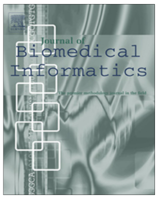
Journal of Biomedical Informatics xxx (2014) xxx–xxx

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 xxxx

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 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 defined in the HIPAA legislation [1]. Two approaches for

* 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 Published by Elsevier Inc.

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