EVALUATION OF PHI HUNTER IN NATURAL LANGUAGE PROCESSING RESEARCH

Posted on January 4, 2015 by Administrator

Category: HIM Operations
Tags: human subjects, informatics, protected health information, research ethics

by Andrew Redd, PhD; Steve Pickard, MBA; Stephane Meystre, MD, PhD; Jeffrey Scehnet, PhD; Dan Bolton, MS; Julia Heavirland, MA; Allison Lynn Weaver, MPH, LADC; Carol Hope, PharmD, MS; and Jennifer Hornung Garvin, PhD, MBA, RHIA, CTR, CPHQ, CCS, FAHIMA

Abstract

Objectives: We introduce and evaluate a new, easily accessible tool using a common statistical analysis and business analytics software suite, SAS, which can be programmed to remove specific protected health information (PHI) from a text document. Removal of PHI is important because the quantity of text documents used for research with natural language processing (NLP) is increasing. When using existing data for research, an investigator must remove all PHI not needed for the research to comply with human subjects' right to privacy. This process is similar, but not identical, to de-identification of a given set of documents.

Materials and methods: PHI Hunter removes PHI from free-form text. It is a set of rules to identify and remove patterns in text. PHI Hunter was applied to 473 Department of Veterans Affairs (VA) text documents randomly drawn from a research corpus stored as unstructured text in VA files.

Results: PHI Hunter performed well with PHI in the form of identification numbers such as Social Security numbers, phone numbers, and medical record numbers. The most commonly missed PHI items were names and locations. Incorrect removal of information occurred with text that looked like identification numbers.

Discussion: PHI Hunter fills a niche role that is related to but not equal to the role of de-identification tools. It gives research staff a tool to reasonably increase patient privacy. It performs well for highly sensitive PHI categories that are rarely used in research, but still shows possible areas for improvement. More development for patterns of text and linked demographic tables from electronic health records (EHRs) would improve the program so that more precise identifiable information can be removed.

Conclusions: PHI Hunter is an accessible tool that can flexibly remove PHI not needed for research. If it can be tailored to the specific data set via linked demographic tables, its performance will improve in each new document set.

Keywords: research ethics, human subjects, protected health information, informatics

Introduction

Unstructured clinical text from the medical record can be a key source of clinical information for healthcare research and quality improvement. In the Department of Veterans Affairs (VA), documents from the electronic health record (EHR) are stored as free text, and the unstructured

clinical information contained within them is not readily accessible for decision support, quality assessment, performance monitoring, and clinical research. The VA's Office of Research and Development (ORD) Health Services Research and Development Service (HSRD) launched major research-related initiatives to address this informatics gap in 2008. The Consortium for Healthcare Informatics Research (CHIR)[1] is a major VA research initiative that was formed to create new tools and methods for natural language processing (NLP). VA Informatics and Computing Infrastructure (VINCI)[2] facilitates the use of data in a secure environment and serves as both a software development environment and a secure location to store and analyze data. This study used VINCI and was conducted within a research project that resulted from the development of tools related to CHIR.

In this study, we introduce and evaluate a software program called PHI Hunter. PHI Hunter was developed for use within VA HSRD research projects to remove only the protected health information (PHI) not relevant to the given research study, thus leaving intact all text necessary to fulfill the objectives of the research. PHI Hunter is a program developed in a common statistical software package, SAS, by one of the authors (S.P.) and was designed to identify and remove PHI from free-form text. It consists of a set of rules or algorithms to identify patterns in text called regular expressions (regexs). Regular expressions are a family of programming languages or dialects that can be used for identifying patterns in text, such as repeating numbers, prefixes and dates. The term *regular expression* typically refers to a specific instance or pattern. To operationalize regular expressions, we needed a specific implementation engine to define the syntax, interpret the defined regex patterns, and perform the search and removal of PHI. We used the Perl compatible regular expressions (PCRE) implementation, a free and open-source implementation commonly used in statistical and business analytics software including SAS/STAT,[3] R (a free, open-source implementation of the S language used in statistics and data analysis),[4] and Perl (a scripting programming language that is well suited for text processing).[5] *PCRE* refers both to the specific dialect of regular expressions and to the open-source engine that processes the patterns and text of the same name. The choice of PCRE maximized the potential for this tool to be reused and generalized by ourselves and others.

PHI Hunter was applied to a corpus of clinical documents in the VA text integration utility (TIU) files, stored in Veterans Health Information Systems and Technology Architecture (VistA), the VA's health information system and EHR system. These documents represent unstructured text. As part of our text research portfolio, we sought to selectively remove true patient identifiers while attempting to minimize removal of PHI needed for the research. We evaluated how well PHI Hunter met this goal. We found that PHI Hunter had good performance for removal of most PHI, and we explored the limits of PHI requirements for research.

Background

The Veteran's Health Administration (VHA)[6] establishes procedures for the use of data for VHA research purposes. The definition of PHI is based on both the Health Insurance Portability and Accountability Act (HIPAA) of 1996[7] and the regulations governing Institutional Review Boards (IRBs)[8] known as "the Common Rule." The HIPAA Privacy Rule's "Safe Harbor" method delineates 18 identifiers as PHI,[9] only some of which were needed to undertake our research (see Table 1). De-identification is the removal of all PHI. De-identification is usually performed manually, making it a costly and time-consuming endeavor. NLP can be used to automatically detect PHI and transform it in clinical documents; this process has been developed and evaluated by several teams described in a review by Meystre and colleagues.[10] Noteworthy recent systems that were not described in the aforementioned review are the MITRE Identification Scrubber Toolkit (MIST),[11] developed by the MITRE Corporation, and a text de-identification system, developed by VA researchers, that is considered to be best-of-breed in text de-identification and is nicknamed BoB.[12] While both of these systems either detect and remove or transform all categories of PHI found in clinical narratives, using sophisticated methods based on machine learning algorithms, dictionaries, pattern matching, and rules, utilization of the systems requires knowledge of these methods as well as access to these customized systems expressly installed and built for that purpose. In this study, we did not seek to develop or use a full de-identification system, but rather to develop a system that is easily implemented and can be used to increase the safeguards that are already in place to protect patient data in text used in human subject research. While de-identification systems are an area of active research, the literature on the specific case of removing specific identifiers based on investigative needs to increase patient privacy in research is limited.

Even though our research is conducted within VINCI, a secure research environment, we wanted to use only the identifiable patient data we needed for our research because the ethical use of patient information for research requires that researchers use only the data that is needed for research.[13] SAS software, a readily accessible tool, was used to create a program using PCRE to remove specific occurrences of PHI that were not needed in our use of clinical text notes. Algorithms, step-by-step procedures for using PCRE with SAS software, were developed to scan documents for PHI and to remove and replace it with marker text denoting that an element had been removed.

Materials and Methods

We used PHI Hunter to identify and remove PHI because the objective of our research was to use only the text necessary to conduct our study. Some categories of PHI were needed to conduct the research and so were not removed from our data (see Table 1). For example, our study required

dates. However other categories of information were removed, including patient, provider, and other names; Social Security numbers; phone numbers; addresses; and other identifiers such as insurance numbers or license plate numbers. To remove this information, PCRE expressions were created to search for and replace the identifiable information. On the basis of a sample of notes, we used our best estimate (heuristic determination) of a set of regexs to identify all unique text patterns of data within each category for removal. A variable was also created for each category to sum the number of occurrences in each document where data were found and replaced with the SAS algorithms.

We piloted PHI Hunter using SAS 9.2 because SAS versions 9.0 and above support PCRE. We chose SAS for this type of process is because SAS is commonly used in many medical research and informatics institutions, allowing transferability of the algorithms among the healthcare community with little modification. The regular expressions also could be used with other statistical analysis programs such as R with little modification. The exact regular expressions used are available in Appendix A, and the specific SAS implementation is provided in Appendix B.

Our system is flexible. It is possible to turn off the categories that are not needed. This program was used to remove up to 6 of the 18 defined PHI elements (see Table 1). For example, as mentioned, service dates were important for research purposes in this study, so this category of PHI removal was not used. However, any of the 18 defined PHI elements could be removed. Operationally, this program works by pulling in the text files, processing the text to remove the PHI, and then outputting a cleaned version of the file in which PHI is replaced by the text "*** PHI Removed ***" (see Figure 1). The files can be in any format that SAS can read, such as comma-delimited files, text files, Microsoft Excel spreadsheets, Microsoft Access databases, and SAS data sets.

To evaluate the performance of PHI Hunter, 500 documents were processed to remove PHI. The data for this evaluation were collected in two stages. Stage 1 identified the missed PHI, or false negatives. Stage 2 identified instances that were falsely identified as PHI, or false positives. In the first stage, the TIU documents were run through the PHI Hunter system to remove all PHI. Next, experienced document reviewers, called annotators, reviewed the documents for manual identification of PHI that was missed by the system. The annotators identified the missed information and classified the instances by type of PHI. The annotations were compared and differences were reviewed by an adjudicator, who resolved differences between the independent reviews of the two annotators. The results were recorded in a spreadsheet.

The goal of the second phase of annotation was to identify instances where words or numbers were falsely identified as PHI, or false positives. Each piece of text that was identified by PHI Hunter was examined in context and determined to be either correctly or incorrectly identified as PHI. The

Windiff.exe utility,[14] a tool that graphically compares the contents of two text files or the contents of two folders that contain text files to verify whether they are the same, was used to identify the differences between the original version and the PHI Hunter version of the files. Two members of the study team reviewed the highlighted differences, recorded in a spreadsheet what PHI Hunter

had removed, and validated whether the information was PHI or not. By comparing the original to the PHI removed, a human annotator was able to classify the removed text with regard to whether the PHI was correctly removed and what kind of PHI it was to derive descriptive statistics as well as to determine false positives.

Through this evaluation, we estimated the performance of PHI Hunter and identified areas of deficiency. The metrics we used were sensitivity (an indicator of a system's ability to correctly identify cases of interest, statistically defined as the probability that the system will select a true case as a predicted case) and positive predictive value (a measure of a system's reliability, or the statistical probability that a predicted case from the system is actually a case). We would ordinarily also calculate specificity, but because of the nature of regexs, no clear unit of analysis is available, and therefore specificity is undefined. We analyzed deficiency through a detailed failure analysis to (1) identify classes of PHI that are most commonly missed, (2) identify the forms of text that are most commonly incorrectly identified as PHI, (3) identify areas where text was incorrectly identified as PHI but could not reasonably be excluded by the algorithm, and (4) identify PHI that was missed but could not reasonably be removed by algorithm.

Results

The results show that in the areas where regular expressions are expected to perform well, the system did indeed perform well. The Social Security number is perhaps the most structured of the information, and indeed the system performed perfectly in extracting Social Security numbers from our corpus. Table 2 shows the performance of the system by PHI type as determined by a human reviewer. Certain instances of data such as dates and times did not qualify as PHI, as reflected in the Table 2. Other categories of PHI where the system performed well were medical record numbers and phone numbers.

The areas where PHI Hunter had the most trouble were names and locations, many of which were often names as well. These areas had low sensitivity. Of the names that were found, there was a high positive predictive value. The false positives involving names were most often due to triggering characters; for example, a phrase such as "need to ask MD" might flag the word "ask" as a name of a provider. The system did not catch 2,245 instances of PHI, from a total of 6,698 instances, for a system overall sensitivity of 66 percent; however, as shown in Table 2, the system performed well and had a wide range of concept-specific sensitivities.

Overall, the system removed 5,475 words or phrases, 4,453 of which were PHI, for a positive predictive value of 81.3 percent (95 percent confidence interval, 80.3 percent–82.4 percent). The false positives were examined to review the reasons for failure of accurate detection. Major causes were abbreviations that included numbers, such as chemical abbreviations, and proximity to triggering phrases such as "MD". Other causes include eponyms such as "Von Willebrand's disease" and "Simpson's test". Similarly, we found that patterns of characters appeared to the program as

identifiers.

Table 3 shows the number of false positives that were triggered by each of the patterns in the system. The vast majority of these are due to an ID pattern. This resulted in several false positives that were similar to ID numbers but were in fact chemical compositions such as “O2” and “CO2.”

False negatives were also examined. The breakdown by percentage from each category of PHI is listed in Table 4. Location false negatives could largely be split into two categories, those with VA location information and those with nonspecific locations smaller than a hospital. VA locations listed the city of a VA clinic with or without indicating it was a clinic. For example; “Salt Lake City”, “Salt Lake City VA”, “SLC”, and “SLC VA” were all present. Nonspecific locations were locations within a hospital that technically meet the criteria for PHI; however, they are nonspecific in that they are locations that many hospitals clinics have. For example, “primary care”, “ER”, “ED”, “emergency room”, and “icu” were all common.

Dates were reexamined and categorized on the basis of the pattern exhibited. It appears that the system can be improved by including previously unrecognized patterns of date information. Table 5 shows the breakdown of these patterns. The table is ordered by a recommendation of inclusion. The “Count” column shows the number of PHI elements that could be captured by including just that pattern in the algorithm. The “New Count” column shows how many elements would be caught that were not caught in any of the patterns above it. The “Cumulative Count” column shows the total number of elements that would be caught by including that pattern and all above it. No attempt was made to discern how many false positives would be generated from these patterns; however, the last pattern, “MMM”, is very short and in our opinion has a high likelihood of false positives. It is included in the table for summarization of the false negatives.

Names do not conform to simple patterns for identification and so are difficult to extract through automated regular expression–based systems. Of the phone number false negatives, 10 consisted of four- or five-digit extensions and 2 were full phone numbers.

Discussion

For the task of removing PHI not needed for research to aid in patient protection, the PHI Hunter system performed well. It was proficient at removing some of the most sensitive information and information that is hardly ever needed for research, particularly Social Security numbers, medical record numbers, other IDs, and phone numbers. Categories in which it performed poorly, such as dates, locations, and names, are often needed for research. (Note that our research remains in the secure VINCI environment to give the best protection to the patients whose records are used in research.)

We discovered in development and evaluation that the patterns may or may not be exclusive to a single category of PHI, which could complicate the issue of overremoval or overscrubbing when

only specific details are intended to be removed. As an example, in the VA, Social Security and medical record numbers look remarkably similar because they are both 9-digit numbers, sometimes including separators and sometimes not. In our research we wished to remove the Social Security number but not the medical record number. When adapting this system to a particular application, care will need to be taken not to remove pieces of information that are not intended to be removed, especially when this information is critical to further tasks.

This project also highlights the importance of context around the elements that are desired to be removed. The inability to separate Social Security numbers from medical record numbers shows that context would be needed to correctly classify each if it were important to remove one but leave the other. Dates show this scenario in the extreme. Most dates in the medical record are PHI, but other dates are often needed for research, and without a contextual analysis these would be impossible to differentiate. Birthdates appear to a computer much the same as clinical visit, procedure, or other dates, and context would be needed for determining if the information was relevant to the research or not.

One possible place for improvement through context is demographic information. If demographic tables that can be linked to the EHR system, then the program can be more precise in the identifiable information that can be removed. In the program, the EHR can be linked to the demographic information. Once linked, patterns could be generated around the demographic information such as patient name, address, and facility for each record. For example, facilities are usually consistent with how a doctor's name might appear in the record. Key phrases such as "ATTENDING PHYSICIAN: " or "COSIGNER: " allow a regular expression to be built around "ATTENDING PHYSICIAN:" and "COSIGNER:" that will remove and replace what comes after the colon. This functionality, however, is beyond the scope of our research, in which we sought a simple solution for removing critical PHI not needed in the research.

The easiest area of improvement would be to improve and expand the patterns that are looked for in text. Dates included many more variations than were originally caught in the development stage. These variations do, however, show a likelihood for false positives or overscrubbing. As discussed by Meystre et al.,[15] any automatic text de-identification could cause overscrubbing. This issue is often related to clinical eponyms, diagnoses, procedures, devices, or anatomical locations that bear personal names, such as "von Willebrand's disease" or "Simpson's test," as mentioned previously. PHI Hunter suffered from this issue, which could be alleviated with additional disambiguation methods, such as those implemented in BoB.[16] Overscrubbing could also alter the clinical document formatting and its readability, but Meystre and colleagues[17] report that this risk is low.

Another issue related to de-identification is the risk of patient "re-identification." The social and clinical information has rich content that could allow linking with other databases of such information that also include patient identification. Since PHI Hunter does not seek full de-identification, this risk

has not been evaluated, but we acknowledge that in this study, in which only selected PHI categories were targeted, this risk could be higher, which is why other safeguards are in place.

As mentioned previously, PHI Hunter does not seek to be a full de-identification system, yet it can perform well for some highly sensitive PHI categories, such as Social Security numbers. It falls short in other categories when compared to systems such as BoB, which has reached a sensitivity of 98.5 percent for patient names. PHI Hunter fills a role of a lightweight solution that increases patient privacy when records are used for research, but does not include the overhead and maintenance costs that come with a full de-identification system or the time requirements of manual de-identification.

Conclusions

This tool is designed to be used in a secure research environment with research staff trained in human subject rights. PHI Hunter is an accessible tool that can flexibly remove specific PHI categories not needed for a research project to improve patient privacy. The performance of PHI Hunter can be improved via linked demographic tables in a given document set. With the increasing use of large document data sets for research, PHI Hunter provides a readily available mechanism to remove PHI.

Acknowledgments

The research reported here was supported by the Department of Veterans Affairs, Veterans Health Administration, Office of Research and Development, Health Services Research and Development Service IBE 09-069. Dr. Redd is a statistician at the VA Salt Lake City Health Care System IDEAS Center. The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the United States government.

Andrew Redd, PhD, is an Assistant Professor at the University of Utah and a statistician in the IDEAS Center at the VA Salt Lake City Health Care System in Salt Lake City, UT.

Steve Pickard, MBA, is a data analyst in the IDEAS Center at the VA Salt Lake City Health Care System in Salt Lake City, UT.

Stephane Meystre, MD, PhD, is an Assistant Professor at the University of Utah and a Research Investigator in the IDEAS Center at the VA Salt Lake City Health Care System in Salt Lake City, UT.

Jeffrey Scehnet, PhD, is a VINCI Staff member at the VA Salt Lake City Health Care System in Salt Lake City, UT.

Dan Bolton, MS, is a statistician in the IDEAS Center at the VA Salt Lake City Health Care System in

Salt Lake City, UT.

Julia Heavirland, MA, is a research coordinator in the IDEAS Center at the VA Salt Lake City Health Care System in Salt Lake City, UT.

Allison Lynn Weaver, MPH, LADC, is an analyst at the Centers for Medicare and Medicaid Baltimore, MD.

Carol Hope, PharmD, MS, is a postdoctoral fellow in the IDEAS Center at the VA Salt Lake City Health Care System in Salt Lake City, UT.

Jennifer Hornung Garvin, PhD, MBA, RHIA, CTR, CPHQ, CCS, FAHIMA, is an Associate Professor at the University of Utah and a core research investigator in the IDEAS Center at the VA Salt Lake City Health Care System in Salt Lake City, UT.

Notes

1. 1. Veterans Health Administration. *Program Announcement for Request for Concept Paper for Service Directed Research: Consortium for Healthcare Informatics Research (CHIR)*. US Department of Veterans Affairs, 2011. Available at http://www.research.va.gov/funding/solicitations/docs/Consortium-Healthcare-Informatics.pdf.
2. US Department of Veterans Affairs. "VA Informatics and Computing Infrastructure (VINCI)." Available at http://www.hsrd.research.va.gov/for_researchers/vinci/.
3. Cody, R. "An Introduction to Perl Regular Expressions in SAS 9." *Proceedings of the 29th Annual SAS Users Group International* (2004).
4. "The R Project for Statistical Computing." Available at http://www.R-project.org/.
5. "The Perl Programming Language." Available at http://www.perl.org/.
6. Veterans Health Administration. *Use of Data and Data Repositories in VHA Research* (VHA Handbook 1200.12). Department of Veterans Affairs, 2009. Available at http://www1.va.gov/vhapublications/ViewPublication.asp?pub_ID=1851.
7. "Health Insurance Portability and Accountability Act of 1996." Public Law 104-191. August 21, 1996. Available at http://aspe.hhs.gov/admnsimp/.
8. Veterans Health Administration. *Use of Data and Data Repositories in VHA Research* (VHA Handbook 1200.12).
9. US Government Printing Office (GPO), Department of Health and Human Services. 45 CFR 164.514, "Other Requirements Relating to Uses and Disclosures of Protected Health Information," 2002.
10. Meystre, S. M., F. J. Friedlin, B. R. South, S. Shen, and M. H. Samore. "Automatic De-identification

of Textual Documents in the Electronic Health Record: A Review of Recent Research." *BMC Medical Research Methodology* 10 (2010): 70.
11. Aberdeen, J. S. Bayer, R. Yeniterzi, B. Wellner, C. Clark, D. Hanauer, B. Malin, and L. Hirschman. "The MITRE Identification Scrubber Toolkit: Design, Training, and Assessment." *International Journal of Medical Informatics* 79, no. 12 (2010): 849–59.
12. Ferrandez, O., B. R. South, S. Shen, F. J. Friedlin, M. H. Samore, and S. M. Meystre. "BoB, a Best-of-Breed Automated Text De-identification System for VHA Clinical Documents." *Journal of the American Medical Informatics Association* 20, no. 1 (2013): 77–83.
13. US Department of Veterans Affairs. "VA Informatics and Computing Infrastructure (VINCI)."
14. Microsoft Support. "How to Use the Windiff.exe Utility." 2012. Available at support.microsoft.com/kb/159214.
15. Meystre, S. M., F. J. Friedlin, B. R. South, S. Shen, and M. H. Samore. "Automatic De-identification of Textual Documents in the Electronic Health Record: A Review of Recent Research."
16. Ferrandez, O., B. R. South, S. Shen, F. J. Friedlin, M. H. Samore, and S. M. Meystre. "BoB, a Best-of-Breed Automated Text De-identification System for VHA Clinical Documents."
17. Meystre, S. M., F. J. Friedlin, B. R. South, S. Shen, and M. H. Samore. "Automatic De-identification of Textual Documents in the Electronic Health Record: A Review of Recent Research."

Printer friendly version of this article.

Andrew Redd, PhD; Steve Pickard, MBA; Stephane Meystre, MD, PhD; Jeffrey Scehnet, PhD; Dan Bolton, MS; Julia Heavirland, MA; Allison Lynn Weaver, MPH, LADC; Carol Hope, PharmD, MS; and Jennifer Hornung Garvin, PhD, MBA, RHIA, CTR, CPHQ, CCS, FAHIMA. "Evaluation of PHI Hunter in Natural Language Processing Research." *Perspectives in Health Information Management* (Winter 2015): 1-15.

There are no comments yet.