A Successful Technique for Removing Names in Pathology Reports Using an Augmented Search and Replace Method

Sean M. Thomas, M.D., Burke Mamlin, M.D., Gunther Schadow, M.D., Ph.D., and Clement McDonald, M.D.
Regenstrief Institute for Health Care, Indianapolis, Indiana

Abstract

The ability to access large amounts of de-identified clinical data would facilitate epidemiologic and retrospective research. Previously described de-identification methods require knowledge of natural language processing or have not been made available to the public. We take advantage of the fact that the vast majority of proper names in pathology reports occur in pairs. In rare cases where one proper name is by itself, it is preceded or followed by an affix that identifies it as a proper name (Mrs., Dr., PhD). We created a tool based on this observation using substitution methods that was easy to implement and was largely based on publicly available data sources. We compiled a Clinical and Common Usage Word (CCUW) list as well as a fairly comprehensive proper name list. Despite the large overlap between these two lists, we were able to refine our methods to achieve accuracy similar to previous attempts at de-identification. Our method found 98.7% of 231 proper names in the narrative sections of pathology reports. Three single proper names were missed out of 1001 pathology reports (0.3%, no first name/last name pairs). It is unlikely that identification could be implied from this information. We will continue to refine our methods, specifically working to improve the quality of our CCUW and proper name lists to obtain higher levels of accuracy.

Introduction

The potential benefit of sharing clinical information across institutions is very large. Access to such data sources would facilitate epidemiologic and other research and would provide larger patient samples than could be obtained at a single institution. These data would be most accessible for users if it could be de-identified.

The Health Insurance Portability and Accountability Act (HIPAA)[1] provides a list of 19 data elements that should be removed to de-identify data. A number of attempts have been made to remove these identifiers both from structured databases [2] and from free textual reports [3,4]. These methods have achieved success rates greater than 90% in removing identifying information, but the details of the algorithms and methods used are not publicly available[3] or require specific knowledge about natural language processing [4]. We wished to determine if the de-identification accuracy of a substitution method that could be easily implemented and reused could meet these levels of de-identification. The system would not require specialized natural language processing and could be built largely from publicly available data sources. We concentrated on the removal of proper names including patient and physician names as well as names of institutions for the purposes of this study, as these are by far the most prevalent identifier in our pathology reports.

Methods

The basis of our strategy was to create a list of proper names that we want to remove from, and a list of Clinical and Common Usage Words (CCUW) that we want to retain in, the target text. The challenge to our method is that a large number of words are ambiguous, or can be used as either a proper name or a CCUW. We know that colors can be proper first names (Violet) and/or last names (White, Green). Many other words that could be used in a clinical report are also proper names (Hood, Mark, Billing). Clinical terms that are proper names (Barrett's esophagus, Schilling's test) are so common that we have a name for them, eponyms. Finally, many of the proper names we found were peculiar, including "Vessel, Cancer, Tissue, Block, And, The". All of these appeared as proper names in the Social Security Death Index. These ambiguous words are a special problem requiring removal when they are used as proper names, but not when they are used as CCUW. We developed simple rules to make this distinction.

A foundation of our method is the assumption that proper names almost always occur in pairs in clinical reports. If a single proper name is used, there will be surrounding cues identifying it as a proper name (Mrs., Dr., Ph.D.) If we can correctly identify one proper name in a pair, or a prefix or suffix that suggests a name, then we can remove the target word even if it is not in our proper name list. For example, if "Mary Snow" is a proper name in our report, the program will find and remove the name Mary since it is in the proper name list. It will then evaluate Snow, and if Snow is not in the CCUW list, it will be removed because it is preceded by a name. However, if Snow is in the CCUW list, it will remain. We want any of the ambiguous words to be excluded from

either list so the program will these syntactic clues in the analysis.

The master CCUW list was built from the Unified Medical Language System (UMLS) word index of words with a Source Restriction Level of 0 or 1 (no restriction or restriction on translating to foreign languages) and the word list from the GNU spellchecking program Ispell. Proper names are capitalized in the Ispell dictionary, making them easy to remove. After combining these two lists and removing duplicates, the list contained 320,000 words.

The master proper name list was composed from three sources. We took (1) proper names from the Ispell dictionary, (2) all patient and physician names from the Regenstrief Medical Record System (RMRS), and (3) all names from the Social Security Death Index (SSDI, 65 million records). We included first and last names as separate proper names. Multipart names were split into individual components (Van Eyck became two names). We combined all these names and removed duplicates. The list of proper names was extremely large and included 1.8 million proper names.

The overlap between the CCUW list and the proper name list included more than 37,000 words. This is depicted in Figure 1. In our first attempts to de-identify documents, we removed any word from the target text that was in the proper name list. Because of the large overlap between the proper name list and the CCUW list, this strategy removed 40% of the words in the target text, rendering it virtually unreadable.

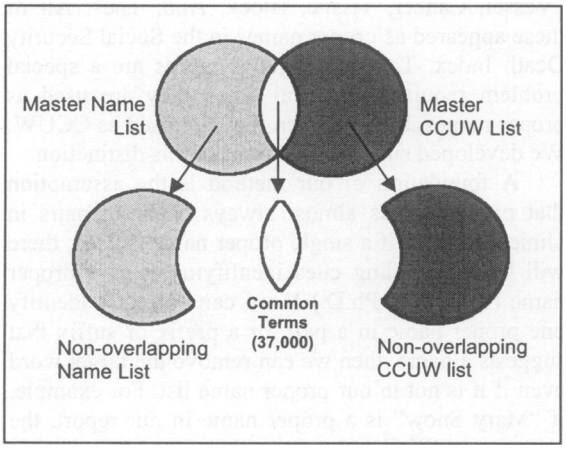

Figure 1. Intersection of the CCUW and Proper name lists.

We refined our proper name list by removing CCUW that were not proper names or ambiguous words. The strategy used to create our final lists is outlined in Figure 2. We started this process by taking the 1219 most frequently occurring words in 200,000 pathology reports. These words represented 90% of the word mass of the reports. Words from this list that were not in the master proper name list were added to the non-overlapping CCUW list (from Figure 1). There were 580 words in this list. The other 639 words were hand reviewed and divided into three categories. Words that were obviously CCUW were added to the CCUW list (such as abdominal, cancer, and vessel). Words that were obviously proper names were added to the proper name list (such as Aaron, Jones and Wagner). Ambiguous words were not included in either list (such as brown and green).

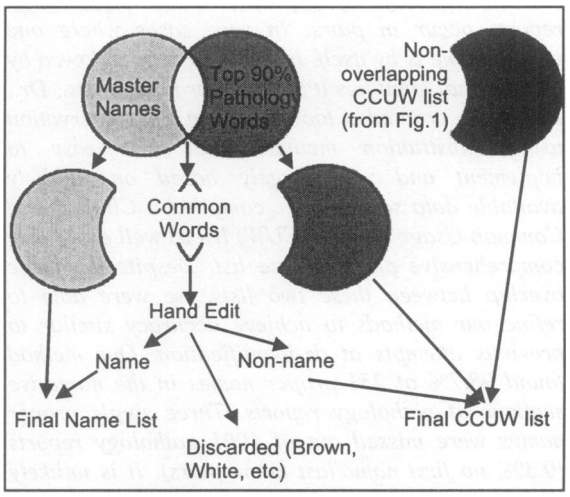

Figure 2. Make-up of the final proper name and CCUW lists.

Our approach protected spurious removal of names from the proper name list. At the same time, we were able to remove CCUW from the proper name list that were very common in pathology reports (especially words such as the, and, vessel and cancer). This greatly improved the readability of the final reports and maintained high accuracy in de-identification.

Search and Replace Algorithm

We initially transform each report into an XML document. Specimen IDs are removed and the header section (which contains a large number of proper names, such as surgeon, pathologist, etc.) is marked up with XML tags separately from report body (description and diagnosis). Our de-identifying software ignores all XML tags.

Our algorithm is depicted in Figure 3. A report is read into memory, and tokenized into words. If the word in question is in the proper name list, it is tagged for removal. Otherwise, if the word is in the CCUW list, the surrounding words are checked for

prefixes and suffixes that suggest a proper name such as Dr., Mr., or MD.

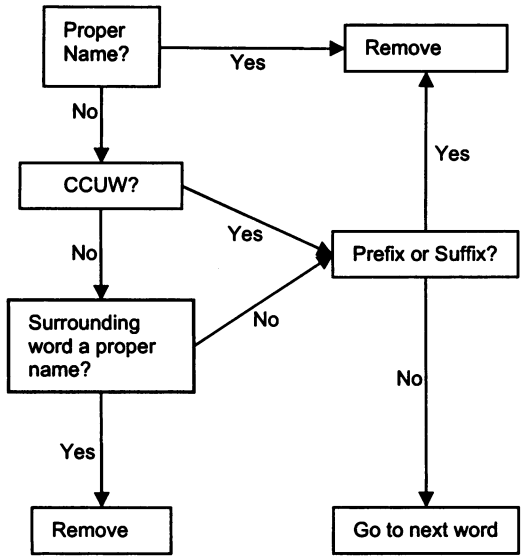

Figure 3. The de-identification algorithm.

If the word is not in either the proper name list or the CCUW list, then we check the surrounding words to check for signs of a proper name. First, the preceding word is checked to see if it is marked as a proper name. The following word is check to see if it is in the proper name list. If either of these conditions is true, the word is marked as a proper name. This allows us to mark combinations where the word in question is not in either list, but is part of a proper name. Finally, the word is checked for a prefix or suffix that suggests a proper name, as described above.

Once every word in the report has been checked, we use simple pattern matching to change dates, specimen numbers, telephone numbers and email addresses. These identifiers were not included in this study.

Evaluation

We randomly selected 143 reports from all pathology reports in the RMRS for Wishard Hospital for each of the seven years from 1995-2001 for a total of 1001 reports. These reports were different than the reports used to create the "Top 90%" list. Two of the authors (SMT, BM) marked all names with an XML tag that identified the word as a proper name.

The de-identifying software marked words that it found in the proper name list as "found", and words that were marked because of surrounding words or affixes as "guessed". We then used an XSL style sheet to count the resulting tagged reports. Our rules for interpretation of the tags are depicted in Table 1.

		Human	
		Tagged	Not Tagged
Program	Found	Correctly Identified Proper Name	CCUW Incorrectly Identified as a Proper Name
	Not Found	Proper Name Not Identified	CCUW Correctly Ignored

Table 1. Interpretation of tags.

For example, a word that was tagged as a proper name and also "found" was considered correctly identified, as were words that were tagged as a proper name but also "guessed" by the program to be a proper name. Using the XSL stylesheet, we were able to view the reports in a Web browser (Figure 4) and easily locate words that were correctly identified or missed.

In the example shown in Figure 4, the last name Short❶ was missed by the algorithm. This is because it was not in the proper name list, but was in the CCUW list. The names John, Ross and Simpson ❷ were in the proper name list and were correctly identified as names. The names Francis and Margo❸ were correctly guessed to be names because they occurred next to names. The word tan❹ was incorrectly identified as a proper name because it is in the proper name list. Finally, the word white❺ was incorrectly identified as a name because it was next to a word that was tagged as a proper name, tan.

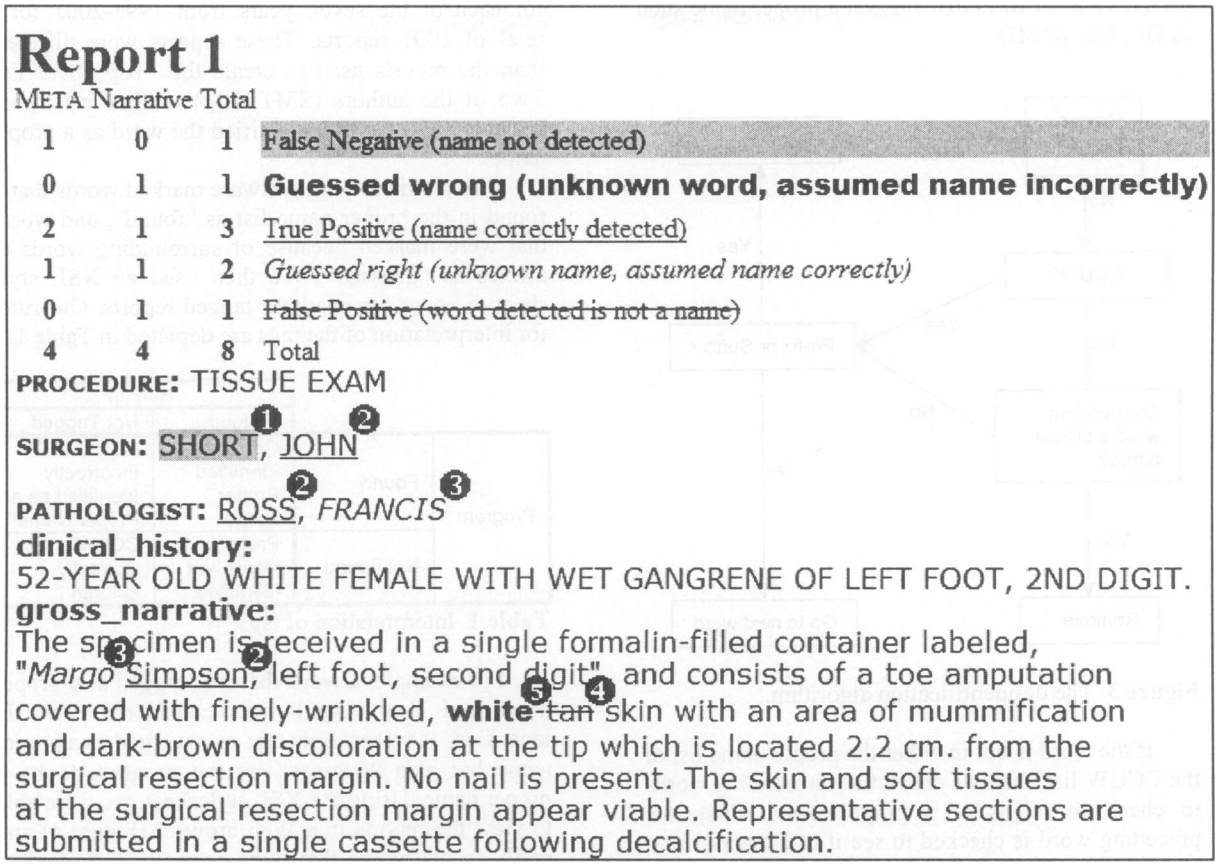

Figure 4. A screenshot of the output of the XSL stylesheet used for reviewing the performance of the software. Names are fictitious. The top portion shows the number of errors by error type and report section.

Results

The results are summarized in Table 2. The sample set consisted of 1001 pathology (surgical pathology and cytology reports). There were a total of 108,092 words that were checked throughout all reports. The authors tagged a total of 7710 proper names in the reports. Of these, 231 were in the narrative sections. The program correctly identified 228 (98.7%) of the proper names in the narrative section, and 7151 (92.7%) of the total proper names. The software incorrectly marked 2063 words as proper names, or 1.9% of the total words.

The three proper names that were missed in the narrative section included one first name and two last names, all of physicians. In the first case, the first name was a commonly used preposition, and the middle and last name were correctly marked for removal. It is unlikely that the physician would be identified based on the missed name. In the other two cases, the last names were missed based on an error during the hand-editing of the intersection set as described in Figure 2. Both names were inadvertently left out of the proper name list. When the error was corrected, the program correctly identified both names. The first names in both cases were correctly identified despite the error in editing, however.

	Narrative Section	Entire Report
Proper Names Correctly Identified	228 (98.7%)	7151 (92.7%)
Proper Names Not Identified	3 (1.3%)	559 (7.3%)
Word Incorrectly Identified as Proper Name	1711	2063 (1.9%)
Total Number of Words Checked		108,092

Table 2. Results by report section.

Discussion

This is a successful approach to the de-identification of free-text pathology reports using substitution methods. By adding the analysis of syntactic clues we improved the success rate of previous search and replace strategies (30-60%) [3], to

correctly identify 98.7% of proper names in the narrative sections of pathology reports. The program missed only three proper names out of 1001 reports bodies (0.3%). Additionally, they were single proper names, never a first/last name pair. It is unlikely that a patient could be identified from these missed proper names.

The strength of our algorithm is the handling of words that do not appear in the CCUW or name list. We use syntactic clues to determine if these ambiguous terms are proper names. In this way, we were able to correctly identify Jack Brown as two proper names, but not incorrectly identify "green-brown tissue" as a proper name. Additionally, we correctly identified common words (words in the CCUW list) that were actually names, such as Dr. Hood or Joe Billing, MD. The characteristics of the reports at our institution make this reliable. It is unknown if this strategy would work with reports from other institutions.

One of the major obstacles to the success of this method was the heterogeneity of the lists. The master name list contains the words "abdominal, absolutely, cancer, vessel, and, the and tissue", while the word list contains "Aaron, Abbey" and other proper names. Clearly, many words are ambiguous and can be either CCUW or proper names.

We concentrated our analysis on the report bodies due to the high yield of useful clinical data. Header information can be stripped off in a preprocessing step with minimal loss to the clinical data integrity. We kept the headers intact for our analysis to increase the number of "targets" for our software. The names that were not identified in the report body by our program were single names (never a first name, last name pair), and unlikely to lead to identification of a patient or sample.

Another disadvantage of our method is the large number of false positives. The reports remained readable by one of the authors despite this, although there is the potential to remove critical words. We favored over-scrubbing over increased specificity to ensure that all names were removed. We plan to pursue methods to increase the performance of the software.

The accuracy of our software can be increased in at least two ways. The brute-force method would be to go through the list of words common to both the name and non-name lists (37,000 words). This is time-consuming, and introduces operator error and bias. A more simple approach is to use our marked up sample set, putting names and non-names into their respective lists. This would require some hand editing so that ambiguous words would not be included in either list.

We are interested in disseminating this tool to other institutions to see what results they achieve. We will also make our CCUW list and proper name list available to the research community. We will investigate other methods of refining our base word lists as these are critical to the success of our algorithm. Additionally, we will add logic to remove other identifying elements (phone numbers, addresses) that while rare, may occur in our reports.

This work was performed at the Regenstrief Institute for Health Care in Indianapolis, Indiana and was supported in part by grants from the National Library of Medicine (T15 LM-7117-05) and National Cancer Institute (1 U01 CA91343-01).

References

1. Department of Health and Human Services OotS. The Health Insurance Portability and Accountability Act of 1996, Standards for Privacy of Individually Identifiable Health Information; Final Rule. Federal Register 65 FR 82462 . 12-28-2000.
2. Sweeney L. Guaranteeing anonymity when sharing medical data, the Datafly System. Proc AMIA Annu Fall Symp 1997;51-55.
3. Sweeney L. Replacing personally-identifying information in medical records, the Scrub system. Proc AMIA Annu Fall Symp 1996;333-337.
4. Ruch P, Baud RH, Rassinoux AM, Bouillon P, Robert G. Medical document anonymization with a semantic lexicon. Proc AMIA Symp 2000;729-733.