Efficient De-Identification of Electronic Patient Records for User Cognitive Testing

Kenric W. Hammond  
Department of Veterans Affairs  
kenric.hammond@va.gov

Efthimis N. Efthimiadis*  
University of Washington  
rjlaundr@uw.edu

Abstract

Cognitive testing of Electronic Health Record (EHR) document presentation and practice will be increasingly valuable as institutions adopt the EHR and begin to experience information overload. Valid testing of clinical decision making requires realistic documents to validly portray clinical scenarios in context. The authors describe a system adapted from public source software that efficiently assists de-identification at considerable savings of time and with good accuracy.

1. Introduction

Electronic health record (EHR) usage will soon become the norm in US healthcare. One inevitable consequence of the ability to access large volumes of textual records will be an urgent need for tools to manage textual information overload. Text is a vital complement to structured and encoded information. Natural language is “natural”: easily learned, flexible and able to express nuance. As such, it is often the most appropriate, easiest and cheapest way to store complex patient care information. When text capability is added to an EHR system, it quickly dominates the data space accessed by most EHR users. Text might not exceed stored images in sheer data volume, but is by far the most accessible form of information in the EHR.

Twenty-five years of EHR usage in the U.S. Department of Veterans Affairs (VA) health care system has resulted in a formidable volume of stored text. For VA clinicians it’s an everyday occurrence to meet a new patient for the first time and discover that the electronic chart contains one or two thousand on-line notes. Given that typical caregivers have only a few minutes to review charts before starting to make clinical decisions, we have made studying how to support efficient access to text notes our research priority. In previous research on the clinician’s perspective on working with text in the EHR, we learned of the cognitive challenge posed by large volumes of text. EHR users told us of their desire for quicker access to important information in text records and their need to separate relevant from irrelevant information [1, 2].

Given the motivation to introduce EHR text, rather little planning went into developing tools to manage text once it became broadly established. There are some possible explanations for this: paper records, for better or worse, were the default model for the first implementations of text in the EHR. In paper charts, notes, written in longhand and viewed by members of the care team with chart access served as mnemonics for the unfolding progress of care, and culminated if the patient was hospitalized, in a dictated, typed discharge summary. The main use of notes, once the summary was complete was archival, permitting reconstruction of a case for a lawsuit or a quality review, but contributing relatively little to supporting contemporaneous decision making. Paper-based notes from office practices were notoriously untidy. In the present era of managed care shared among many providers in a health care organization, legible and clear records are necessary. When an EHR is present, legibility is no longer a barrier, and access to narrative records is no longer limited to physical proximity to the chart. Changes in health care delivery organization and increased accessibility of the EHR have elevated the prominence and broadened the “readership” of the EHR “progress note” in a host of important ways, including:

- Increased prospective surveillance of care.
- Increased reliance on documentation as the basis for reimbursement.
- Reliance on documentation for attestation of adherence to required procedures in care
- Reliance on documentation to furnish proof of regulatory compliance.
- Increased interest in mining text to yield data for research.

*deceased
Our research also identified that in addition to providing a record of the evolving course of patient care, a textual information system can also serve as a dynamic communication system to coordinate care, a means of exerting administrative control over care provider behavior, a medical education resource, and a resource for communicating with the patient. Freely available and abundant patient text has disrupted the paradigm of the paper chart. The result is a relatively undifferentiated and unplanned resource, and only recently have questions emerged about how to put it in some order.

Experimental study may assist understanding opportunities and methods to manage text. Our approach has been to apply a cognitive work analysis perspective informed by users’ reports. Recently we have evaluated the impact of using statistical relevance filtering of displayed textual documents. Initial results demonstrated that users found documents scoring high in relevance as having higher quality and utility. We conducted this research using actual, de-identified patient records.

The next step in evaluating this will be to directly assess whether clinical decision making can be enhanced by relevance filtering. Proper study requires many actual clinical records for hypothesis testing. As EHR development continues, more and more research questions will shift from how to implement the EHR to how best to implement it. The design of user interfaces and information retrieval strategies will become more important. Remarkably little empirical data exists. One of the main reasons for this is that implementing an EHR is very costly in organizational resources. Once members of the health care organization adopt a system, it’s hard to change the system, and considerations of the cost of changing impede progress. Accordingly, testing interfaces “in vivo” is quite difficult, and comparing interfaces with outcomes from one organization’s EHR to another’s is methodologically difficult. Preferable would be developing a way to test user interfaces experimentally yet realistically.

One approach is use of simulation. In a previously reported study we piloted use of a web page that closely resembles the VA’s EHR, permitting remote presentation of actual patient data to human research subjects who navigate, review and respond to the record[3]. Because we were presenting data to individuals not involved in the care of the patients whose records were sampled, several considerations were required:

- Protecting patient privacy: The Health Insurance Portability and Accountability Act (HIPAA) forbids disclosure of any information which may be used to identify patients. HIPAA regulations list 17 types of identifiers which should be excluded from texts.
- Protecting identities of other individuals mentioned in the record. These could not only identify the patient, but could also identify treating staff, whom we also treated as research subjects.
- Preserving an intact clinical story: Because the cognitive research addresses the understandability of the record, maintaining a consistent, but disguised chronology of events and observations is essential.

Thus, de-identification is a necessary step to enabling valid EHR interface research.

The current work was motivated by desire for a de-identification method more efficient than a manual approach used in previously reported experiment that studied users’ appraisals of clinical document quality. For the reasons listed above this study also required de-identification of a set documents belonging to a single patient [4, 5]. At the time, we lacked both an automated system and the resources to develop one, so a manual approach was used to replace all potential patient identifiers with dummy data. Names of staff and other individuals and institutional, alphanumeric and geographic identifiers were also disguised. All dates were shifted by a fixed number of days, and time stamps were truncated to the nearest hour. This was accomplished using the search/replace feature of a commercial word processor. Text highlighting marked changes to avoid making multiple changes. Expectably, the effort was time-consuming. Date shifting was especially problematic. Elimination of errors required repeated audit and edit cycles. More than 40 hours of labor were required to de-identify 30 documents. Because VA patient records contain hundreds of documents, it is desirable to investigate how users cope with large record sets. However, the cost of de-identifying large document sets poses a significant constraint on conducting such study. A more practical solution was sought.

Much of the published research in medical record de-identification methodology is motivated by the desire to enable sharing reference sets of patient care documents to allow investigators to improve natural language processing (NLP) and data mining methods while maintaining patient privacy [6-8]. Few health care settings have access to investigators with skills in NLP, and few computer science or linguistics departments have access to health records. Clinical decision making, cognitive and human interface research are other areas that require de-identification technology for somewhat different reasons. In addition to suppressing patient and staff identifiers, studies of
this type also require that de-identified texts present a coherent and realistic clinical story.

Methods

The source of documents was a corpus of 3.7 million clinical notes belonging to approximately 88,000 patients produced over three years at a single institution and stored in a relational database. To test decision making related to common chronic conditions, cases concurrently diagnosed with coronary artery disease, Posttraumatic Stress Disorder, diabetes and hypertension were identified. Six of these cases, accounting for 585 documents, were selected for the de-identification experiment. Document text and metadata (date, author, note title and treatment location) were extracted from the data base into an XML document.

Time stamps, to the nearest second in many cases, were a frequent potential identifier. These were zeroed out using the regular expression replacement feature in Notepad++. 1,497 time stamps were thus found and zeroed. 21 laboratory accession numbers, consisting of a lab test code followed by strings representing the date and serial number were similarly de-identified by this method. The resulting XML object was processed using public source Harvard-MIT de-identification software described by Neamatullah et al. [9]. This PERL-based system, using dictionary lookups and regular expressions, was developed to process general medical text such as nursing notes and discharge summaries. For use in the VA, lookup lists were augmented with surnames and given names from VA patient and staff databases and lists of local hospital and geographic terms. The program’s regular expression tools for finding dates and numeric identifiers were used unmodified. The cleaned XML patient records were processed by the Harvard-MIT system to mark potential identifiers and dates. The resulting annotations were passed to a Plone-based program to create a manual editing environment (http://plone.org).

For the editing interface, the cleaned XML document was segmented using the Lucene whitespace tokenizer, and tokens were joined to corresponding annotations from the de-identification program. 12,695 distinct (case-sensitive) tokens, 72,321 tokens in all, occurred in the corpus. For the editing task, tokens in each document set belonging to a patient were grouped, ranked by frequency, and flagged as to whether they conveyed potential identifying information. Shifted dates were displayed as action buttones alongside the original date. Token instances could be edited by editing a “replacement text” field, and results of changes could be viewed. The editor allowed string searching and enabled sorting the list by “potential identifier flag”, term, and term frequency (Figures 1-3). When editing was complete, a synthesized document incorporating the replacements was output for manual audit, and a de-identified XML object created for loading into the simulated EHR interface. An example of final output document is shown in figure 4.

The editing screen displays text, potential identifier flag, rank, replacement text and links to source documents. Here, “Lnm” replaces the surname “Smith” in an output document. Terms are case sensitive and the entire list is searchable. Ability to sort by term, rank, and identifier flag aids efficiency in processing the list.
This view of the editing screen shows several replacement terms. Items that resemble dates have a button displaying a suggested replacement date, which is inserted if the button is clicked. Efficiency is gained when multiple occurrences of dates in text are encountered. When a date-like form that might not be a true date occurs, the record link can be clicked to check. In this case, “8/10” occurred in the phrase “current pain level 8/10”, and was left unchanged.

This is the display shown when a link is selected. This example illustrates how frequently dates, many generated from the data base, appear in text.
An experiment was conducted wherein 585 documents belonging to 6 patients were reviewed and de-identified using the system. The goal was to spend approximately one hour editing each electronic chart.

**Results**

The mean number of documents per case was 97.5. Edit time was measured to the nearest 5 minutes and mean edit time per case was 65 minutes (range: 60-75 minutes, SD 7.46). Output from the experiment was manually audited for missed instances of identifying information, and 101 omission errors were discovered. Audit time required approximately 4 hours. Table 1 shows results of the experiment and Table 2 shows the category distribution of true positive and false negative tagging.

Inspection of Table 1 shows that the de-identification algorithm labeled 13,290 term occurrences as positive. Most of these were common words which could also be interpreted as proper names. False negatives were not exhaustively assessed because of their lack of saliency, but the most common patterns missed by the algorithm and identified by the reviewer were numeric identifiers, solitary month abbreviations (e.g. “DEC”) and solitary references to years (e.g., “1999”). Sorting the token list and searching for potential missed items made it possible to correct many occurrences, but in some instances, global changes generated errors in the final output. For example, example after the capitalized token ‘May’ had been heuristically substituted with ‘June’, occasional output phrases of the form “Activity
Restrictions: June return to pre-hospital activities immediately” were generated. We chose to ignore such errors because they were few, because they did not breach privacy and because they appeared to interfere minimally with human understanding of the text.

The most frequent sources of error not discovered in the editing sessions and found by audit were hyphenated last names (staff names) and multiword hospital names (places). These errors were corrected when found.

<table>
<thead>
<tr>
<th>Raw Machine Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>#labeled Positive</td>
</tr>
<tr>
<td>13,290</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Results after Human Editing</th>
</tr>
</thead>
<tbody>
<tr>
<td>#labeled Positive**</td>
</tr>
<tr>
<td>3,568</td>
</tr>
</tbody>
</table>

*not fully assessed; ** estimated

Table 1. Results of De-identification

<table>
<thead>
<tr>
<th>Distribution of identifiers among the 3,567 true positive tagged identifiers (after editing):</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dates</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>1,403</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Distribution of false negative items:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dates</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>13</td>
</tr>
</tbody>
</table>

Table 2: Category distribution of tagged and missed identifiers.

These results certainly indicate areas where the de-identification algorithm could have been improved, but because our primary intent was to improve de-identification efficiency in order to use a larger set of records for cognitive testing, we did not attempt further refinements of the algorithm. This is certainly a future possibility, using standard scripting procedures. The de-identification system made it practical to prepare a realistic testing corpus ten times larger than the one previously used and, as before, the result was approved for research use by a hospital privacy officer.

In contrast to the manual approach use of this system gained efficiency by employing tokenization, which permitted bulk editing, and afforded a convenient way to identify and change the very prevalent date forms. As shown in Table 2, date forms were the most frequent “identifier” found, and using the date “button” accounted for much of the efficiency gain seen.
Discussion

Research investigation of patient records is limited by the extensive presence of potentially identifying information in medical texts. The very comprehensive HIPAA standard, combined with a high prevalence of HIPAA-proscribed identifying information presents a well known challenge to those who would wish to analyze textual data. Accordingly, considerable attention has been devoted to developing heuristic de-identification methods [6-9] and much progress has been made. Investigators who have reported on the performance of de-identification systems developed and trained on corpora available to them tend to report accuracies of 98% or above. Also reported is the experience that when a method developed at one site is applied “out of the box” in a new setting, performance degrades. The raw results we found (precision 0.19, recall 0.70) with the Harvard-MIT system are similar.

The high rate of false positive terms identified by the system resulted from the large number of tokens that could be interpreted as names, e.g. “He”, “Long”, etc. Many of these “ambiguous names” were relatively common words that are also relatively rare surnames. In large data bases, rare and difficult names (e.g., the surname “With”) are found, and present special difficulties. For record sets of the size studied, the false positive rate was not a barrier to efficient review, but could pose difficulty if the goal were to accomplish fully automated de-identification.

In this investigation, the primary goal was to find a practical method to de-identify a modestly large number of patient documents, not to develop a high performance de-identification tool. De-identification was a rate- and resource-limiting step in the sequence leading to our research objective of presenting realistic, yet de-identified patient records to test subjects. By adapting an effective existing tool, and integrating that tool with ability to view the source record, our goal was met. Using a manual method, a similar de-identification task required 40 hours to de-identify and audit 30 documents (.75 documents/hour). With the present method, 585 records were de-identified and audited in 10.5 hours (55.7 documents/hour), a 74-fold performance gain. Further improvements to the system are possible, but were not required in the present application.

Conclusion

For the planned cognitive testing application, scrupulous de-identification is required for institutional and ethical reasons. In addition, because adequate clinical understanding of the record is necessary, maintaining proper temporal relativity is important. The system we evolved represents a compromise between performance and expediency. Other than augmenting lookup tables, the de-identification component of Harvard-MIT system was unchanged. Integrating that system’s output with a Plone-based editor was straightforward, and resulted in a system capable of efficient high-quality de-identification of patient record sets. Because it requires human intervention, our method is suitable for application to hundreds to a few thousand patient documents, but not the many more required, for example, if data were to be shared among institutions to conduct large-scale epidemiologic research. Combined with simulator technology replicating the EHR experience, the de-identification method described surmounts important logistical barriers to cognitive testing with clinicians.

Acknowledgement

This work was supported by the United States Department of Veterans Affairs Health Services Research and Development Service, the VA Puget Sound Health Care System and the University of Washington Information School. 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.

References
