De-identification of Address, Date, and Alphanumeric Identifiers in Narrative Clinical Reports

Mehmet Kayaalp, MD, PhD, Allen C. Browne, MS,
Zeyno A. Dodd, PhD, Pamela Sagan, RN, Clement J. McDonald, MD
Lister Hill National Center for Biomedical Communications,
U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD

Abstract

Introduction: *The Privacy Rule of Health Insurance Portability and Accountability Act requires that clinical documents be stripped of personally identifying information before they can be released to researchers and others. We have been developing a software application, NLM Scrubber, to de-identify narrative clinical reports.* ***Methods****: We compared NLM Scrubber with MIT's and MITRE's de-identification systems on 3,093 clinical reports about 1,636 patients. The performance of each system was analyzed on address, date, and alphanumeric identifier recognition separately. Their overall performance on de-identification and on conservation of the remaining clinical text was analyzed as well.* ***Results****: NLM Scrubber's sensitivity on de-identifying these identifiers was 99%. It's specificity on conserving the text with no personal identifiers was 99% as well.* ***Conclusion****: The current version of the system recognizes and redacts patient names, alphanumeric identifiers, addresses and dates. We plan to make the system available prior to the AMIA Annual Symposium in 2014.*

1. Introduction

Electronic health records are treasure troves for clinical scientists because with the availability of high volumes of electronic reports, clinicians are no longer limited to a cohort of their patients and can easily test their hypotheses on much larger samples. Access to those records, however, is not easy and involves overcoming a number of institutional barriers. These barriers have been raised purposefully to ensure that only the right person could access the private information of the patient. While these barriers had been the primary means to protect patient privacy, the requirements were so difficult to attain that they also became a barrier to scientific progress. Having seen both sides of the issue, the U.S. Congress enacted Health Insurance Portability and Accountability Act (HIPAA) in 1991, which tasked the U.S. Department of Health and Human Services (HHS) with regulating access to health records while protecting the health information of individuals.

The Privacy Rule of HIPAA requires that clinical documents be stripped of personally identifying information before they can be released to researchers and others. There have been several attempts to de-identify clinical text data automatically via software with an upward trend of performance, yet the clinical research community still considers human verification and validation necessary prior to the release of any automatically de-identified clinical data.

We have been developing a software system to automatically de-identify clinical records, which we have named NLM Scrubber. The current version of the system recognizes and redacts patient names, alphanumeric identifiers, addresses, and dates. The primary goal in clinical text de-identification is to raise the sensitivity performance of the system to an acceptable level so that de-identified data can be used with no need of human verification. We designed NLM Scrubber with that goal in mind and this study is a new step forward to reach that goal.

2. Background

As defined by HHS, Protected Health Information (PHI) comprises a subset of the health information of an individual who is the subject of the health record *and* personally identifiable information[*] (PII), including demographic information, collected from the same individual to be used by the health care provider, health plan, employer or health clearinghouse. PII is any information that distinguishes or traces an individual's identity such as

[*] The text of CFR 45 § 164.514 uses the term *individually* identifiable information instead of *personally* identifiable information. One possible reason is that the meaning of the legal term *person* also includes entities other than natural person (human) such as trust, estate, partnership, corporation, and professional association among others. Since personally identifiable information and its acronym PII are more widely known terms, we used them here instead.

name, social security number, date of birth or biometric records and any other information such as medical, financial and employment information that is linkable to an individual.[2][3]

Table 1. Per HIPAA Privacy Rule, the following identifiers must be deleted from PHI to fully de-identify health information. (*) As of 2010, there were 18 sets of zip codes with distinct initial three digits whose corresponding population sizes were less than or equal to 20,000.[1]

1. Names
2. All geographic subdivisions smaller than a state, except the first two digits of the zip code of the postal address. The third digit of the zip code can also be left intact, only if the size of the population in the area of the censored two digits is greater than 20,000 according to the most recent census data.(*)
3. All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older.
4. Telephone numbers.
5. Fax numbers.
6. Electronic mail addresses.
7. Social security numbers.
8. Medical record numbers.
9. Health plan beneficiary numbers.
10. Account numbers.
11. Certificate/license numbers.
12. Vehicle identifiers and serial numbers, including license plate numbers.
13. Device identifiers and serial numbers.
14. Web universal resource locators (URLs).
15. Internet Protocol (IP) address numbers.
16. Biometric identifiers, including fingerprints and voiceprints.
17. Full-face photographic images and any comparable images.
18. Any other unique identifying number, characteristic, or code, except the ones that may be generated by the covered entity for re-identification.

HHS developed the Privacy Rule, defining certain identifiers as part of PHI, which should be de-identified before health records are accessed for research purposes (see Table 1). Note that the health information dissociated from those identifiers of the individual is not considered PHI. According to the Privacy Rule the identifiers in Table 1 that belong to the individual or relatives, employers or household members of the individual, should not be present in any de-identified health records.[4]

2.1 Current Clinical Text De-identification Systems

De-identification of structured data is a fairly straightforward process, where the content of fields defined as containing PHI (e.g., name and birth date fields) be removed and made inaccessible to researchers. De-identification of unstructured data or free text, on the other hand, is a challenge. Because of the idiosyncrasies of any natural language, including English, the utterances of information are not always predictable and we have to devise intelligent tools to recognize those words and phrases containing PHI.

A thorough review of 18 clinical text de-identification systems has been published recently.[5] Since then only four additional systems have appeared in major journals.[6-9] These 22 systems can be categorized in two groups based on their target documents: general purpose vs. niche (specialized) de-identifiers. They can also be classified in terms of their underlying methodologies, which use either symbolic or machine learning approaches. Symbolic approaches mainly rely on rules, regular expressions, and lookup tables (also referred to as dictionaries or gazetteers). The availability of a de-identification system is another important characteristic; some are freely available, some are commercial products, and others have not been made available.

Currently, there are only five freely available systems, three of which were specialized to de-identify surgical pathology reports only.[10-12] The other two systems are general purpose de-identification systems developed by researchers at MIT and MITRE. MIT's system took a symbolic approach; whereas, MITRE's is a machine learning system using conditional random fields.

The name of MIT's system was not mentioned in their publication[13] but the filename of the code was *deid.pl*. Since there is another (commercial) system with the same name, De-ID, to prevent any confusion, we here call MIT's system "MITdeid". MITdeid provides various features that are closely tuned to clinical setting, such as accepting a list of provider names of the institute and the full name of the patient per report.

MITRE's system, MIST, was developed to demonstrate how an existing conditional random field program designed for generic use could be repurposed quickly as a successful clinical text de-identification system.[14] MIST has proven itself as one of the most successful systems in the i2b2 competition in 2006.[15] As a machine learning system, MIST requires a training dataset. The current version of the system does not store the constructed model and has to be re-trained before each testing session.

As part of Clinical Text De-identification Project at NLM, we studied personal name recognition in great depth. Since we previously reported the results of the name scrubbing performance of our system elsewhere, we do not repeat them here.[16]

2.2 Contribution of NLM Scrubber

There have been several attempts to de-identify clinical text data automatically via software, but all those systems that are freely available had failed to scrub some identifiers; showing the need for improvement. As part of HHS, NLM started the clinical text de-identification project to respond to this need and promote scientific progress by enabling the research community to access large amount of health information that does not contain personal identifiers. NLM Scrubber will be freely available at the time of this publication. The goal of NLM Scrubber is to facilitate the production of de-identified clinical text data, and minimize (if not eliminate) the burden of manual de-identification for the clinical research community.

3. Methods

In this section, we present (1) how we select and process the clinical text data from a large corpus of clinical reports in order to reliably study and develop robust de-identification methods; (2) the methods and components of our de-identification tool; and (3) our evaluation methods.

3.1 Data Selection

A hospital information system may retain many copies of a given report (e.g., the initial report, the report with addenda or corrections) and HL7 messages logs that we used do the same. A sample with duplicate reports may inflate the magnitude of events observed in the study. Duplicate reports in a de-identification test set would distort the analysis of tests and the success of a de-identifying program.

To ensure the reliability of the study results, we devised a random sampling method to exclude any redundant reports. For each randomly selected patient, we collected all reports generated during a particular visit of the patient, grouped them by report types, sorted each group by report filing time, and took only the most recent, presumably the most developed report in that group.

Our sampling method relies on the assumption that each visit is associated with a unique visit number and reports of the same type in two different visits are sufficiently dissimilar. Note that this assumption may not always hold. After performing the sampling, we sorted reports of each patient by word counts. The manual comparison of the reports that were similar in size helped us discover partially duplicate records from distinct visits of the same patients. We eliminated the earlier reports from these sets. This approach may inadvertently eliminate some non-duplicate reports, but in the final analysis, it yields an unbiased, large spectrum of reports per visit with distinct report types.

In this study, we used 3093 distinct clinical reports about 1636 distinct patients of the Clinical Center at NIH. The maximum number of reports per patient was 20.

We developed NLM Scrubber using a training set of 1140 clinical reports from the same institution. Unlike the study test data, retrieval of the training data was done in several iterations over a long period of time in an ad hoc and not truly randomized manner. The patient cohorts in the training and test data were mutually exclusive.

3.2 PII Recognition Methods

We collapse 18 types of identifiers defined by HIPAA into four distinct categories: 1) names, 2) addresses, 3) dates (incl. ages), and 4) alphanumeric identifiers (e.g., medical record numbers, report identifiers, telephone numbers). We reported our results with the first category, names, previously [16]. Here we report our results with the other three categories.

3.3 Alphanumeric identifier recognition

We define an alphanumeric string as a string of characters containing at least one or more digits. It may or may not contain other characters. Alphanumeric Identifier Recognizer (AIR) has a two-prong approach: It detects patterns that correspond to alphanumeric strings such as phone and social security numbers that need to be labeled as alphanumeric-Id, but it also detects patterns of known clinical entities such as lab values that need to be preserved. AIR also attempts to distinguish alphanumeric strings from date-like patterns so that dates would not be mislabeled.

If a token of alphanumeric string contains only a single digit, AIR ignores it completely; otherwise, it analyzes the content of token t and its context. If t is preceded by a token containing certain strings such as number, protocol, or # sign, it labels t as an alphanumeric identifier. If an alphanumeric token containing a sequence of two or more upper case letters is followed by certain tokens such as “protocol”, it is labeled as alphanumeric identifier. A 9–10 digit number pattern with or without delimiters in between is also defined as alphanumeric identifiers (i.e., phone or social security numbers).

AIR also checks numerous conditions (e.g., a number followed by a unit of measure) that may indicate that token t be a valid piece of clinical data to be conserved and marks most other alphanumeric strings as alphanumeric identifiers.

3.4 Date and age recognition

Algorithms for identifying dates and ages are based on a set of regular expressions to detect the corresponding patterns. Some date patterns are listed in Table 2. For example, string 07-08-2012 would be identified using the pattern DD*MM*YYYY, where “*” is a delimiter and D, M, Y are digits such that YYYY should be greater than 1900 and less than the current year, $1 \leq$ DD ≤ 31 and $1 \leq$ MM ≤ 12.

Table 2. Date Patterns: D, M, Y, h, and m are date, month, year, hour and minute digits; * is a delimiter; MONTH, HOLIDAY are literal values of month and holiday incl. their abbreviations; X? denotes that X is optional; | indicates distinct disjunctive patterns

Pattern	Example	Pattern	Example
YYYY*MM*DD	2012-08-07	DD*MM?	07-08, 07-8
DD*MM*YYYY	07-08-2012	YYYYMMDD	20120708
MM*DD*YYYY	08-07-2012	YYYYMMDDhhmm	201207081215
MM*DD*YY	08-07-12	YYYY	2012
M*DD*YY	8-07-12	DD?*?MONTH	7-August, 7August, 7 Aug
YYYY*YYYY	2011-2012	MONTH*YY(YY)?	August.2012, August’12
DD*MM*YY	07-08-12	(early|mid|late)*YYYY	Mid-2012
M*D*YY	8-7-12	YYYY*?MONTH	2012/August, 2012Aug
M*DD*YY	8-07-12	’YY*?MONTH	’12-August, ‘12Aug
MM*YYYY	08-2012	DD?MONTH*YY	7August’12
DD*MM*DD*MM	07-08/08-08	MONTH*DD?	Aug7, August 7
MM?*DD	08-07, 8-07	MONTH	Aug, August
DD?*MM	07-08, 7-08	HOLIDAY	Christmas, Easter
MM*DD?	08-07, 08-8		

Unlike date patterns, age patterns are more involved. For example, age patterns may be required to catch phrases like “on his **ninety-third** birthday” or “in his late **90s**”.We classified alphanumeric age expressions and labeled them with specific names (see Table 3). The corresponding patterns are recognized through regular expressions.

Whenever a date (age) regular expression is matched with the tokens in the text, those tokens are labeled as date (age).

Table 3. Alphanumeric Age Expression Classes

Expression Classes	Examples
AGE-WITH-SUCCEEDING-MARKER	he was [93 years-old]
AGE-WITH-PRECEDING-MARKER	at the [age of 93],
AGE-WITH-APPENDED-UOM	his father, [93yo], has
AGE-FRACTION-EXPR	he is [5-years and 3-months] old
AGE-FROM-PHRASE-CONTEXT	she [was nearly 93].
AGE-BIRTHDAY-CONTEXT	on his [ninety-third birthday]
AGE-DECADE-CONTEXT	in his late [90s]
AGE-SIMPLE-CATCH-ALL	(as 93)
AGE-COMPOUND-CATCH-ALL	(93 and 90)

3.5 Address identifier recognition

Addresses are recognized mostly via the "shapes" component of dTagger, a specialized part-of-speech tagger extended with limited pattern tagging abilities for entities, such as addresses.[17] The dTagger searches address terms in various lexicons, which contain city and states names as well as street types and their abbreviations (e.g., Avenue, Alley, Blvd, and Circle). In its current format, the recognizer is difficult to maintain and will be revised before the release of the software package; therefore, we do not provide any further specifications of the soon-outgoing recognizer in this report.

3.6 Redaction

The redactor removes the PHI content in a post-processing step, where it replaces the removed text with a corresponding standard PHI label. For example, John would be replaced with [NAME]. If two distinct recognizers (e.g., both date and alphanumeric-Id recognizers) tag the same token as PHI, the redactor labels the content as [PHI] instead of choosing one tag over another.

3.7 Evaluation Methods

We evaluated the NLM Scrubber on a test set of 3,093 dictated narrative reports generated at the NIH Clinical Center. The set was annotated by two experts, a linguist and a registered nurse, producing the gold standard for the test data. Following NLM Scrubber's run on the study data, we compared the resulting tags against the gold standard and evaluated them in terms of sensitivity, specificity, precision, and F_2 measures. We also evaluated the privacy risks due to the revealed PHI tokens.

We tested the scrubbing performance of two of the most prominent and freely available de-identification systems, MIST and MITdeid against the same data and used the same evaluation approach for all of them.

Since MIST is a machine learning system, it requires training before testing. We used our held-out set of 1,140 annotated reports as training data for MIST. After testing MIST extensively on the training dataset using various parameterizations and based on consultations with its developers, we decided to run it with a bias of –4, which greatly favors sensitivity over precision but not to the extent that the results become unreliable. We appreciate the generous assistance we received from many members of the MIST developer team at different phases of our study.

3.8 Evaluation of differently tokenized results

Most de-identification systems come with their own tokenizers producing different sets of incompatible results. In order to compare the results and to report token misalignment errors, evaluators devised terminology such as colliding tokens, boundary detection failure and partially tagged tokens. For example, Deleger et al. reported that partially tagged PHIs due to boundary detection errors were 13% of all tagging errors.[7] Some researchers in the NLP community also use complex alignment schemas to remedy the problem.[18] When tokens produced by different systems do not match, the evaluation gets complicated and the differences between results become obscured. The situation gets more complicated as the number of systems to be compared increases. In the literature, we have not seen any proposed solution to the problem for robust evaluation of de-identification systems.

In this study, we align all outputs to be compared to the tokens of the gold standard. This method simplifies the evaluation without introducing any bias favoring one system over another: (1) We re-tokenize all outputs using the same tokenization scheme that the gold standard annotation has adopted. (2) When a token t in a system output does not correspond one-to-one to a gold standard token t_G, one of the following three scenarios is observed: (a) t may be a proper substring of t_G; (b) t_G may be a proper substring of t; or (c) t and t_G may overlap partially. After re-tokenization, the string of characters in t is distributed into a sequence of one or more tokens. We tag the resulting sequence of tokens with the original tag of t. (3) If t was tagged with a set of multiple tags originally, we apply them simultaneously to all tokens in the resulting sequence.

3.9 Evaluation of system's labeling of single tokens

Accurately distinguishing patient identifiers (e.g., telephone numbers and addresses) from provider identifiers is too difficult to attain for any text de-identification system. Recall that provider identifiers are not PHI. By design, NLM Scrubber (as well as most other clinical text de-identification systems) attempts to de-identify all personal identifiers regardless of whom they belong to. Throwing such a wide net does not degrade the quality of the de-identified text, since provider identifiers usually have no information value for clinical scientists.

Although this approach simplifies our task of catching *all* patient identifiers, it also complicates the evaluation process significantly. How should we evaluate labels of tokens related to providers? For example, if the system labels the physician's phone number as an alphanumeric identifier, should we count this label as true positive (TP)? Conversely, if the system misses a provider's phone number, should we count the instance as false negative (FN)?

When the aim is to protect patient privacy, the performance of a de-identification system should indicate the level of protection it provides. Inflating the TP count with the inclusion of de-identified non-PHI tokens would distort the actual performance. Note that in clinical reports, physician phone numbers are mentioned very frequently but patient phone numbers are extremely rare. Consider the case where a system de-identifies *all* physician phone numbers but misses a few, rarely-occurring patient phone numbers. Had we counted physician phone numbers as TPs, they would totally wash out the system's failure vis-à-vis the missed patient phone numbers. The resulting statistics would give a false diminutive impression on the system's failure on protecting patient privacy.

When we create the gold standard, should we then avoid labeling provider identifiers as PII? In that scenario, whenever a de-identification system recognizes a provider identifier as PII, we would have to count it as false positive (FP) and when it misses the provider identifier, we would have to count it as true negative (TN). Note, the design of our system requires de-identifying all personal identifiers. It would be incongruous to reward the system with a TN count when it misses a provider identifier or to penalize it with a FP count when it finds a provider identifier by following its design requirement faithfully. De-identifying a provider token (or keeping it intact in the text) has no effect on the actual performance of privacy protection or clinical information preservation; thus, we excluded provider tokens from performance evaluation.

The main purpose of our false positive rate analysis is to obtain an indirect measure about the level of preservation of scientific information present in clinical reports from de-identification. Most clinical information of any scientific value is represented in tokens that are not labeled in our gold standard annotation. The only exception to this rule is tokens representing non-PHI age (i.e., age < 90). The rate of preservation of non-labeled tokens and non-PHI age tokens is an indicator for the rate of preservation of clinical information.

If a system assigns a PII label to a token that actually is a patient identifier, we consider the decision as TP. If the system fails to label a patient identifier as PII, we call it as FN. A PII label would be FP if the token has no label in the gold standard (i.e., it is not PII) or if it denotes a non-PHI patient's age, because consequently the token would be redacted and the information would not be available to clinical scientists. If a non-labeled token or a non-PHI age token is labeled as non-PII, we consider it as TN.

3.10 Nonparametric analytic methods

Confidence intervals are staples of biostatics where samples usually come from a well-known parametric distribution and observations are random variables distributed independently and identically, which are not applicable to words in our dataset.

To estimate confidence intervals (CIs) in this study, we adopted a nonparametric bootstrap method,[19] bias-corrected, accelerated (*BC_a*) percentile intervals as implemented in package *boot* in R.[20] Through a bootstrap resampling strategy, we could truly simulate our initial sampling method. For each bootstrap sample, we randomly selected a patient and then included all reports (hence all associated token sequences) of the patient into the sample. We repeated this process until reaching the same number of patients in our original test data.

We computed statistical significance for the scores where CIs were overlapping, based on Wilcoxon paired signed test with Pratt's adjustments,[21] using the package *coin* in R.[22] This method is more suitable than bootstrap based *p* value estimation because it can successfully take into account that two sets of compared results are paired datasets.

These methods can be used in a wide-range of computational linguistic studies and provide a strong analytic footing for comparisons of different study results. We previously used them to compare and analyze information extraction performances of various systems.[23]

4. Results

In Table 4 on Alphanumeric-Id, Address, and Date rows, we reported each identifier recognition performance separately. For example, on the first row, the alphanumeric identifier recognition performance of NLM-S was isolated from the effects of date and address recognizers of NLM-S on the TP and FN counts. On the PHI rows, however, we reported the total effect of all three recognizers of each system.

Note that for any given system, neither TP nor FN counts on the PHI row are direct sum of the other three rows. For example, the total FN count of NLM-S (excluding PHI row) was 2 + 48 + 311 = 361, but the total missed PHI token count was 201, because some of the missed tokens by one recognizer (e.g., by Date recognizer) were caught by others (e.g., by Alphanumeric identifier recognizer). While we analyze each recognizer in isolation (before evaluating overall de-identification performance), it is important not to lose sight from the overall picture.

4.1 Alphanumeric Identifiers

In alphanumeric identifiers, NLM Scrubber (NLM-S) performance was clearly superior to others. It missed only two tokens, one of which was "406," a 3-digit area code of a telephone number, which should be considered non-PII since the area it covers is the entire state of Montana.[24] The other missed token was a protocol number, which is considered a low risk to privacy as the necessary information to re-identify the patient is not publicly available and such information is usually given to the patient's health care providers only.[25] MITdeid did not produce a viable alphanumeric de-identification on this dataset.

4.2 Addresses and Dates

In both addresses and dates, MIST results yield the best sensitivity scores, but on addresses, the sensitivity score difference between NLM-S and MIST was not statistically significant. After reviewing the false negative cases of NLM-S, we observed that most of the NLM-S's "missed address tokens" were actually non-PHI tokens such as geographical direction (e.g., Northern), state name abbreviation (e.g., VA), large city names in other countries (e.g., Beijing) and country names (e.g., England). None of the missed address tokens revealed a street address, but three of the revealed address tokens may cause some privacy concerns. They were Falls Church (Falls Church, VA: pop. 12,751) and Takoma (Takoma Park, MD: pop. 17,021).

Table 4. De-identification Sensitivity Results of NLM Scrubber (NLM-S), MIT's de-identification system (MITdeid), and MIST: Bold fonts denote the best results among the three, which are also statistically significant if their confidence intervals are written in bold fonts.

Identifier	Gold	System	TP	FN	Sensitivity
Alpha-Numeric-Id	4165	**NLM-S**	4163	2	**1.000 (0.998,1.000)**
		MITdeid	1444	2721	0.347 (0.334,0.359)
		MIST	4091	74	0.982 (0.977,0.986)
Address	292	NLM-S	244	48	0.836 (0.769,0.888)
		MITdeid	129	163	0.442 (0.371,0.510)
		MIST	250	42	**0.856** (0.791,0.905)
Date	29134	NLM-S	28823	311	0.989 (0.984,0.992)
		MITdeid	27595	1539	0.947 (0.942,0.951)
		MIST	28906	228	**0.992 (0.988,0.994)**
PHI	33591	**NLM-S**	33390	201	**0.994 (0.992,0.995)**
		MITdeid	29347	4244	0.874 (0.868,0.879)
		MIST	33310	281	0.992 (0.988,0.994)

In dates, MIST missed fewer date tokens than the other two systems and the differences were statistically significant. The sensitivity performance difference between MIST and NLM-S was 0.003, which however was statistically significant. NLM-S requires further sensitivity improvement on dates. None of the dates by NLM-S was tagged as PHI-Age (i.e., age > 89) in the gold standard.

Although trailing behind the other two systems, MITdeid showed strong sensitivity performance on dates (0.947), but not on addresses (0.442).

4.3 Overall Performance

It is not uncommon that a system tags a PHI token (e.g., a date) with a wrong label (e.g., an alphanumeric identifier). In such cases, there is neither a leakage of PHI nor a loss of clinical information. The PHI row in Table 4 indicates that there were a total of 33,591 PHI alphanumeric, address, and date tokens, of which NLM-S missed only 201, MIST 281 (40% more than NLM-S), and MITdeid 4,244 (21 times as many as NLM-S). NLM-S was clearly superior in overall sensitivity.

The decomposition of the revealed PHI tokens by PII types is displayed in Table 5. Note that the superiority of MIST on date and address de-identification that we observed in Table 4 is not present when we focus only on missed PHI instead of the accuracy of the classification into specific categories.

Table 5. Decompositions of Revealed PHI tokens by System

Tag	PII Type	Gold	NLM-S	MIST	MITdeid
AlphaNumericId	**AlphaNumeric-Id**	3502	0	24	1885
	Protocol-Id	660	1	3	659
	Telecom	3	1	0	0
Address	**Address**	292	48	40	163
Date	**Date**	29124	151	207	1532
	Age 90+	10	0	7	5
All		**33591**	**201**	**281**	**4244**

Although we do not have direct information about the inadvertent loss of clinical information due to over-de-identification, measures such as specificity, precision and F_2 may be used instead as indirect indicators (see Table 6).

Table 6. False Positive (FP), specificity, precision, and F_2 measures of the de-identification systems

System	FP	Specificity	Precision	F_2
NLM-S	5370	0.9950 (0.9947,0.9953)	0.861 (0.853,0.869)	0.964 (0.962,0.967)
MITdeid	2143	**0.9980 (0.9978,0.9982)**	**0.932 (0.926,0.938)**	0.885 (0.880,0.890)
MIST	3143	0.9971 (0.9967,0.9974)	0.914 (0.903,0.922)	**0.975 (0.971,0.978)**

MITdeid was superior in specificity and precision. Due to its low sensitivity score, however, its F_2 measure was not on par with others. In terms of F_2 measure, MIST did better than the other systems and NLM-S was a close second with a 0.011 difference.

5. Discussion

NLM-S revealed much fewer personal identifier tokens than the other two systems (see Table 5) and none of those revealed tokens were informative enough to disclose the identity of any patient. Our primary goal and our main criterion for success are to eliminate all patient related PII tokens when possible. As seen in results, NLM-S incurred a substantial number of false positives in order to catch the maximum number of identifiers, but the specificity penalty it incurred as a result was not higher than 0.005, which means that out of 1000 words only 5 of them were inadvertently deleted. This level of specificity does not have any significant adverse effect on the preservation of clinical information and on the readability of the resulting text. For keeping the trust of the U.S. Public to the research community, we have to continue working on improving the sensitivity of NLM-S especially on dates and addresses even if it costs us more false positives to achieve that. On the other hand, we are also cognizant of the needs of the research community and have to pay great attention to false positive rates and to the effective conservation of clinical information in the upcoming versions of NLM-S.

In our study data, NLM-S has recognized more PHI tokens than MIST and MITdeid have, which are the only freely available, general-purpose clinical text de-identification systems at the time of this study. Our risk analysis indicates that the revealed tokens would not cause any substantial risk to the patient privacy. Only three instances of address identifiers revealed the home city of three distinct patients, where the population sizes were less than 20,000 but greater than 12,500. Population size 20,000 was devised by the Privacy Rule as a threshold for further censoring zip codes (see Table 1).

MIST was clearly the second best performing system of this study. Due to their underlying methodological power, probabilistic machine learning systems do very well in this domain. Given that we devised our system based on the characteristics of the clinical corpus in our hand, we should not be surprised if MIST outperforms NLM-S in another clinical dataset with different characteristics.

As we indicated in one of our earlier studies,[23] probabilistic machine learning and symbolic linguistic methods are not an either-or proposition, a good NLP system should incorporate methods of both paradigms and reap the benefits of both worlds. We plan to develop a robust machine learning component to our scrubber so that it could perform well on a variety of reports from different origins.

Acknowledgements

We are grateful to the Scientific Counselors of LHNCBC and Dr. Olivier Bodenreider, Chief of Cognitive Science Branch at LHNCBC for their generous inputs and contributions to the project and to an earlier, extended version of this text. We are grateful to Dr. Jon McKeeby, CIO, Clinical Center at NIH and his staff for their help in obtaining and interpreting the clinical data. We thank Guy Divita, Dr. Yanna Kang, Selcuk Ozturk, and Shuang Cai for their contributions to the project. We also thank Drs. Lynette Hirschman, Samuel Bayer and Ben Wellner as well as John Aberdeen of MITRE, for their generous help and offers to test MIST on our study data.

Funding

This work was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

Competing Interests

The first author receives royalties from University of Pittsburgh for his contribution to a de-identification project. NLM’s Ethics Office reviewed and approved his appointment.

References

1. U.S. Census Bureau. ZIP code tabulation areas, 2010.
2. Department of Health and Human Services. Public Welfare; Administrative Data Standards and Related Requirements; General Administrative Requirements; General Provisions; Definitions. 45 CFR § 160.103.
3. McCallister E, Grance T, Scarfone K. Guide to protecting the confidentiality of personally identifiable information (PII). Recommendations of the National Institute of Standards and Technology. U.S. Department of Commerce, NIST, 2010.
4. Department of Health and Human Services. Public Welfare; Administrative Data Standards and Related Requirements; Security and Privacy; Privacy of Individually Identifiable Health Information; Other Requirements Relating to Uses and Disclosures of Protected Health Information. 45 CFR § 164.514.
5. Meystre S, Friedlin F, South B, Shen S, Samore M. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Medical Research Methodology 2010;10(1):70.
6. Benton A, Hill S, Ungar L, Chung A, Leonard C, Freeman C, et al. A system for de-identifying medical message board text. *BMC bioinformatics* 2011;12 Suppl 3:S2.
7. Deleger L, Molnar K, Savova G, Xia F, Lingren T, Li Q, et al. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. J Am Med Inform Assoc 2013;20(1):84-94.

8. Fernandes AC, Cloete D, Broadbent MT, Hayes RD, Chang CK, Jackson RG, et al. Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records. *BMC Med Inform Decis Mak* 2013;13:71.
9. Ferrández O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. BoB, a best-of-breed automated text de-identification system for VHA clinical documents. J Am Med Inform Assn 2013;20(1):77-83.
10. Beckwith BA, Mahaadevan R, Balis UJ, Kuo F. Development and evaluation of an open source software tool for deidentification of pathology reports. BMC Med Inform Decis Mak 2006;6:12.
11. Berman JJ. Concept-match medical data scrubbing. How pathology text can be used in research. Arch Pathol Lab Med 2003;127(6):680-686.
12. Gardner J, Xiong L. HIDE: An integrated system for health information de-identification. Proceedings of the 2008 21st IEEE International Symposium on Computer-Based Medical Systems 2008:254-259.
13. Neamatullah I. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak 2008;8:32.
14. Wellner B, Huyck M, Mardis S, Aberdeen J, Morgan A, Peshkin L, et al. Rapidly retargetable approaches to de-identification in medical records. J Am Med Inform Assn 2007;14(5):564–573.
15. Uzuner Ö, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assn 2007;14(5):550-563.
16. Kayaalp M, Browne AC, Callaghan FM, Dodd ZA, Divita G, Ozturk S, et al. The pattern of name tokens in narrative clinical text and a comparison of five systems for redacting them. J Am Med Inform Assn 2013.
17. Divita G, Browne AC, Loane R. dTagger: a POS tagger. *AMIA Annu Symp Proc* 2006:200-3.
18. NTT system description for the WMT 2006 shared task. Workshop on Statistical Machine Translation; 2006; New York, NY. Association for Computational Linguistics.
19. Efron B. Better bootstrap confidence interval. Journal of the American Statistical Association 1987;82(397):171-185.
20. Davison AC, Hinkley DV. Bootstrap methods and their application: Cambridge University Press, 1997.
21. Pratt JW. Remarks on zero and ties in the Wilcoxon signed rank procedures. Journal of the American Statistical Association 1959;54(287):655-667.
22. Hothorn T, Hornik K, van de Wiel MA, Zeileis A. Implementing a class of permutation test: the coin package. Journal of Statistical Software 2008;28(8):1–23.
23. Kang YS, Kayaalp M. Extracting laboratory test information from biomedical text. Journal of Pathology Informatics 2013;4(1):23-35. URL: http://www.jpathinformatics.org/ text.asp?2013/4/1/23/117450. Accessed in 9/3/2013.
24. Wikipedia. Area code 406: Wikipedia, 2013. URL: http://en.wikipedia.org/wiki/Area_ code_406. Accessed in 8/20/2013.
25. Office of Civil Rights. Guidance regarding methods for de-identification of protected health information in accordance with health insurance portability and accountability act (HIPAA) privacy rule, 2012.