Journal of Biomedical Informatics 50 (2014) 142–150

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Text de-identification for privacy protection: A study of its impact on clinical text information content

Stéphane M. Meystre [a,b,*], Óscar Ferrández [c], F. Jeffrey Friedlin [d], Brett R. South [c], Shuying Shen [a,b,e], Matthew H. Samore [a,b,e]

[a] Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, United States
[b] VA Health Care System, Salt Lake City, UT, United States
[c] Nuance Communications Inc., Burlington, MA, United States
[d] Regenstrief Institute, Inc., Indianapolis, IN, United States
[e] Department of Internal Medicine, University of Utah, Salt Lake City, UT, United States

ARTICLE INFO

Article history:
Received 25 August 2013
Accepted 23 January 2014
Available online 3 February 2014

Keywords:
Natural Language Processing
Medical informatics
Confidentiality, patient data privacy
De-identification, Anonymization,
Electronic health records
United States department of veterans affairs

ABSTRACT

As more and more electronic clinical information is becoming easier to access for secondary uses such as clinical research, approaches that enable faster and more collaborative research while protecting patient privacy and confidentiality are becoming more important. Clinical text de-identification offers such advantages but is typically a tedious manual process. Automated Natural Language Processing (NLP) methods can alleviate this process, but their impact on subsequent uses of the automatically de-identified clinical narratives has only barely been investigated.

In the context of a larger project to develop and investigate automated text de-identification for Veterans Health Administration (VHA) clinical notes, we studied the impact of automated text de-identification on clinical information in a stepwise manner. Our approach started with a high-level assessment of clinical notes informativeness and formatting, and ended with a detailed study of the overlap of select clinical information types and Protected Health Information (PHI). To investigate the informativeness (i.e., document type information, select clinical data types, and interpretation or conclusion) of VHA clinical notes, we used five different existing text de-identification systems. The informativeness was only minimally altered by these systems while formatting was only modified by one system. To examine the impact of de-identification on clinical information extraction, we compared counts of SNOMED-CT concepts found by an open source information extraction application in the original (i.e., not de-identified) version of a corpus of VHA clinical notes, and in the same corpus after de-identification. Only about 1.2–3% less SNOMED-CT concepts were found in de-identified versions of our corpus, and many of these concepts were PHI that was erroneously identified as clinical information. To study this impact in more details and assess how generalizable our findings were, we examined the overlap between select clinical information annotated in the 2010 i2b2 NLP challenge corpus and automatic PHI annotations from our best-of-breed VHA clinical text de-identification system (nicknamed 'BoB'). Overall, only 0.81% of the clinical information exactly overlapped with PHI, and 1.78% partly overlapped.

We conclude that automated text de-identification's impact on clinical information is small, but not negligible, and that improved clinical acronyms and eponyms disambiguation could significantly reduce this impact.

1. Introduction

As Electronic Health Records (EHR) are being deployed throughout the U.S. healthcare system, more and more electronic clinical information is becoming easier to access for secondary uses such as clinical research. This evolution offers tremendous potentials, but also equally growing concern for patient confidentiality and privacy breaches. Secondary uses of clinical information for research purposes require patient informed consent, a requirement often difficult to fulfill, especially with research involving larger patient populations. This patient informed consent requirement can be waived if the patient EHR content is de-identified, as

[*] Corresponding author. Address: University of Utah, Department of Biomedical Informatics, 26 S 2000 E, HSEB suite 5700, Salt Lake City, UT 84112, United States. Fax: +1 801 581 4297.

E-mail address: stephane.meystre@hsc.utah.edu (S.M. Meystre).

http://dx.doi.org/10.1016/j.jbi.2014.01.011
1532-0464/© 2014 Elsevier Inc. All rights reserved.

defined in the HIPAA legislation [1]. Two approaches for de-identification are proposed: the "Safe Harbor" method, requiring removal of Protected Health Information (PHI), or the statistical method. Both methods typically involve significant human resources to manually examine EHR content and de-identify it. The former (i.e., "Safe Harbor" method) can also be applied automatically on clinical narrative text, using Natural Language Processing (NLP) methods, and therefore allowing for faster and cheaper de-identification of clinical text [2]. NLP methods have been shown to allow for high accuracy, [3–5] but they could also erroneously categorize clinical information as PHI, or introduce new misleading information when replacing the detected PHI with other information. These issues are also shared with manual de-identification approaches, and could imply reducing the information content of clinical notes, and the accuracy of subsequent automated processes such as information extraction.

The Veterans Healthcare Administration Consortium for Healthcare Informatics Research (CHIR) is a multi-disciplinary group of collaborating investigators affiliated with VHA sites across the U.S. The objectives of the CHIR are to improve the health of veterans through foundational and applied informatics research, advancing the effective use of unstructured text and other types of clinical data in the EHR. Building methods and tools that can be used to automatically de-identify VHA clinical documents is of paramount importance in the development of this initiative. In the context of the CHIR, the de-identification project focused on investigating the current state of the art of automatic clinical text de-identification [2], on developing a best-of-breed de-identification application for VHA clinical documents [3], and on evaluating its impact on subsequent text analysis tasks and the risk for re-identification of this text.

This paper presents our effort to study the impact approaches for preserving patient privacy, specifically automated clinical text de-identification, can have on clinical text informativeness, and on subsequent uses of clinical text such as information extraction.

2. Background

In the United States, current regulations require patient informed consent when using clinical information for research purposes, but this requirement can be waived if the information is de-identified, or if patient consent is not possible (e.g., data mining of retrospective records). For clinical data to be considered de-identified, the "Safe Harbor" method defined in the Health Insurance Portability and Accountability Act (HIPAA; codified as 45 CFR §160 and 164) requires 18 categories of Protected Health Information to be removed [6]. These categories include names, dates (except the year), addresses, telephone and fax numbers, e-mail addresses, social security numbers, other personal identifiers, etc.

Several text de-identification applications have been developed previously, starting with Sweeny's Scrub system [7]. These applications target a variable selection of PHI, ranging from patient names only [8], to all PHI categories defined in the Safe Harbor method, or even everything that was not recognized as clinical information [9]. Most applications focused on only one or two specific clinical document types, such as pathology reports and discharge summaries, and only few systems were evaluated with a more heterogeneous document corpus [7,8,10]. Existing text de-identification applications are mostly based on two different groups of methodologies: pattern matching and machine learning. Many applications combine both approaches for different types of PHI, but the majority uses no machine learning and relies only on pattern matching, rules, and dictionaries. These resources are typically manually crafted, at the cost of months of work by experienced domain experts, and with limited generalizability. An advantage of these methods is that they require little or no annotated training data, and can be easily and quickly modified to improve performance by adding rules, dictionary terms, or regular expressions. Most recent applications tend to be based more on machine learning methods. A large corpus of annotated text is required to train these machine learning algorithms, a resource that also requires significant work by domain experts, even if text annotation is often considered to be easier than knowledge engineering. Annotated corpora can also be shared, such as during the i2b2 de-identification challenge [11]. This challenge allowed for several text de-identification systems development and methods evaluation. A detailed review of earlier research in this domain was published in 2010 [2]. A noteworthy more recent system is the MITRE Identification Scrubber Toolkit (MIST [4]), based on machine learning algorithms and offering a user interface easing the system local adaptation.

We evaluated a selection of these existing systems in the context of our CHIR de-identification project [12], and this study demonstrated an important need for customization to PHI formats specific to VHA documents. It also provided us with detailed insight about the best performing methods and resources for each category of PHI. This knowledge guided our development of a "best-of-breed" (hence the nickname 'BoB') text de-identification system for VHA clinical documents, a system we evaluated with different corpora, and a system that reached excellent performance for VHA clinical documents de-identification [3].

As already mentioned, there is a risk that text de-identification has an adverse effect on subsequent uses of the text like information extraction, but this risk has barely been investigated. To our knowledge, only one published study investigated this risk, and only for medication names [5]. In that study, two different systems were used to automatically de-identify 3503 clinical notes from the Cincinnati Children's Hospital Medical Center: MIST [4], and a locally developed system based on similar methods. An automated information extraction system [13] was used to extract medication names from these notes, before and after de-identification. No significant differences in medication names extraction performance were observed.

The impact of text de-identification on the information content of clinical documents, and on the degree to which the document's key clinical data and the overall meaning and understanding of the document were retained, has not been reported in scientific publications.

3. Methods

Our study of the impact of automatic text de-identification on clinical notes information content was based on a stepwise approach, starting with a high-level analysis of the impact on clinical note interpretability and formatting, and ending with a detailed analysis of the impact on specific clinical information types (Fig. 1). Each step was driven by a research question, and consisted in one of the studies described below.

The experiments presented here were based on two different corpora of clinical notes: the 2010 i2b2 NLP challenge corpus ([14] briefly presented below in Section 3.3.1), and a corpus of VHA clinical notes. The latter was a subset of a reference standard that consisted of 800 manually de-identified clinical documents. These documents were selected using a stratified random sampling approach of the 100 most frequent clinical note types available in a large VHA research database. More details are available in [3].

Each document was annotated by two reviewers, with disagreements adjudicated by a third reviewer. A fourth and final reviewer examined any ambiguous or adjudicated cases the third reviewer marked as needing further clarification. These tasks used annotation guidelines and schemata based on the 18 PHI classes defined

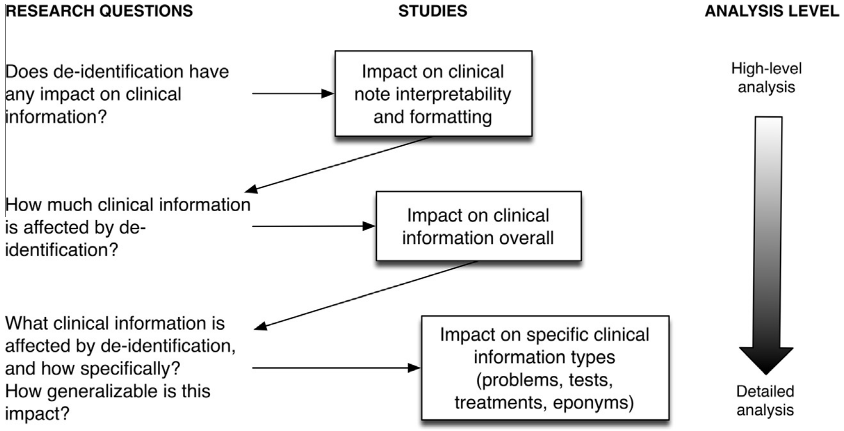

RESEARCH QUESTIONS
STUDIES
ANALYSIS LEVEL
Does de-identification have any impact on clinical information?
Impact on clinical note interpretability and formatting
High-level analysis
How much clinical information is affected by de-identification?
Impact on clinical information overall
What clinical information is affected by de-identification, and how specifically? How generalizable is this impact?
Impact on specific clinical information types (problems, tests, treatments, eponyms)
Detailed analysis

Fig. 1. Stepwise approach research questions and related studies.

in the HIPAA "Safe Harbor" regulation. Going beyond what is defined in this regulation, we adopted a more conservative approach and included annotation classes for organizations (Other Organization Names), mentions of health care facilities (Healthcare Unit Name), information specific to armed forces (Deployment), all states, counties and countries, and all date annotations including the year. It was our interpretation that these types of information could be considered under the 18th PHI category described by HIPAA as "any other unique identifying number, characteristic, or code". Our reference standard was further enriched with annotations of clinical eponyms, (i.e., clinical information bearing person or location names). All annotation tasks were accomplished with Knowtator [15], an open source annotation tool.

3.1. De-identification impact on clinical note interpretability and formatting

After the evaluation of the methods and resources used by a selection of existing text de-identification systems with our VHA corpus [12], we examined a subset of these documents, as automatically de-identified by five selected systems (listed below). We selected a random subset of 50 documents that were manually annotated to identify PHI and clinical eponyms, as explained above. This 50-document set contained 1205 PHI annotations and was also manually annotated for report types, and conclusions or interpretations. Clinical eponyms consist of mentions of diseases, procedures, devices, or anatomy that contain proper names of persons or locations. Examples include Alzheimer's disease, Nissen's fundoplication, Swan-Ganz catheter, or Achilles tendon. Report type is defined as an explicit statement in the report such as "Discharge Summary". We define conclusion or interpretation as a summary or final interpretation of either the findings or the overall impression of the report. These are typically in the form of a single sentence or paragraph. Five freely available text de-identification systems were used in our evaluation: (1) Medical De-identification System (MeDS [16]), (2) HMS Scrubber (HMS [17]), (3) Health Information DE-identification (HIDE [18]{Gardner:2010vj}), (4) MITRE Identification Scrubber Toolkit (MIST [4]), and (5) the MIT System (MIT [19]). Systems 1, 2 and 5 are rule-based systems while the others are mostly machine learning based.

We performed the following evaluations to assess the readability and the degree to which pertinent clinical data was retained in the automatically de-identified reports.

3.1.1. Report interpretability assessment

To indirectly assess informativeness of de-identified documents, we defined the "interpretability score" (IS) for each text de-identification system. The IS consists in a three point scoring system:

IS1: significant clinical data retained (1 = yes, 0.5 = partial, 0 = no).
IS2: type of report retained (1 = yes, 0.5 = partial, 0 = no).
IS3: conclusion or interpretation retained (1 = yes, 0.5 = partial, 0 = no).

To calculate the IS for each system, we manually reviewed all 50 de-identified reports in our sub-corpus, and compared them to their original version (i.e., not de-identified). An IS score was calculated for every system and for every report. For IS1, we gave the system a 1 when all significant clinical data was retained, a 0.5 when only a portion of the significant clinical data was retained, and a 0 when no significant clinical data was retained. We used the same method for IS2 and IS3, and added each document score to obtain the system score. For each component of the IS, a score of 50 was the maximum score, and the maximum total IS score that could be obtained for each system was 150. For this evaluation, all three criteria (IS1, IS2, IS3) are weighted equally and we consider them to be *generally* equivalent in importance. We acknowledge that, depending on what clinical data related to each criteria was deleted, the three criteria may not always be equal (in terms of clinical relevance) in every instance. However in an effort to keep the scoring scheme consistent, easily interpretable, and as free from human judgment as possible, we chose this scoring system and accept some slight degree of over-generalization.

3.1.2. Report formatting alterations assessment

We performed another evaluation where we assessed the degree to which the original report formatting was retained in the de-identified reports. Report formatting consists of syntactic features such as word capitalization, line spacing and indentation, punctuation, paragraph breaks and line numbering. For this assessment, we manually reviewed and compared the original report to the de-identified report for each system and determined whether there were differences in the syntactic features detailed above. This evaluation was qualitative and descriptive only and detailed statistics on formatting alterations were not recorded.

3.2. Impact of de-identification on clinical information overall

To extend our evaluation of the impact of text de-identification on subsequent information extraction from VHA clinical documents, we used an indirect approach to assess how much clinical information was affected. We compared the output of an open source clinical information extraction application (cTAKES [20]) extracting all SNOMED-CT concepts. We used cTAKES to extract these clinical concepts from our VHA corpus de-identified with various methods and approaches to hide PHI.

We started by creating seven different versions of a random subset of 300 documents from our VHA corpus. One version was not de-identified, and the other six versions were de-identified with two different methods for PHI false positives filtering in BoB, our "best-of-breed" text de-identification system (see details in [3]), and with three different approaches to hide PHI: resynthesized PHI (i.e., replaced with realistic surrogates, like replacing "J. Smith" with "P. Herbert"), a general 'PHI' tag for all PHI found in the notes (e.g., replacing "J. Smith" with <PHI>), and tags that also indicate the category of PHI (e.g., replacing "J. Smith" with <PHI-Name>). These corpus versions were as follows:

(1) Original corpus (not de-identified).
(2) De-identified using four binary SVM filters and resynthesized PHI.
(3) De-identified using four binary SVM filters and a general 'PHI' tag.
(4) De-identified using four binary SVM filters and PHI tags that included categories.
(5) De-identified using one multiclass SVM filter and resynthesized PHI.
(6) De-identified using one multiclass SVM filter and a general 'PHI' tag.
(7) De-identified using one multiclass SVM filter and PHI tags that included categories.

We used cTAKES (version 3.0.0, `AggregatePlaintextUMLSProcessor` analysis engine) with its default configuration, limited to the SNOMED-CT dictionary, and ran it with the seven different versions of our corpus. We then performed a pair-wise comparison of our original corpus and each de-identified version of our corpus, counting all SNOMED-CT concepts extracted from each document in the corpus.

We started with a calculation of the proportion of count difference for each concept between the original and each de-identified version of our corpus (count difference divided by count in original version), and averaged it across all concepts. For statistical analysis, we used a null hypothesis that stated there was no concept count proportion difference between the original and a de-identified version of our corpus, a level of significance of 0.05, and two different methods: a paired Student's *t*-test (2-tailed) to compare corpus versions overall, and a log-likelihood ratio comparison method [21] to compare concepts in each corpus version. We used the Bonferroni correction for multiple comparisons.

The log-likelihood ratio method was originally developed to compare word frequencies between different corpora, and we adapted it for concept frequencies. We started by calculating the expected count E_i of each concept in each corpus under a model of homogeneity for frequencies. Comparing each de-identified version of the corpus with the original corpus, we have two expected counts E_1 and E_2 (Eq. (1); a = count of concept in de-identified corpus, b = count of concept in original corpus, c = count of all concepts in de-identified corpus, d = count of all concepts in original corpus).

$$\text{Expected count} \quad E_1 = c(a+b)/(c+d)$$
$$E_2 = d(a+b)/(c+d) \tag{1}$$

The log-likelihood value LL is then calculated for each concept in each de-identified version of our corpus, as specified in Eq. (2) (and ignoring null values).

$$\text{Log-likelihood value} \quad LL = 2[(a \cdot \ln(a/E_1)) + (b \cdot \ln(b/E_2))] \tag{2}$$

3.3. Impact of de-identification on specific clinical information types

To assess how generalizable our findings with our VHA corpus were, and examine the impact of de-identification in more details, we used the 2010 i2b2 NLP challenge corpus as well as our VHA corpus.

3.3.1. Impact on problems, tests, and treatments

The 2010 i2b2 challenge corpus is available upon a Data Use Agreement for research purposes, which makes our methodology reproducible for other researchers. The i2b2 2010 NLP challenge was focused on extracting medical problems, tests, and treatments, as well as assessing their local context (e.g., "...*denied* chest pain"), and extracting specific relations between these concepts [14]. This makes the i2b2 2010 NLP challenge corpus an ideal reference standard for our study

In this study, the first step consisted in the de-identification of the documents. We used 'BoB,' [3] our VHA clinical text de-identification application, to automatically annotate all PHI that could appear in the i2b2 corpus documents. BoB also annotates clinical eponyms, which are not PHI but could easily be confused with sensitive PHI classifiers such as person names.

Once we had the documents de-identified, we analyzed the overlap of our automatic PHI annotations with the 2010 i2b2 NLP challenge reference standard problem, test, and treatment annotations. High overlap rates would jeopardize subsequent uses of the documents in other tasks such as clinical information extraction.

Evaluating the performance of BoB in the de-identification task is out of the scope of this paper. Moreover, the 2010 i2b2 NLP challenge corpus does not contain PHI annotations, which would make this evaluation difficult. For details about BoB's performance on the de-identification task, please see [3].

3.3.2. Impact on clinical eponyms

We used our VHA corpus to perform a second evaluation where we assessed the number of instances when the de-identification systems we evaluated (mentioned in Section 3.1) recognized clinical eponyms as PHI. To perform this evaluation, we first identified all human annotations of clinical eponyms in four categories: (1) anatomy, (2) devices, (3) diseases, and (4) procedures. We then identified the number of times each system identified an annotation of one of these categories as PHI.

4. Results

4.1. De-identification impact on clinical note interpretability and formatting

4.1.1. Report interpretability assessment

As reflected in Table 1, not all 50 training documents had explicitly stated report types. Approximately 30% of the training documents did not explicitly state a report type before de-identification. Where no report type was explicitly stated, each system was given an IS2 score of one. Of all reports processed by the 5 systems, only 3 de-identified reports (2 generated by HMS Scrubber and 1 by MeDS) had the report type removed. Also, not all reports

Table 1
Interpretability calculations for each system.

System	Interpretability score (maximum score = 50 for each category)			
	IS1 (clinical data)	IS2 (report type)	IS3 (conclusion)	Total score (percent of total)
MIT	48	50	50	148 (99%)
MIST	48	50	50	148 (99%)
HIDE	46	50	50	146 (97%)
HMS	43	48	50	141 (94%)
MeDS	39	49	50	138 (92%)

had what could reasonably be defined as a conclusion. For example, reports such as a "Treatment Plan Weekly Update" or an "Informed Consent Form" did not have interpretations or conclusions as defined in Section 3.1. Approximately 40% of the reports were missing interpretations or conclusions before de-identification. None of the de-identification systems removed any portion of a report's conclusion or interpretation.

4.1.2. Report formatting alterations assessment

When comparing the syntactic features of the original report to the de-identified reports, we found that most systems retained nearly all original report formatting in the de-identified reports. The MeDS system removed some report formatting (some line breaks) in approximately 5% of de-identified reports. In all reports processed by the HMS Scrubber system, all line spacing, paragraph breaks and indentation was removed.

Fig. 2 shows a side-by-side comparison of an original report with formatting (on the left) and a de-identified report (on the right) by the HMS Scrubber system with most syntactic features removed.

4.2. Impact of de-identification on clinical information overall

When used with the seven versions of our corpus, cTAKES found an average of 136976.6 (0.95 CI: 136892.4–137060.7) SNOMED-CT concepts in the different versions of our corpus, reaching an average of 456.6 (0.95 CI: 456.31–456.87) concepts per document. These concepts correspond to 8794 distinct SNOMED-CT concepts. The most frequent concept was found 1117 times in the not de-identified corpus, and 926 different concepts were found only in the various de-identified versions of our corpus. On average, distinct concepts were found 15.71 times in the corpus. The distribution of concept counts is depicted in Fig. 3.

The probability associated with the *t*-test allowed us to reject our null hypothesis when comparing most de-identified versions of the corpus with the original corpus (Table 2). Only versions de-identified using one multiclass SVM filter and resynthesized PHI or tags indicating the PHI category (C5 and C7 in Table 2) were not significantly different from the original corpus.

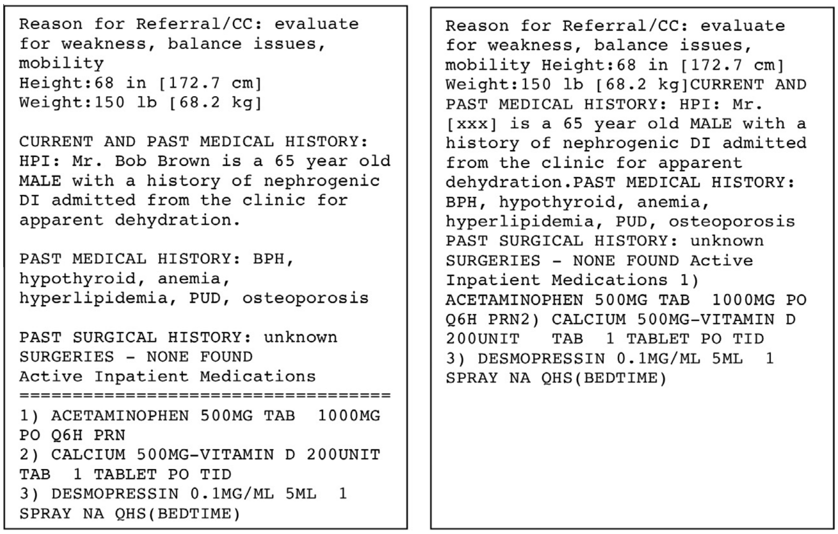
To determine the statistical significance of the log-likelihood value, we referred it to the chi-squared distribution (with one degree of freedom). All log-likelihood values above 6.96 (would have been 3.84 without Bonferroni correction) could then allow rejecting our null hypothesis. As seen in Table 2, none of the de-identified version of our corpus had a significant average log-likelihood value.

Even if no corpus version was different overall, a few concepts had an average log-likelihood value above 6.96, with counts that were therefore significantly different between the original and a de-identified version of our corpus (Table 3). The de-identification process caused a few false negatives (i.e., concepts found in the original version of the corpus, but not in a de-identified version), and sometimes false positives (i.e., concepts found in a de-identified version of the corpus, but not in the original version).

4.3. Impact of de-identification on specific clinical information types

4.3.1. Impact on problems, tests, and treatments

The 2010 i2b2 NLP challenge corpus includes a total of 47,685 annotations: 19,667 medical problems, 13,833 tests, and 14,185 treatment terms. As shown in Table 4, 386 PHI annotations exactly overlapped with problem, test, or treatment annotations, which means that overall, 0.81% of the relevant clinical data was mistakenly de-identified. When considering partial overlaps, this percentage slightly increased to 1.78%. In both cases, exact and partial overlaps, treatment was the category most affected by



```
Reason for Referral/CC: evaluate
for weakness, balance issues,
mobility
Height:68 in [172.7 cm]
Weight:150 lb [68.2 kg]

CURRENT AND PAST MEDICAL HISTORY:
HPI: Mr. Bob Brown is a 65 year old
MALE with a history of nephrogenic
DI admitted from the clinic for
apparent dehydration.

PAST MEDICAL HISTORY: BPH,
hypothyroid, anemia,
hyperlipidemia, PUD, osteoporosis

PAST SURGICAL HISTORY: unknown
SURGERIES - NONE FOUND
Active Inpatient Medications
===================================
1) ACETAMINOPHEN 500MG TAB  1000MG
PO Q6H PRN
2) CALCIUM 500MG-VITAMIN D 200UNIT
TAB  1 TABLET PO TID
3) DESMOPRESSIN 0.1MG/ML 5ML  1
SPRAY NA QHS(BEDTIME)
```

```
Reason for Referral/CC: evaluate
for weakness, balance issues,
mobility Height:68 in [172.7 cm]
Weight:150 lb [68.2 kg]CURRENT AND
PAST MEDICAL HISTORY: HPI: Mr.
[xxx] is a 65 year old MALE with a
history of nephrogenic DI admitted
from the clinic for apparent
dehydration.PAST MEDICAL HISTORY:
BPH, hypothyroid, anemia,
hyperlipidemia, PUD, osteoporosis
PAST SURGICAL HISTORY: unknown
SURGERIES - NONE FOUND Active
Inpatient Medications 1)
ACETAMINOPHEN 500MG TAB  1000MG PO
Q6H PRN2) CALCIUM 500MG-VITAMIN D
200UNIT   TAB  1 TABLET PO TID
3) DESMOPRESSIN 0.1MG/ML 5ML  1
SPRAY NA QHS(BEDTIME)
```

Fig. 2. Side-by-side comparison of an original report (left) and a de-identified report (right) with most syntactic features removed (all PHI in the original report is fictitious).

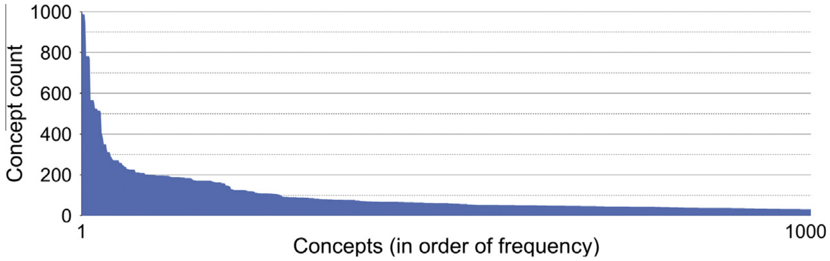


Fig. 3. SNOMED-CT concepts distribution in the original corpus.

Table 2
SNOMED-CT concept count differences between corpora.

	C1	C2	C3	C4	C5	C6	C7
All concepts count	138,137	136,990	136,473	136,574	137,482	135,883	137,297
All concepts count difference with original corpus (C1)		−0.83%	−1.20%	−1.13%	−0.47%	−1.63%	−0.61%
Average concept count difference with original corpus (C1)		−2.28%	−2.75%	−3.04%	−1.18%	−1.87%	−1.59%
Concept count difference 95% confidence int.		−1.20% – −3.36%	−1.70% – −3.81%	−1.98% – −4.11%	0% – −2.28%	−0.94% – −2.79%	−0.51% – −2.66%
Probability associated with *t*-test		0.0014*	<0.001*	0.002*	0.068	<0.001*	0.022
Average log-likelihood value		0.275	0.277	0.322	0.275	0.211	0.261

C1 = Original corpus (not de-identified); C2 = De-identified using four binary SVM filters and resynthesized PHI; C3 = De-identified using four binary SVM filters and a general 'PHI' tag; C4 = De-identified using four binary SVM filters and tags that included PHI categories; C5 = De-identified using one multiclass SVM filter and resynthesized PHI; C6 = De-identified using one multiclass SVM filter and a general 'PHI' tag; C7 = De-identified using one multiclass SVM filter and tags that included PHI categories.

* Statistically significant pair-wise comparison with the original corpus (C1).

de-identification errors; 1.98% of treatment annotations were completely de-identified, while 3.4% were partially de-identified.

Note that in order to reach the aforementioned counts, our de-identification system automatically correctly reclassified 112 clinical eponyms, which helped exclude common ambiguous clinical terms from the PHI category such as 'Parkinson', 'Pfannenstiel', 'Holter', 'Foley', 'Whipple', or 'Roux'.

Table 5 illustrates the overlap rates of problems, tests, and treatments, with regard to each PHI category detected by our de-identification system. As expected, the majority of overlaps correspond to ambiguities with the *Person Name* PHI category. Specifically, 88% of the total count of exact overlaps relate to this category, and 76% when considering partial overlaps. Other PHI categories such as *Healthcare Unit* and *State/Country* also included some overlap with clinical data, although less frequently.

4.3.2. Impact on clinical eponyms

Human annotators found a total of 65 clinical eponym annotations in the 50 documents VHA sub-corpus. Table 6 shows the number of eponyms each system identified as PHI. We found that most systems misclassified approximately 10% of the eponyms (such as devices, procedures and diseases) as PHI while two systems (MeDS and HMS Scrubber) misclassified eponyms to a much larger degree.

5. Discussion

Our stepwise study of the impact of de-identification on clinical information showed that the overall impact on the information content of clinical documents was minimal, but not negligible.

When looking at the interpretability and formatting of clinical notes, most key clinical data and the overall meaning and understanding of the document were retained. MeDS generated the highest number of reports (11 of 50) where a portion of the clinical data was removed from the report, but in all of these cases only a very small proportion of the entire clinical data was removed and the overall interpretation or meaning of the report was retained. Clinical narrative reports often contain syntactic features such as line spacing, paragraph breaks and section headers in order to make the documents more human readable. Very few systems removed or modified the syntax or formatting of reports. One system (HMS Scrubber) however, did remove all line spacing, paragraph breaks and indentation from the report, essentially formatting the report as one long, unending paragraph as shown in Fig. 2. Subjectively, the de-identified report with most syntactic features removed in Fig. 2 is much more difficult for the human reader to interpret and understand. Clearly, line spacing, indentations and location of empty lines appeared to play a very important role in human readability of a medical report. We found that when these formatting features were removed and the report was displayed as one long, unending paragraph, it became much more difficult for the human to quickly scan the report and identify the type of report and the conclusion for example. Since many NLP systems frequently make use of syntactic features when processing a report, there is the potential that the accuracy of automated concept extraction would suffer when features such as line spacing and indentations are removed during de-identification. Interestingly, it was the opinion of the reviewer performing this evaluation that removal of report formatting played a greater role in decreasing understanding and comprehension of a report than did the occasional removal of isolated medical data, which as stated below, could often be inferred correctly.

Our assessment of the impact of de-identification on clinical information overall in VHA clinical notes also demonstrated only a small impact (loss of about 1.2–3% of SNOMED-CT concepts), although we found significant differences in counts of concepts between the original version of our corpus (i.e., not de-identified) and most de-identified versions of the corpus.

When examining each concept separately, we realized that 34 of them had a log-likelihood value above 6.96. This significant count difference between the original and the de-identified version of our corpus was often due to PHI in the original version being erroneously recognized as a SNOMED-CT concept, and sometimes parts of the PHI tag or resynthesized PHI in the de-identified

Table 3
SNOMED-CT concepts with significant count differences (*).

SNOMED concepts (code)	Matched terms (correct meaning)	Concept count in each corpus version							LL value average (and extremes)
		C1	C2	C3	C4	C5	C6	C7	
Vertebral artery (85234005)	VA (Veterans Admin.)	112	36*	0*	0*	4*	0*	0*	82.91 (40.32–125.50)
Carney complex (239132009)	Carney Hospital	53	64	61	382*	61	59	63	48.02 (0.43–284.40)
Phencyclidine measurement (20857001)	PCP (Primary Care Provider)	47	8*	6*	6*	7*	7*	6*	33.80 (30.30–35.79)
Pneumonia, pneumocystis carinii (415125002)	PCP (Primary Care Provider)	42	6*	5*	4*	7*	8*	6*	30.27 (24.79–36.16)
Diabetic retinopathy (4855003)	DR (Doctor)	51	24*	15*	17*	26*	28	26*	11.69 (6.42–20.31)
Endoplasmic reticulum (33761008)	ER (Emergency Room)	20	4*	4*	4*	5*	4*	5*	10.82 (9.54–11.51)
Mixed antiglobulin reaction test (117360002)	MAR	3	15*	0	0	12*	0	0	7.33 (5.83–8.83)

version being erroneously recognized as a SNOMED-CT concept. In the former case, examples were 'VA' (i.e., "Veterans Administration") recognized as "vertebral artery," 'PCP' (i.e., "Primary Care Provider") recognized as "pneumocystis carinii pneumonia," 'DR' (i.e., 'Doctor') recognized as "diabetic retinopathy," or 'ER' (i.e., "Emergency Room") recognized as "endoplasmic reticulum."

The versions of our corpus de-identified with BoB using one multiclass SVM to filter PHI candidates had the lowest impact on clinical information, especially when replacing PHI with resynthesized realistic surrogates. In this case, about 1.18% less SNOMED-CT concepts were found after de-identification. Since an average of 456.6 concepts were found in each document, this means that an average of about 5 concepts were "lost" in each document because of the de-identification process. Knowing that almost all concepts that were significantly less (or more) frequent after de-identification were generated because of cTAKES errors (erroneously disambiguated acronyms), this impact would eventually be even more limited.

When studying the details and generalizability of our assessment of the impact of de-identification on clinical information, the overlap between PHI and select clinical information was also very limited. Our analysis revealed that many de-identification errors were found in the "medications" section of the i2b2 corpus documents, where a list of medications taken by the patient is mentioned. The lack of context in these sections made our de-identification system mark some of the medications as PHI. Examples of treatment wrongly annotated as PHI were 'Colace', 'Lopressor', and 'Senna.' Regarding problem and test annotations, the overlap was much less frequent although the system still misinterpreted some annotations as PHI, such as 'E. Coli', 'Fournier', 'Addison' as problems, and 'Apgars', and 'Papanicolaou' as tests. The format of these annotations, such as *Initial. LastName* for 'E. Coli', as well as common person names and last names that are genuine clinical data, were the main causes of errors. These eponyms were also problematic in our VHA corpus. All systems identified some clinical eponyms such as anatomic locations, devices, diseases and procedures, as PHI. Two systems (MeDS and HMS Scrubber) identified a significantly larger number of eponyms as PHI. However, in reports where medical data was identified as PHI and removed, it was never to the degree where the overall meaning, interpretation and readability of the report were compromised. In fact, we found that it was frequently possible for medical experts to correctly infer what medical data was removed based on the context of the report and the words surrounding it. For example a trained medical expert would likely be able to infer that in the phrase "the patient was having trouble urinating, so we inserted a <PHI> catheter" the PHI removed was in fact "Foley," a type of medical catheter.

To sum up, our results point out that more accurate de-identification techniques for ambiguous medical terms such as the ones mentioned above are needed. This is still an unsolved challenge affecting the entire NLP research community, and it involves investigating and solving semantic variability and ambiguity. Nevertheless, the low overall rate of relevant clinical data that was damaged by the de-identification process leads us to think that, although not negligible, the impact of de-identification would not be very significant.

This study demonstrates that even an efficient text de-identification system like BoB can cause clinical information to be mistakenly considered as PHI and hidden or removed. This overlap is small, but not negligible. Another recent detailed study focused on the impact of text de-identification on the subsequent automatic extraction of medication names, and found no significant impact [5], but medications represent only a small part of the clinical information found in clinical notes, and a minority of the overlapping information we analyzed. This outcome is probably due to the

Table 4
Problem, test, and treatment annotations overlap with PHI detected by our de-identification system.

i2b2 concepts	Total	Detected as PHI (exact match)	%	Detected as PHI (partial match)	%
Problem	19,667	65	0.33	187	0.95
Test	13,833	40	0.29	180	1.30
Treatment	14,185	281	1.98	482	3.40
Overall	47,685	386	0.81	849	1.78

Table 5
Problem, test, and treatment annotations overlap rates for each PHI type.

	Exact match				Partial match			
PHI type	Problem	Test	Treat.	Overall	Problem	Test	Treat.	Overall
Person name	0.31%	0.25%	1.73%	0.72%	0.82%	0.74%	2.70%	1.36%
Street/City	0	0	0.01%	0	0.01%	0.01%	0.02%	0.01%
State/Country	0.01%	0.01%	0.01%	0.01%	0.06%	0.09%	0.13%	0.09%
Deployment	0	0	0	0	0	0	0	0
ZIP code	0	0	0	0	0	0	0	0
Healthcare Unit	0	0	0.19%	0.06%	0	0.12%	0.37%	0.15%
Other Org Name	0	0	0.01%	0	0	0.07%	0.11%	0.05%
Date	0	0	0	0	0.02%	0.14%	0.01%	0.05%
Age > 89	0	0	0	0	0	0	0	0
Phone Number	0	0	0	0	0	0	0	0
Electronic Address	0	0	0	0	0	0	0	0
SSN	0	0	0	0	0	0	0	0
Other ID Number	0.01%	0.03%	0.03%	0.02%	0.04%	0.13%	0.06%	0.07%

Table 6
Number of PHI annotations identified by each system as clinical eponyms.

System	Eponyms classified as PHI	
	Count	% Of all eponyms
MIT	8	12.31
MIST	7	10.77
HIDE	9	13.85
HMS	26	40.00
MeDS	32	49.23

limited “resemblance” of medication names with PHI. On the other hand, several clinical information categories resemble PHI, and clinical eponyms are a good example of such resemblance. In a previous study of clinical narratives at the VHA, we found that across all annotated documents, eponyms represented 3.5% of all annotations, while proper names of persons represented only 13% of all annotations [22]. Eponyms for procedures were most prevalent (45%), devices and diseases were less common (18% and 30%), anatomical structures were least common. Average Inter-Annotator Agreement (between human annotators) was only moderate for annotation of eponyms (74%), but high for proper names of persons (93%). Therefore, the risk to wrongly annotate eponyms as some other PHI type is high. Automated de-identification systems should not only ensure that patient names are reliably detected, but also that eponyms are retained in de-identified clinical documents to the extent that is possible. Based on the various observations discussed above, we plan to focus our future research efforts on the automatic disambiguation of clinical eponyms and abbreviations, and on PHI detection accuracy improvements in general.

Acknowledgments

We thank Abhisek Trivedi and Matthew Maw for their help with these studies.

Research supported by VA HSR HIR 08-374. Views expressed are those of authors and not necessarily those of the Department of VA or affiliated institutions.

References

[1] CFR Title 45 Subtitle A Part 46: Protection of Human Subjects [Internet]. GPO; October 1, 2008. <http://www.access.gpo.gov/nara/cfr/waisidx_08/45cfr46_08.html>.

[2] Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Med Res Methodol 2010;10:70.

[3] Ferrandez O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. BoB, a best-of-breed automated text de-identification system for VHA clinical documents. J Am Med Inform Assoc 2013;20(1):77–83.

[4] Aberdeen J, Bayer S, Yeniterzi R, Wellner B, Clark C, Hanauer D, et al. The MITRE identification Scrubber Toolkit: design, training, and assessment. Int J Med Inform 2010;79(12):849–59.

[5] Deleger L, Molnar K, Savova G, Xia F, Lingren T, Li Q, et al. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. J Am Med Inform Assoc 2012.

[6] CFR Title 45 Subtitle A Part 164: Security and Privacy [Internet]. GPO; October 1, 2008. <http://www.access.gpo.gov/nara/cfr/waisidx_08/45cfr164_08.html>.

[7] Sweeney L. Replacing personally-identifying information in medical records, the Scrub system. Proc AMIA Annu Fall Symp 1996:333–7.

[8] Taira RK, Bui AA, Kangarloo H. Identification of patient name references within medical documents using semantic selectional restrictions. AMIA Annu Symp Proc 2002:757–61.

[9] Berman JJ. Concept-match medical data scrubbing. How pathology text can be used in research. Arch Pathol Lab Med 2003;127(6):680–6.

[10] Ruch P, Baud RH, Rassinoux AM, Bouillon P, Robert G. Medical document anonymization with a semantic lexicon. AMIA Annu Symp Proc 2000:729–33.

[11] Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc 2007;14(5):550–63.

[12] Ferrandez O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. Evaluating current automatic de-identification methods with Veteran’s health administration clinical documents. BMC Med Res Methodol 2012;12(1):109.

[13] Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc 2010;17(5):514–8.

[14] Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 2011;18(5):552–6.

[15] Knowtator: a protégé plug-in for annotated corpus construction. Association for Computational Linguistics; 2006. <http://dl.acm.org/citation.cfm?id=1225791>.

[16] Friedlin FJ, McDonald CJ. A software tool for removing patient identifying information from clinical documents. J Am Med Inform Assoc 2008;15(5):601–10.

[17] Beckwith BA, Mahaadevan R, Balis UJ, Kuo F. Development and evaluation of an open source software tool for deidentification of pathology reports. BMC Med Inform Decis Mak 2006;6:12.

[18] HIDE: An integrated system for health information DE-identification. CBMS 2008 June 17; 254–9.

[19] Neamatullah I, Douglass MM, Lehman LW, Reisner A, Villarroel M, Long WJ, et al. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak 2008;8(1):32.

[20] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507–13. 2010 ed., September–October.

[21] Rayson P, Garside R. Comparing corpora using frequency profiling. association for computational linguistics. In: Proceedings of the workshop on comparing corpora. 2000;p. 1–6.

[22] South B, Shen S, Maw M, Ferrandez O, Friedlin FJ, Meystre S. Prevalence estimates of clinical eponyms in de-identified clinical documents. AMIA Summits Transl Sci Proc, CRI; 2012. p. 136.