of the direct and quasi-identifiers is more-or-less the same in the data set. In practice this cannot be ensured and therefore we need a more robust approach for evaluation.

Table 3
Comparison of masking recall and conventional recall on i2b2 data.

	ID	Names	Organizations	All Quasi-identifiers
Masking recall	0.8396	0.7909	0.4517	0.847
Conventional recall	0.7796	0.7892	0. 3820	0.1649

2.3.3. One or more leaks of direct identifiers

As noted earlier, for direct identifiers we assumed that a leak of a single value in a document would result in the patient being re-identified. To be precise we are concerned about *at least one* of the direct identifiers leaking from the de-identification process. We also need to evaluate this in a manner that accounts for the different frequencies of different types of identifiers. Let s_i be the number of documents that a particular identifier i appears in, and n the total number of documents. Then we can define the probability that a direct identifier is missed or leaks given that it actually appears in the corpus being evaluated as:

$$Pr(\text{leak}, \text{appears}) = Pr(\text{leak}|\text{appears}) \times Pr(\text{appears}). \tag{9}$$

Which gives the probability that a leak will occur given that the identifier actually appears in the data. The probability that direct identifier i leaks and appears in a document is given by:

$$w_i(1 - r_i), \tag{10}$$

where $w_i = s_i/n$ and r_i is the all-or-nothing recall. The probability that a document will leak *at least one* direct identifier is therefore given by:

$$1 - \prod_i (1 - w_i(1 - r_i)) \tag{11}$$

This gives us the combined probability of a leak for all direct identifiers. Since each direct identifier type is dealt with independently, the frequency with which specific direct identifiers appear in the data set will not affect this calculation directly (except when computing the confidence intervals).

2.3.4. Quasi-identifier risk

For quasi-identifiers, a single value is not necessarily uniquely identifying. However, there is evidence that, in a number of jurisdictions, two quasi-identifiers such as the date of birth and the ZIP or postal code, are unique across most of the population [11–15]. For example, that uniqueness approaches 100% in Canada and the Netherlands [11–13], and is closer to 63% in the US [14]. We therefore make the conservative assumption that at least two quasi-identifiers must leak in the same document to re-identify a patient.

Let m be the number of times, on average, that a quasi-identifier value in a document is repeated (i.e., the average number of instances per quasi-identifier value). Also, let r_q be the micro-average recall computed across all quasi-identifiers. Then the probability of at least one quasi-identifier instance being leaked would be given by $1 - (r_q)^m$. This means that the more instances that a quasi-identifier has in a document, the greater the likelihood that there will be a leak.

Finally, let n_q be the average number of distinct quasi-identifier values per document. Since we do not know which two or more quasi-identifiers will be leaked, we need to account for all combinations of 2 or more leaks. This can be represented as a binomial distribution with n_q trials:

$$Pr(X \geqslant 2) \quad \text{for } X \sim B(n_q, 1 - (r_q)^m) \tag{12}$$

where $B(a, b)$ is a binomial distribution with a trials and b probability of success. This is a suitable distribution even when the population is known to be finite. The values for m and n_q are computed from the data.

The expression in equation (12) assumes that the instances for the same quasi-identifier are protected independently. In practice, this is a conservative assumption since the ability to detect one instance of a quasi-identifier could be quite similar across all instances of that quasi-identifier in a document. For example, the recall for a date of birth will be the same for all instances of date of birth. A less conservative approach for modeling of at least two quasi-identifiers leaking would then be $Pr(X \geqslant 2) \quad \text{for } X \sim B(n_q, 1 - r_q)$. We nevertheless err on the conservative side because the recall will also depend on the context in which a quasi-identifier is used and how it is expressed, and that will not necessarily always be the same across all instances. For example, the name of a facility may be "The Ottawa Hospital", "TOH", and "the general hospital in Ottawa" and all of these instances refer to the same quasi-identifier but will have different recall values.

In the i2b2 data set the proportion of documents with at least two leaked quasi-identifiers was 0.3704, and the probability as expressed in equation (12) was 0.467. Therefore, we can see in this example that equation (12) sets an upper bound on the risk and errs on the conservative side.

2.3.5. Re-synthesis recall

It is common practice to replace the elements in text that are annotated by the de-identification tool as direct or quasi- identifiers with fake values. These would be randomly generated values that are substituted for the original values. Such a *re-synthesis* of the original text ensures that the de-identified text looks realistic.

It has been shown that an adversary who attempts to re-identify individuals from a re-synthesized document has difficulty in determining which identifiers are re-synthesized ones versus original ones that were missed by the de-identification tool [16,17]. For example, if the de-identified text has the names "James" and "Alan" in the document, there will be uncertainty as to which one of these is the real name of the patient. For this reason, re-synthesis allows leaks to be *hiding in plain sight*.

The probability that a document will leak at least one *direct identifier* that is recognized by an adversary, and therefore the probability of re-identification, is given by:

$$\begin{aligned} Pr(\text{recognize}, \text{leak}, \text{appears}) = {} & Pr(\text{recognize}|\text{leak}, \text{appears}) \\ & \times Pr(\text{leak}|\text{appears}) \\ & \times Pr(\text{appears}) \end{aligned} \tag{13}$$

Let h be the probability a leaked identifier value is successfully *hiding in plain sight*, i.e., the probability that an adversary can correctly determine whether an identifier is an original one that was leaked versus one what was re-synthesized. The above formulation for *direct identifiers* can be computed as:

$$1 - \prod_i (1 - h \times w_i(1 - r_i)) \tag{14}$$

where r_i is the all-or-nothing recall for direct identifiers. For *quasi-identifiers* we have:

$$Pr(X \geqslant 2) \quad \text{for } X \sim B(n_q, h(1 - (r_q)^m)) \tag{15}$$

Based on previous experiments [16] a reasonable value can be computed as $h = 0.1$, which also errs on the more conservative side given that some studies found that $h = 0$ [17].

2.3.6. Strict recall

Equation (14) could result in quite small values of recall giving seemingly acceptable levels of re-identification probability. For example, if we use $h = 0.1$ from [16], $w_i = 1$, and $r_i = 0.4$, then the overall probability of re-identification with re-synthesis would be 0.06, even though the value of r_i is quite low. Furthermore, with a low value for r_i the density of identifiers that have leaked will be high and it is not clear that the h value from these previous studies would still hold. Therefore, we need to specify a minimal value for the recall values in order to use the re-synthesis adjustment. This adjusts the equations above for those recall values above 0.9, versus those below 0.9. For direct identifiers we have:

$$1 - \prod_{\{i|r_i \geqslant 0.9\}} (1 - h \times w_i(1 - r_i)) \prod_{\{i|r_i < 0.9\}} (1 - w_i(1 - r_i)) \tag{16}$$

In this case we assumed that a high recall of 0.9 for direct identifiers would be necessary for the published h value to hold. We use a slightly lower cutoff value than is reported in the literature [16] because the literature uses micro-average recall all the time rather than all-or-nothing, and this will result in inflated recall values. Therefore, the lower threshold is an attempt to adjust for that.

Note the impact of w, the probability a direct identifier appears in a document, will have on the overall risk from direct identifiers. On the one hand $w < 1$ will decrease risk, possibly even countering for the loss of the factor $h = 0.1$ when recall is below 0.9; on the other hand w will increase variance for recall (which depends on $s_i = n \times w_i$). In order to justify the use of the factor h we need to ensure it is significantly greater than or equal to 0.9 (see the discussion of confidence intervals in Section 2.3.8).

And for quasi-identifiers,

$$\begin{aligned} & Pr(X \geqslant 2 \text{ if } r_q \geqslant 0.7, \text{ or } Y \geqslant 2 \text{ if } r_q < 0.7) \\ & \text{for} \\ & X \sim B(n_q, h(1 - r_q^m)), Y \sim B(n_q, (1 - (r_q)^m)), \end{aligned} \tag{17}$$

where 0.7 is the minimum recall value. This is the value that we have used in our analysis based on our subjective judgement and what would be acceptable to the institution releasing the data in our study, but it is a parameter that can be adjusted by the analyst.

2.3.7. Accounting for attempted attack

If a de-identified text document is going to be disclosed publicly, then the results in equations (16) and (17) would be the correct ones to use. However, for non-public data releases it is necessary to take into account the probability that an adversary will actually attempt to re-identify an individual in the data set [18]. Considering the probability of attempt is common disclosure control practice for health data and has been included in recent guidance and standards [19–22].

This can be modeled as follows for direct identifiers:

$$\begin{aligned} & Pr(\text{reid}, \text{attempt}, \text{leak}, \text{appears}) \\ & \quad = Pr(attempt|leak, appears) \\ & \qquad \times \left(1 - \prod_{\{i|r_i \geqslant 0.9\}} (1 - h \times w_i(1 - r_i)) \prod_{\{i|r_i < 0.9\}} (1 - w_i(1 - r_i))\right) \end{aligned} \tag{18}$$

And for quasi-identifiers:

$$\begin{aligned} & Pr(reid, attempt, leak, appears) = Pr(attempt|leak, appears) \\ & \times Pr(X \geqslant 2 \text{ if } r_q \geqslant 0.7, \text{ or } Y \geqslant 2 \text{ if } r_q < 0.7) \\ & \text{for} \\ & X \sim B(n_q, h(1 - (r_q)^m)), Y \sim B(n_q, (1 - (r_q)^m)) \end{aligned} \tag{19}$$

A scheme based on subjective probability that has been in use for a number of years to evaluate the probability of re-identification for health data has been developed for computing a value for $Pr(attempt|■)$ [8]. This uses checklists to evaluate the security and privacy practices of the data recipient, the types of contractual controls in place, and the motives and (technical and financial) capacity of the data recipient to re-identify the data set.

2.3.8. Confidence intervals

In the literature it has been assumed thus far that the computed recall value is an accurate point estimate, and typically no confidence interval was computed for it. However, because during validation studies the computed value is an estimate of recall, it is important to report the confidence interval around that estimate as well. That confidence interval will be affected by, for example, the sample size of the corpus and the frequency of identifiers in the data.

Therefore the recall can then be represented by a normal distribution with the observed value as the mean and the estimate of the variance would be $r_i(1 - r_i)/s_i$. Similarly, the weight w_i can then be represented by a normal distribution with the observed value as the mean and the estimate of the variance would be $w_i(1 - w_i)/n$.

Because each identifier will have a different frequency in the data, the computations of recall will have different accuracy, and this needs to be accounted for in an evaluation framework. For example, a direct identifier that appears in 1000 documents will have a recall value that is computed more accurately after evaluation than a direct identifier that only appears in 10 documents. We therefore need to account for this uncertainty.

Document frequency and all-or-nothing recall can be treated as proportion estimates; document frequency is the estimated proportion of documents with a particular type of PHI and all-or-nothing recall the estimated proportion of documents correctly annotated. Proportion estimates follow a binomial distribution since they are modeled as Bernoulli trials, however it is common practice to approximate this with a normal distribution [23].

The value of $Pr(\text{attempt})$ can also be represented as a triangular distribution which is a common approach to represent uncertainty with subjective probabilities [24,25]. The counts n_q and m can be represented as Poisson distributions given that there will be variation in their values across documents as well.

The variable weight and recall values can be represented as normal distributions denoted by $N(a, b)$, where a is the mean and b is the standard deviation. The triangular distribution is given by $Triang(a, b, c)$ where b is the most likely value and a and c the minimum and maximum values. Therefore, we can then formulate the overall probability distribution for direct identifiers as follows:

$$
\begin{aligned}
&Pr(\text{reid}, \text{attempt}, \text{leak}, \text{appears}) \\
&= \left(1 - \prod_{\{i|r_i \geqslant 0.9\}} (1 - h \times W_i(1 - R_i)) \prod_{\{i|r_i < 0.9\}} (1 - W_i(1 - R_i))\right) \times A \\
&\text{for} \\
&W_i \sim N(w_i, \sqrt{w_i(1 - w_i)/n}), R_i \sim N(r_i, \sqrt{r_i(1 - r_i)/s_i}), \\
&A \sim \text{Triang}(a, b, c) \qquad (20)
\end{aligned}
$$

And for quasi-identifiers:

$$
\begin{aligned}
&Pr(\text{reid}, \text{attempt}, \text{leak}, \text{appears}) \\
&= Pr(X \geqslant 2 \text{ if } r_q \geqslant 0.7, \text{ or } Y \geqslant 2 \text{ if } r_q < 0.7) \times A \\
&\text{for} \\
&X \sim B(N_q, h(1 - (R_q)^M)), Y \sim B(N_q, (1 - (R_q)^M)), \qquad (21) \\
&A \sim \text{Triang}(a, b, c) \\
&\text{where} \\
&R_q \sim \text{N}(r_q, \sqrt{r_q(1 - r_q)/s_q}), N_q \sim \text{Pois}(n_q), M \sim \text{Pois}(m)
\end{aligned}
$$

The distribution of the terms in equations (20) and (21) can be computed using a Monte Carlo simulation and the 95% confidence interval for the overall probability of re-identification derived from that empirical distribution [25].

2.3.9. Setting thresholds

In this section we discuss how to evaluate the re-identification probability distribution by comparing it to an appropriate threshold for each of the direct and quasi- identifiers.

2.3.9.1. Evaluating the distribution for direct identifiers. For direct identifiers, we create a benchmark or *threshold distribution* and compare the actual distribution obtained from this data with that threshold distribution. This threshold distribution is derived from existing practices in the literature. If the actual distribution does not cover a risk greater than what is covered by the threshold distribution, then we have sufficient evidence to conclude that the actual risk is the same as or lower than the threshold risk and is therefore considered acceptably low. This is illustrated in panel (a) of Fig. 1. In other words we need the upper confidence limit of the actual distribution to be less than or equal to the upper confidence limit of the threshold distribution. Otherwise we cannot conclude that the risk is lower than the threshold distribution, or that the risk is acceptably low. As illustrated in panel (b) of Fig. 1, the upper confidence limit of the actual distribution is greater than the upper confidence limit of the benchmark distribution, and therefore we cannot conclude that the actual risk is less than or equal to the benchmark distribution.

This can be thought of in terms of a null hypothesis, where the actual risk is greater than the benchmark distribution. In panel (a) we can reject this null hypothesis and conclude that the actual risk is not greater than the threshold distribution, but in panel (b) there is insufficient evidence to reject it and we therefore conclude that the actual risk may be greater than the threshold distribution.

For the benchmark distribution we need to determine an acceptable recall for direct identifiers that will result in a measure of risk that is equivalent to existing standards. The authors in [26] recommended that a recall of at least 0.95 would be acceptable for direct identifiers. We have extended this criteria to all-or-nothing recall, which is more conservative than these authors had intended since they were referring to micro-average recall.

When constructing the benchmark distribution we assume that $w = 1$, the worst case in terms of risk in that it assumes that all of the direct identifiers are present in each document. By examining the literature review in Appendix A we see that the smallest data set that was used to evaluate a rule-based de-identification tool or the testing data set for a machine-learning based tool was 220 documents. We therefore assume that $n = 220$ for the benchmark distribution.

When we put these values into equation (20) we obtain a conservative benchmark probability distribution that reflects what has been considered acceptable performance for the detection and removal of direct identifiers. Note that if a particular data set has $n < 220$ then this would result in an actual confidence interval that is wider than the benchmark distribution, increasing the chance that the actual risk may cover a risk that is greater than the benchmark distribution. Therefore, we do not set minimal data set sizes for evaluations because that is already accounted for.

When $w < 1$ the overall risk from direct identifiers will decrease, but this will also increase variability because recall depends on $s_i = n \times w_i$. In this case the actual distribution may cover a risk that is greater than the benchmark distribution, and we would not conclude that the risk is acceptably low.

Now that we have a conservative benchmark distribution, we can perform the comparisons illustrated in Fig. 1 to determine if the actual distribution covers a risk greater than the benchmark, and therefore decide if the actual re-identification risk is acceptable.

2.3.9.2. Evaluating the distribution for quasi-identifiers. Previous work has suggested a fixed 85% recall threshold for quasi-identifiers in the automated de-identification literature [27]. However, a fixed recall value for quasi-identifiers would be quite inconsistent with how the re-identification risk from quasi-identifiers in structured data sets are evaluated, as illustrated below.

The benchmark for acceptable probability of re-identification is determined by a threshold computed from the sensitivity of the data, the potential subjective and objective harm that can affect a patient if there was an inappropriate disclosure of their data or re-identification, and the extent to which the patient had consented for their information to be used for the anticipated secondary purposes [8,28]. These are the same criteria that are used to determine the acceptable probability of re-identification for quasi-identifiers in structured data sets.

When considering these three criteria, there are some strong precedents for choosing a probability value that is acceptable. Historically, data custodians have used the “cell size of five” rule as a threshold for deciding whether data has a low risk of re-identification [29–44]. This rule has been applied originally to

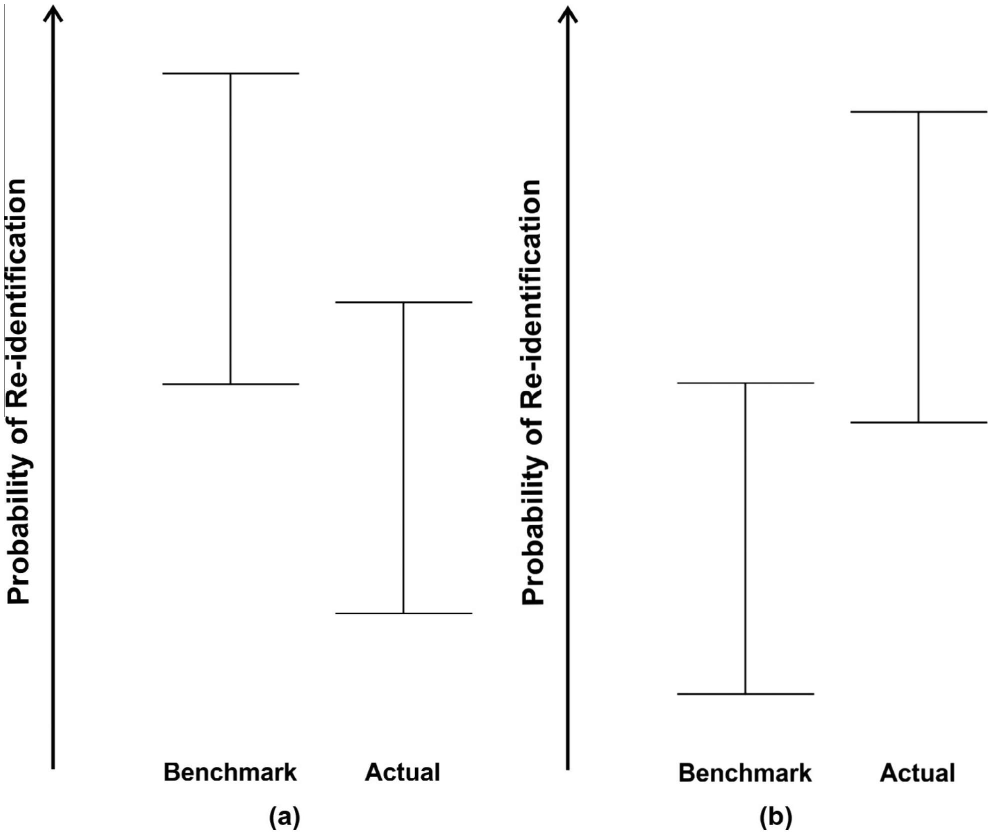
Probability of Re-identification
Benchmark
Actual
(a)
Probability of Re-identification
Benchmark
Actual
(b)

Fig. 1. Illustration of comparing the benchmark with the actual distributions when (a) the actual risk is no worse that the acceptable risk defined by the benchmark distribution and (b) the actual risk may be worse than the acceptable risk defined by the benchmark distribution.

count data in tables. Count data, however, can be easily converted to individual-level data—therefore these two representations are in effect the same thing. A minimum "cell size of five" rule would translate into a maximum probability of re-identifying a single record of 0.2. Some custodians use a cell size of 3 [45–49], which is equivalent to a probability of re-identifying a single record of 0.33. For the public release of data a cell size of 11 has been used in the US [50–54], and a cell size of 20 for public Canadian and US patient data [55,56]. Cell sizes from 5 to 30 have been used across the US to protect student's personally identifying information [57]. Other cell sizes such as 4 [58], 6 [59–62], 10 [63], 16 [63], and 20 [63] have been used in different scenarios within varying countries.

Once an appropriate value is determined from within this range using the three criteria and the checklist and scoring scheme in [8], we can derive the following inequality from equation (21):

$$\begin{aligned} &Pr(X \geqslant 2 \text{ if } r_q \geqslant 0.7, \text{ or } Y \geqslant 2 \text{ if } r_q < 0.7) \times A \leqslant \tau \\ &\text{for} \\ &X \sim B(N_q, h(1-(R_q)^M)), Y \sim B(N_q, (1-(R_q)^M)), \\ &A \sim \text{Triang}(a,b,c) \\ &\text{where} \\ &R_q \sim \text{N}\left(r_q, \sqrt{r_q(1-r_q)/s_q}\right), N_q \sim \text{Pois}(n_q), M \sim \text{Pois}(m) \end{aligned} \tag{22}$$

and τ is the threshold probability. If the inequality is met then the risk of re-identification is considered acceptable. The upper confidence limit of the 95% confidence interval needs to be below the threshold value to be able to conclude that the risk is acceptably small.

2.3.10. Summary

The framework that we have presented above for calculating the probability of re-identification from a de-identified text document provides more precise modeling of the risks from an adversary. They may result in higher probability calculations than under existing approaches in some instances, or smaller values in other instances. Nevertheless, they represent a more accurate way to assess the probability of re-identification than current approaches.

We have also presented techniques to account for the uncertainty in the estimated values and comparing the computed risk values to benchmarks or thresholds in a formal manner. This would allow a precise determination of whether the actual probability of re-identification is acceptably small. These techniques account for the corpus size that is used to perform the evaluations.

The notation used in formulating our framework is summarized in Table 4. The application of the overall model in a hypothetical context is described in Appendix C, which shows how the equations can be used in practice.

3. Methods

3.1. Data set

Our purpose in the empirical application of the evaluation framework is to illustrate its use on a real data set, and show how to interpret the results. We applied the evaluation framework to a data set from the University of Michigan Medical School. The data comes in four groups, one is a random assortment of documents from the full collection of over 80 million, while the other

Table 4
Summary of notation.

Notation	Definition
s_i	The number of documents that a particular direct identifier *i* appears in
r_i	The all-or-nothing recall for direct identifier *i*
n	The number of documents
r_q	The micro-average recall computed across all quasi-identifiers
m	The average number of times that a quasi-identifier value in a document is repeated – the average number of instances per quasi-identifier value
n_q	The average number of unique quasi-identifier values per document
s_q	The number of documents that a quasi-identifier appears in. In most cases this will be the same as *n*
h	The HIPS factor

three are a stratified random sample of three documents types: Social Work Notes, History and Physical Notes, and Progress Notes.

Each document is between 1 and 2 pages in length and has different emphasis that is evident in the content and organization of the document. The random group allows us to analyze each stratum against a general representation of the overall corpus.

There are 30 documents in each group for a total of 120 expert annotated documents. The entire corpus was annotated by a single expert, and subsequently reviewed by a second expert. Where there was disagreement the two experts met and reached consensus on the appropriate annotation to use.

3.2. De-identification

The de-identification was performed with the rule-based engine that was described elsewhere [9], Ch. Free–Form Text. Because this was a rule-based de-identification engine, no training data set was required to construct a model before applying it. The de-identification engine was applied “out-of-the-box” without modification or customization.

The set of direct and quasi- identifiers that were targeted for extraction in these documents are consistent with those that are typically used in the literature [1]. These include: ID’s, phone numbers, people names, email addresses, street addresses, organization names, ZIP codes, ages, country, state, and city.

We will compare our risk assessment results with those that would be obtained using a typical contemporary micro-average evaluation of recall. This will illustrate the difference between the proposed evaluation framework and the current baseline.

3.3. Risk thresholds

For the purposes of our case study, we will use a threshold based on the commonly used “cell size of five” rule, which is equivalent to a probability of re-identification of 0.2 for quasi-identifiers. The upper confidence limit of the quasi-identifier confidence interval needs to be below that value. In the case of direct identifiers the data confidence interval is compared with the benchmark confidence interval.

4. Results

4.1. Data summary

Table 5 contains information on the data element type (annotation) frequency by document and the number of instances of annotations found in the corpus. The table refers to particular annotation sets: the gold standard which was expertly annotated and reviewed. The “document” column indicates the number of documents containing that annotation, while the “annotations” column represents the number instances (individual annotations).

Table 5
Number of annotations in the evaluation corpus.

	#Documents	#Annotations
ID	105	438
Phone number	34	78
Street name	10	14
Names (first, last, middle)	118	1332
Dates	111	703
Organizations	50	110
Age	2	3
ZIP code	9	13
Country	2	3
State	32	47
City	43	78

Table 6
Summary of results that would be obtained by a more traditional micro-average recall calculation (and the leak rate, which is one minus the recall).

	Micro-average recall	Probability of a leak
Direct identifiers	0.9758	0.0242
Quasi-identifiers	0.8757	0.1243

4.2. Evaluation results

The evaluation results are split into two sets. First, the results using a more traditional micro-average recall are shown in Table 6.

In the second set of results we show the 95% confidence intervals for the probability of re-identification using our evaluation framework. In this case we have a mean value for direct identifiers of 0.0074 and a mean value for quasi-identifiers of 0.0022. The confidence interval for the direct identifiers from the data, and compared to the benchmark, is illustrated in Fig. 2. This shows that the upper confidence limit for the re-identification risk from the data is below the upper confidence limit of the benchmark distribution, and therefore we can conclude that the risk of re-identification for direct identifiers is acceptably small.

In Fig. 3 we show the 95% confidence interval for the quasi-identifiers. The upper confidence limit is below the 0.2 threshold that we are using in our example. Therefore we can conclude that the risk of re-identification for quasi-identifiers is acceptably small.

The comparison of these two sets of results shows that the numeric outcomes of the evaluation will be different, and that our evaluation framework, because it takes context into account, will often be less pessimistic about the real risks. Again, as noted

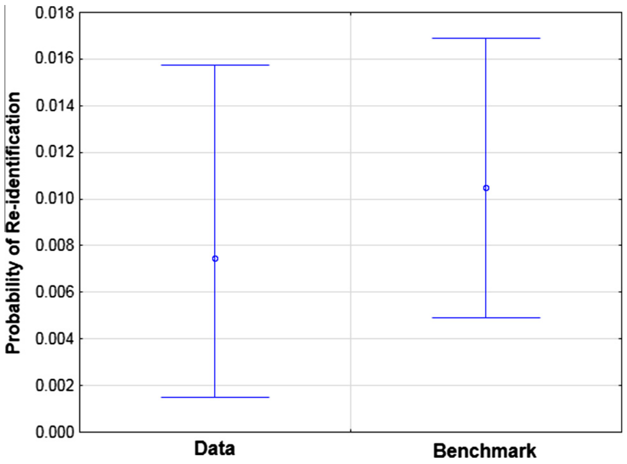

Fig. 2. The 95% confidence intervals for the probability of re-identification for direct identifiers using our evaluation scheme.

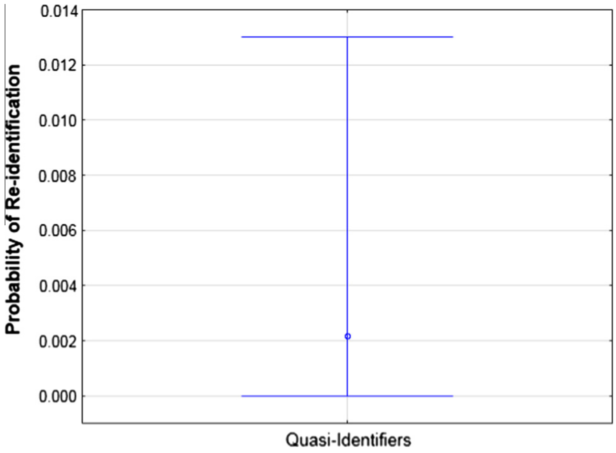

Fig. 3. The 95% confidence intervals for the probability of re-identification for quasi-identifiers using our evaluation scheme.

earlier, this will not always be the case. However, we would expect to see differences in the numerical values and the conclusions about the risk of re-identification.

5. Discussion

5.1. Summary

In this paper we have presented a new framework for evaluating the performance of free-form text de-identification tools that accounts for the many subtleties and distributional variances that one sees in real data sets. It attempts to correct for poorly distributed evaluation corpora, accounts for the data release context, and avoids the often optimistic assumptions about re-identification that are made using the more conventional evaluation approach. This framework provides arguably a more realistic estimate of the true probability of re-identification. The framework was illustrated on a heterogeneous corpus of documents from the University of Michigan medical school.

The application of this framework to the de-identification of clinical reports from clinical trials, as required by the European Medicines Agency, is described further in Appendix B.

5.2. Limitations

Our framework does not consider the precision of the de-identification tool used. Our focus has been on the risk of re-identification only. However, in practice precision would need to be considered as well when evaluating real de-identification systems. That we focused on recall and the risk of re-identification is not intended to diminish the importance of considering precision when evaluating de-identification solutions.

Furthermore, in very rare diseases the risk of re-identification may still be present with a single quasi-identifier. In future work, we will consider the implications of disease frequency in the global population and re-identification risks.

Competing interests

KEE is the founder of and has financial interests in Privacy Analytics, a University of Ottawa and Children's Hospital of Eastern Ontario spin-off company which develops de-identification software for the health sector. MS, GM, and VK are employed and have a financial interest in Privacy Analytics.

Ethics approval

This study received ethics approval from the Children's Hospital of Eastern Ontario REB and the University of Michigan IRB.

Acknowledgments

We wish to thank Stephane Meystre for reviewing an earlier version of this paper and providing us with valuable feedback.

This work was partially funded by a MITACS grant, and a Discovery grant from NSERC.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2016.07.015.

References

[1] S.M. Meystre, F.J. Friedlin, B.R. South, S. Shen, M.H. Samore, Automatic de-identification of textual documents in the electronic health record: a review of recent research, BMC Med. Res. Methodol. 10 (1) (2010) 70.
[2] O. Uzuner, Y. Luo, P. Szolovits, Evaluating the state-of-the-art in automatic de-identification, J. Am. Med. Inform. Assoc. 14 (5) (Oct. 2007) 550–563.
[3] C.J.V. Rijsbergen, Information Retrieval, Butterworths, 1979.
[4] L. Hirschman, J. Aberdeen, Measuring risk and information preservation: toward new metrics for de-identification of clinical texts, in: Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents, 2010, pp. 72–75.
[5] G. Duncan, M. Elliot, G. Salazar, Statistical Confidentiality – Principles and Practice, Springer, 2011.
[6] L. Willenborg, T. de Waal, Statistical Disclosure Control in Practice, Springer-Verlag, New York, 1996.
[7] L. Willenborg, T. de Waal, Elements of statistical disclosure control, Springer-Verlag, New York, 2001.
[8] K. El Emam, Guide to the De-Identification of Personal Health Information, CRC Press (Auerbach), 2013.
[9] K. El Emam, Anonymizing Health Data: Case Studies and Methods to Get You Started, O'Reilly, 2013.
[10] O. Ferrández, B.R. South, S. Shen, F.J. Friedlin, M.H. Samore, S.M. Meystre, BoB, a best-of-breed automated text de-identification system for VHA clinical documents, J. Am. Med. Inform. Assoc. 20 (1) (2013) 77–83.
[11] Alberta Health Services, Likelihood of Identifying Individuals in De-identified Record-Level Data, Public Health Surveillance Bulletin, Bulletin Number 2, June 2013.
[12] Matthijs R. Koot, Guido van 't Noordende, Cees de Laat, A study on the re-identifiability of Dutch citizens, in: PETS Symposium, 2010.
[13] K. El Emam, D. Buckeridge, R. Tamblyn, A. Neisa, E. Jonker, A. Verma, The re-identification risk of Canadians from longitudinal demographics, BMC Med. Inform. Decis. Mak. 11 (1) (2011) 46.
[14] P. Golle, Revisiting the Uniqueness of Simple Demographics in the US Population, in: Proceedings of the 5th ACM Workshop on Privacy in Electronic Society, New York, NY, 2006, pp. 77–80.
[15] L. Sweeney, Uniqueness of Simple Demographics in the U.S. Population, Carnegie Mellon University, Pittsburgh, PA, WP-4, 2000.
[16] D. Carrell, B. Malin, J. Aberdeen, S. Bayer, C. Clark, B. Wellner, L. Hirschman, Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text, J. Am. Med. Inform. Assoc. 20 (2) (2013) 342–348.
[17] S. Meystre, S. Shen, D. Hofmann, A. Gundlapalli, Can physicians recognize their own patients in de-identified notes?, Studies Health Technol Infor. 205 (2014) 778.
[18] C. Marsh, C. Skinner, S. Arber, B. Penhale, S. Openshaw, J. Hobcraft, D. Lievesley, N. Walford, The case for samples of anonymized records from the 1991 census, J. R. Stat. Soc. Ser. A (Stat. Soc.) 154 (2) (1991) 305–340.
[19] Health Information Trust Alliance, HITRUST De-Identification Framework, HITRUST Alliance, 2015.
[20] Institute of Medicine, Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk, Washington, DC, 2015.
[21] PhUSE De-Identification Working Group, De-Identification Standards for CDISC SDTM 3.2, 2015.
[22] The Expert Panel on Timely Access to Health and Social Data for Health Research and Health System Innovation, Accessing Health and Health-Related Data in Canada, Council of Canadian Academies, 2015.
[23] A. Agresti, B. Coull, Approximate is better than 'exact' for interval estimation of binomial proportions, Am. Stat. 52 (2) (1998) 119–126.
[24] M.G. Morgan, M. Henrion, M. Small, Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis, Reprint ed., Cambridge University Press, Cambridge; New York, 1992.

[25] D. Vose, Risk Analysis: A Quantitative Guide, third ed., Wiley, Chichester, England; Hoboken, NJ, 2008.
[26] O. Ferrández, B.R. South, S. Shen, F.J. Friedlin, M.H. Samore, S.M. Meystre, Evaluating current automatic de-identification methods with Veteran's health administration clinical documents, BMC Med. Res. Methodol. 12 (1) (Jul. 2012) 109.
[27] K. El Emam, Risk-based de-identification of health data, IEEE Secur. Priv. 8 (3) (2010) 64–67.
[28] Cancer Care Ontario Data Use and Disclosure Policy, Cancer Care Ontario, 2005.
[29] Security and confidentiality policies and procedures, Health Quality Council, 2004.
[30] Privacy code, Manitoba Center for Health Policy, 2002.
[31] "Privacy code", Health Quality Council, 2004.
[32] Subcommittee on Disclosure Limitation Methodology, Working paper 22: Report on statistical disclosure control, Office of Management and Budget, 1994.
[33] Statistics Canada, Therapeutic Abortion Survey. 2007.
[34] Office of the Information and Privacy Commissioner of British Columbia, Order No. 261-1998, 1998.
[35] Office of the Information and Privacy Commissioner of Ontario, Order P-644, 1994.
[36] L. Alexander, T. Jabine, Access to social security microdata files for research and statistical purposes, Soc. Secur. Bull. 41 (8) (1978) 3–17.
[37] Ministry of Health and Long Term care (Ontario), Corporate Policy 3-1-21, 1984.
[38] Department of Health and Human Services, Standards for privacy of individually identifiable health information, 2000.
[39] Department of Health and Human Services, Standards for privacy of individually identifiable health information, Federal Register, 2000.
[40] Guidelines for the Public Release of Public Health Data, New Hampshire Department of Health and Human Services, September 2008.
[41] Research Data Center, Disclosure Manual – Preventing Disclosure: Rules for Researchers, National Center for Health Statistics, March 2012.
[42] Public Health England, PHE HIV & STI Data Sharing Policy, May 2014.
[43] Data Confidentiality/IRB Policy, Institute for Community Health.
[44] A. de Waal, L. Willenborg, A view on statistical disclosure control for microdata, Survey Methodol. 22 (1) (1996) 95–103.
[45] G. Duncan, T. Jabine, S. de Wolf, Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics, National Academies Press, 1993.
[46] National Center for Education Statistics, NCES Statistical Standards, US Department of Education, 2003.
[47] Office of the Privacy Commissioner of Quebec (CAI), Chenard v. Ministere de l'agriculture, des pecheries et de l'alimentation (141), 1997.
[48] Centers for Disease Control and Prevention, Integrated Guidelines for Developing Epidemiologic Profiles: HIV Prevention and Ryan White CARE Act Community Planning, 2004.
[49] Centers for Medicare and Medicaid Services, BSA Inpatient Claims PUF, 2011.
[50] 2008 Basic Stand Alone Medicare Claims Public Use Files. [Online]. Available: <http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/BSAPUFS/Downloads/2008_BSA_PUF_Disclaimer.pdf>.
[51] E. Erdem, S.I. Prada, Creation of Public Use Files: Lessons Learned from the Comparative Effectiveness Research Public Use Files Data Pilot Project, 13-September-2011. [Online]. Available: <http://mpra.ub.uni-muenchen.de/35478/> (accessed: 09-November-2012).
[52] P. Baier, S. Hinkins, F. Scheuren, The Electronic Health Records Incentive Program Eligible Professionals Public Use File, 26-April-2012.
[53] Instructions for Completing the Limited Data Set ATA use Agreement (DUA) (CMS-R-0235L), Department of Health & Human Services.
[54] K. El Emam, D. Paton, F. Dankar, G. Koru, De-identifying a public use microdata file from the Canadian national discharge abstract database, BMC Med. Inform. Decis. Mak. 11 (53) (2011).
[55] K. El Emam, L. Arbuckle, G. Koru, B. Eze, L. Gaudette, E. Neri, S. Rose, J. Howard, J. Gluck, De-identification methods for open health data: the case of the heritage health prize claims dataset, J. Med. Internet Res. 14 (1) (Feb. 2012) e33.
[56] National Center for Education Statistics, SLDS Technical Brief 3: Statistical Methods for Protecting Personally Identifiable Information in Aggregate Reporting, Institute of Education Sciences, 2010.
[57] Policy for Disclosure of Reportable Disease Information, Iowa Department of Public Health, November 2005.
[58] Health & Social Care Information Centre, The HES Protocol, June 2009.
[59] 2013 Census confidentiality rules and how they are applied, Statistics New Zealand, 2013.
[60] Department of Public Health Confidentiality Procedures, Massachusetts Executive Office of Health and Human Services, October 2012.
[61] MIYHS Data Release Policy, Office of Substance Abuse and Mental Health Services, the Maine Center for Disease Control and Prevention, and the Department of Education, February 2010.
[62] Data Suppression Decision Rules Work Group, Report of Guidelines for Data Result Suppression, Utah Department of Health, Oct. 2009.
[63] Technical Notes: Statistical Methods: Suppression of Rates and Counts, United States Cancer Statistics, September 2014.