Computer-Assisted De-identification of Free Text in the MIMIC II Database

M Douglass[1], GD Clifford[1], A Reisner[1], GB Moody[1], RG Mark[1]

[1]Harvard-MIT Division of Health Sciences & Technology, Cambridge MA, USA

Abstract

Medical researchers are legally required to protect patients' privacy by removing personally identifiable information from medical records before sharing the data with other researchers. We present an evaluation of methods for computer-assisted removal and replacement of protected health information (PHI) from free-text nursing notes collected in the intensive care unit as part of the MIMIC II project [1]. A semi-automated method was developed to allow clinicians to highlight PHI on the screen of a tablet PC and to compare and combine the selections of different experts reading the same notes. An analysis of the performance of three human expert de-identifiers and of an automated system is presented. Expert adjudication demonstrated that inter-human variability was high, with few false positives and many false negatives. The sensitivity of human experts working alone ranged from 0.63 to 0.93, with an average of 0.81, and the average positive predictive value was 0.98. An algorithm generated few false negatives but many false positives. Its sensitivity was 0.85, but its positive predictive value was only 0.37.

Even highly competent and motivated human experts make errors in de-identification, suggesting that multiple independent reviews are necessary to achieve acceptable levels of de-identification. Our preliminary results indicate that at least some automated methods may be as sensitive as human experts, but additional development is needed to reduce their false positive rate.

The de-identified database of nursing notes was re-identified with realistic surrogate (but unprotected) dates, serial numbers, names, and phrases to provide a gold standard database of over 2600 notes (approximately 340,000 words) with over 1700 instances of PHI. This reference gold standard database of nursing notes and the Java source code used to evaluate algorithm performance will be made freely available on Physionet [2, 3] in order to facilitate the development and validation of future de-identification algorithms.

1. Introduction

Patients expect their personal medical data to be shared only among the clinicians and others directly concerned with their case. When using the medical data for research purposes, we must continue to respect and preserve patient confidentiality. The de-identification process removes all explicit personal health information in order to dissociate the individual from his medical record, while still preserving all the medically relevant information about the patient.

In the United States, the guidelines for protecting the confidentiality of health care information have been established in the Health Information Portability and Accountability Act (HIPAA) [4]. Records are said to be de-identified when the risk is very small that the information can be used alone or in combination with other reasonably available information to re-identify the individuals. This risk can be calculated and documented statistically for all the records, or we can use the safe harbor approach and show that every record is free of the 18 types of identifiers listed in the law. Those identifiers include: names of patients and clinicians, all geographic subdivisions smaller than a state, all elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; all ages over 89, telephone and fax numbers, social security numbers, and medical record numbers. Such data is known as protected health information (PHI).

The MIMIC II database [1] contains medical records for over three thousand patients from the intensive care unit of a local hospital. Eventually the entire database will be made publicly available, but first all the PHI must be removed from the patient medical records, as required by HIPAA. Removing PHI by hand is a time-consuming and expensive task which may be prone to serious error. We are developing algorithms to perform the de-identification task automatically.

Algorithms for the removal of personal health information have been developed by other researchers, including Sweeney's Scrub system tested on clinical notes and correspondence [5]; the MEDTAG framework used on patient records including post-operative reports, laboratory and test results, and discharge summaries [6]; and Gupta's de-identification algorithm developed for surgical pathology reports [7]. However, these approaches concerned more highly structured data than the free-text medical notes we used in our study.

Before developing an algorithm, a representative corpus must be de-identified as fully as possible to present a "gold standard" against which to test algorithms. Furthermore, the performance of individual humans must be evaluated to compare to algorithmic performance. It is highly likely that the performance of any particular algorithm is dependent upon the statistical nature of the subject text. The analysis in this article concerns free-text nursing notes, which contain many spelling mistakes, subject specific abbreviations, and grammatical anomalies.

The aim of this study is to evaluate the accuracy of PHI removal of a consensus of up to four human experts, and to provide an annotated open-source corpus that will act as a resource for the development and evaluation of de-identification algorithms, together with the software used to perform the de-identification and adjudication.

2. Methods

A Java-based interface was developed to facilitate the identification of PHI within the corpus of nursing notes by three independent clinicians. A fourth clinician adjudicated the union of the three clinicians' choices to determine inter-expert variance and to create a "gold standard". Finally a previously published simple algorithm was passed over the corpus to further reduce the possibility of human error.

2.1. The corpus

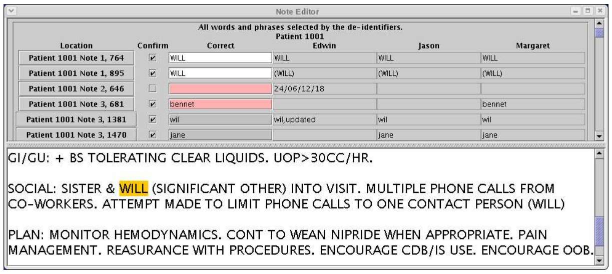
Medical data is collected as part of the MIMIC II project from all patients admitted to the intensive care units of a local hospital [1]. The nursing progress notes are unstructured free text typed by the nurses at least twice a day, and include observations about the patient's medical history, his current physical and psychological state, medications being administered, laboratory test results, notes about visitors, and other information about the patient's state. In these notes, the nurses frequently employ technical terminology, non-standard abbreviations, ungrammatical statements, misspellings, and incorrect punctuation and capitalization.

The corpus we used includes notes from 148 randomly selected patients. There are a total of 2,646 notes, with a total word count of 339,150. Of those notes, 119 have been manually "enriched" to include PHI that is especially difficult to identify (such as "foley catheter" and "Parkinson's disease") and to include more instances of infrequently appearing types of PHI.

To determine the approximate corpus size needed, a standard sample size estimate [8] can be used.

$$N = p(1-p)\left(\frac{Z(1-\frac{\alpha}{2})}{E}\right) \qquad (1)$$

where E is the margin of error, p is the population proportion, and $Z(1-\frac{\alpha}{2})$ reflects the desired level of confidence. Since we wish to distinguish between a 90% and 93% accuracy level, $E = 0.03$ and $Z(1-\frac{\alpha}{2}) = 1.96$ (from tables). A conservative value for p is 0.5, which maximizes the value of N in equation 1 (see [8]). Following this formula, at least 1068 instances of PHI are required in our testing database.



All words and phrases selected by the de-identifiers.
Patient 1001

Location	Confirm	Correct	Edwin	Jason	Margaret
Patient 1001 Note 1, 764	☑	WILL	WILL	WILL	WILL
Patient 1001 Note 1, 895	☑	WILL	(WILL)	(WILL)	(WILL)
Patient 1001 Note 2, 646	□		24/06/12/18		
Patient 1001 Note 3, 681	☑	bennet			bennet
Patient 1001 Note 3, 1381	☑	wil	wil,updated	wil	wil
Patient 1001 Note 3, 1470	☑	jane		jane	jane

GI/GU: + BS TOLERATING CLEAR LIQUIDS. UOP>30CC/HR.

SOCIAL: SISTER & WILL (SIGNIFICANT OTHER) INTO VISIT. MULTIPLE PHONE CALLS FROM CO-WORKERS. ATTEMPT MADE TO LIMIT PHONE CALLS TO ONE CONTACT PERSON (WILL)

PLAN: MONITOR HEMODYNAMICS. CONT TO WEAN NIPRIDE WHEN APPROPRIATE. PAIN MANAGEMENT. REASURANCE WITH PROCEDURES. ENCOURAGE CDB/IS USE. ENCOURAGE OOB.

Figure 1. Screen shot of the comparison mode showing how to make a consensus based on the selections of 3 clinicians. The first column displays the location of the PHI. The second column of check-boxes confirms that the word should be classified as PHI. The third column of text fields is the word being classified. The last three columns show what the clinicians selected at that place in the text. The bottom half of the display shows the note context for the PHI selection "WILL". (The PHI in this display has been replaced with surrogate data to conform with HIPAA.)

2.2. Labeling by clinicians

Medical house officers from local hospitals were recruited to locate and label the PHI in the nursing note corpus. Every clinician was able to read about 80,000 words in a 4 to 5 hour time block, including breaks. They were paid $50 per hour, with the additional incentive of a $200 bonus for the best performer in a group of 6 de-identifiers. A total of 11 different expert clinicians independently scored 20.8% to 43.3% of the corpus.

Each clinician was given a text definition and examples of what is defined by HIPAA as PHI. They were encouraged to make a best guess for ambiguous cases. A Java application was designed to display the nursing note text in an easily readable format and to collect locations of the PHI identified by each clinician. The software was run on a tablet PC, and clinicians located PHI by tapping the word on the screen with the tablet's pen. The locations of the PHI in every note were written to a file.

2.3. De-Identification algorithm

A simple automated de-identification algorithm written in Perl was developed for in-house use [9]. First it uses pattern-matching to identify potential dates, telephone numbers, social security numbers, and other protected types of identification numbers. Next it uses look-up tables to identify potential locations and patient, clinician, and hospital names. Finally the algorithm applies several simple context-based rules, such as the word following "Dr" will often be the doctor's last name. See [9] for further details.

2.4. "Gold Standard" formation

The nursing notes corpus was separated into four sets approximately equal in size, and each set of notes was de-identified by three clinicians independently. A subset of the data was de-identified by four clinicians, but no advantage was found by adding the fourth person. The PHI selections of multiple doctors looking at the same notes were combined using software developed for this project. In the Java interface as shown in Figure 1, the selections of all clinicians for each note are displayed, and a suggestion for the correct text is generated based on the majority response. A clinician from our group reviewed the selected PHI and checked the context of each selection in the original note text in order to make the final decision as to whether a word should be classified as PHI. Finally we ran our algorithm on the same nursing note text and went through the results to identify any PHI not found by the clinicians. This PHI was also verified by a clinician. By the time a note is pronounced completely de-identified, four different clinicians and one algorithm have looked at the text.

For comparison purposes, we created consensuses without an outsider adjudicator for two clinician subsets and for three clinicians. The unadjudicated consensuses were created by taking the union of all selections. Most of the errors made during human de-identification are false negatives (FN), so taking the union minimizes the number of missed FNs.

2.5. Evaluating performance

The selections of a single de-identifier are compared to the completely de-identified gold standard to generate statistics on the sensitivity and positive predictive value for that de-identifier's performance. We adjudicated the evaluation to decide when to count agreements and disagreements as separate instances. The software parses every word as a separate instance. For example, someone missing "New York City" has missed only one instance of PHI (a city name) and should not be penalized for missing three separate instances.

Table 1. De-identification Performance for humans and for an automated algorithm. The "gold standard" is the adjudicated union of the algorithm and three independent human experts. PPV = Positive Predictive Value.

		Min	Max	Mean
1 person	Sensitivity	0.63	0.94	0.81
	PPV	0.95	1.0	0.98
2 people	Sensitivity	0.89	0.98	0.94
	PPV	0.95	0.99	0.97
3 people	Sensitivity	0.98	0.99	0.98
	PPV	0.95	0.99	0.97
Algorithm	Sensitivity	-	-	0.85
	PPV	-	-	0.37

2.6. Re-identification

In order to make the labeled corpus available to the public and conform with HIPAA regulations, the PHI must be removed and replaced with authentic-looking surrogate data. All the dates in a given record were shifted by the same random number of weeks and years, but the days of the week were preserved. The names were replaced with names adapted from publicly available lists of Boston residents with randomly swapped first and last names, in order to get a wide variety of ethnicities and non-standard or unusually spelled last names. Locations were replaced from randomly selected small towns in a different part of the country. The hospital-specific terms, like names of buildings and special wards, were given fictitious names for a fictitious hospital.

A Perl algorithm used the locations of all the PHI to extract the protected text, classify it according to type of PHI, and then suggest an appropriate but still randomly chosen replacement for the text. A Java graphical user interface displayed the suggested replacements and allowed a reviewer to edit or replace the text. The reviewer could also examine the original context to verify that the replacement was reasonable. The capitalization was adjusted based on the surrounding text. Most of the corpus is untouched during the re-identification process, so all the relevant medical information is preserved. The locations of all the surrogate PHI were recorded for use in future algorithm testing.

3. Results

We documented the performance of single clinicians' selections, the union of two clinicians' selections, and the union of the selections of three clinicians reading through the corpus. The statistics are displayed in Table 1. Individual performance varied greatly, with the sensitivity ranging from 0.63 to 0.94. When combining all the selections made by two people, the sensitivity increased to an average of 0.94 without seriously affecting the positive predictive value. The union of three had an even higher sensitivity. The number of FNs for an individual is high and the number of false positives (FP) is low. Having more people look at the notes reduces the number of combined FNs while adding only a small number of FPs.

The algorithm had a sensitivity of 0.85, which is better than the average human although less than the union of two humans, but it had a very low positive predictive value of 0.37 since it identifies many FPs. However, the algorithm does detect most PHI, and it even detected PHI not found by any of the human de-identifiers.

4. Discussion

The results show the limitations of human de-identification of medical data. The combined efforts of four clinicians were needed to completely de-identify the test corpus of the nursing notes to a level of 98% (100% included adjudicated algorithm results combined with the human results). The simple algorithm therefore found another 2%. Tools have been developed to facilitate the process of using a team of humans to perform the task, but human de-identification is still a very time- and manpower-intensive process. There is a clear need for accurate, fully automated de-identification algorithms.

The simple algorithm evaluated here is an early draft and is far from perfect, but it already has a higher sensitivity than the average human de-identifier. The algorithm's high false positive rate can be improved with more sophisticated contextual rules, and we expect that considerable improvement in sensitivity will also be achievable. It seems reasonable to expect that in the near future the performance of automated de-identification algorithms will significantly surpass that of multiple human de-identifiers.

The re-identified reference database will be publicly available on Physionet [2, 3] for the use of the research community. The corpus contains nursing notes from 148 patients, a total of 2,646 notes, a total word count of 339,150, and the corpus includes 1,776 instances of PHI.

5. Conclusion

We have created tools to be used for the evaluation of different methods of de-identification of intensive care unit nursing notes. A single human expert cannot reliably remove all the PHI from a large data set. The software we developed for recording and combining the selections from manual de-identification of text allows a team of clinicians to collaborate to completely de-identify medical records. The gold standard database of re-identified nursing notes along with the locations of the known PHI in the corpus can be used for testing and evaluating automated de-identification algorithms. Automated de-identification algorithms will almost certainly become critical tools for researchers preparing to share text-based medical records with the research community.

Acknowledgments

This publication was made possible by Grant Number R01 EB001659 from the National Institute of Biomedical Imaging and Bioengineering.

References

[1] Saeed M, Lieu C, Raber G, Mark R. Mimic ii: A massive temporal icu patient database to support research in intelligent patient monitoring. Computers in Cardiology 2002;29:641–644.

[2] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. Circulations 2000;101(23):e215–e220.

[3] http://www.physionet.org/.

[4] Health insurance portability and accountability act of 1996.

[5] Sweeney L. Replacing personally-identifying information in medical records, the scrub system. Proc AMIA Symp 1996; 333–337.

[6] Ruch P, Baud R, Rassinoux AM, Bouillon P, Robert G. Medical document anonymization with a semantic lexicon. Proc AMIA Symp 2000;729–733.

[7] Gupta D, Saul M, Gilbertson J. Evaluation of a de-identification software engine: Progress towards sharing clinical documents and pathology reports. Am J Clin Pathol 2004;121(2):176–186.

[8] D'Angostino RB. Introductory Applied Biostatistics. Houghton Mifflin College Div., 2001.

[9] Levine J. De-identification of ICU Patient Records. 77 Mass. Av. Cambridge, MA, USA: MIT Press, 2003. MEng Thesis.

Address for correspondence:

Margaret Douglass
Laboratory for Computational Physiology
Harvard-MIT Division of Health Sciences & Technology
Rm E25-505, 45 Carleton St.,
Cambridge MA 02142 USA
douglass@mit.edu