The Challenges of Creating a Gold Standard for De-identification Research

Allen C. Browne, MS, Mehmet Kayaalp, MD, PhD,
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

We created a Gold Standard corpus comprised over 20,000 records of annotated narrative clinical reports for use in the training and evaluation of NLM Scrubber, a de-identification software system for medical records. Our experience with designing the corpus demonstrated the conceptual complexity of the task.

Introduction

At NLM we have developed a de-identification software system for medical records called NLM Scrubber intended to automatically de-identify clinical reports in compliance with the Privacy Rule of Health Insurance Portability and Accountability Act (HIPAA).[1] In the course of developing the scrubber, we needed to create a manually annotated corpus of medical records to serve as a "Gold Standard" for testing and evaluation. Annotation is needed to demark identifiers that should be found and scrubbed by the de-identification system and to provide enough information to facilitate both evaluation and further development. The nature of the markup in this corpus determines the kind of evaluation that can be undertaken. Ultimately, our corpus consisted of a set of 21,849 tagged clinical narrative reports. In each report, the identifiers meeting the HIPPA requirements for PII (personally identifying information) had to be hand-tagged. We had to make three types of decisions in the process of developing the tag-set. First, the HIPAA rules had to be interpreted and applied to the actual items found in the records. Second, we needed to identify the items themselves as well as their boundaries and third the internal structure of the items needed to be considered. This paper discusses our approach used in this process, outlines our conclusions and discusses alternatives.

Applying the HIPPA Rule

The HIPPA Privacy Rule requires that clinical documents be stripped of personally identifying information before they can be released to researchers and others. HIPAA Privacy Rule describes 18 identifiers that should be scrubbed in the de-identification of medical records.

Some of these descriptions seem quite specific; For example, "[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." But even in this seemingly well-defined rule, there is room for interpretation. Do partial dates without mention of the year count as dates e.g. "July" or "July 23rd"; do special days like "Christmas" or "New Year's" count as dates?

The Boundaries and Structure of identifiers

Identifiers do not always appear in discrete packages with clear borders. Are titles like "Mrs.", "Dr.", "Col." or "Adm." part of the name? What about name suffixes like "Jr." or "III" or titles like "MD", "Ph.D." or "Esq."? Some of them, such as "Mrs.", do not have much identifying information value; whereas, others such as "Adm." May have because of their occurrence in the population. Identifiers can also be conjoined in ways that obscure their structure. "The 5th, 6th and 18th of June 1965" seems to contain three full dates but it also contains lexical material that is not really part of any date, "and" in this case. The question/process of delimiting items is further complicated by the internal structure of those items.

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.[2]

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.

Are single letter initials inside a full name significant, e.g. "Q" in "John Q. Public"? In other words, if the scrubber only misses a middle initial, would the middle initial reveal the identity of the person? If not, should we tag or not tag the middle initials?

Methods

As described in Kayaalp, et. al. 2014,[2] we selected a random sample of patient records from the NIH Clinical Center using a method that prevented duplicate records. The selection involved randomly choosing 7,571 patients, collecting all of their records and removing duplicate records. A linguist and a registered nurse on our research team used VTT (Visual Text Tagger), a freely available text tagging system developed at NLM[3] to annotate PII in each record.[3,4] VTT uses a stand-off method to annotate texts so that both the original text and its formatting are preserved.[5] VTT facilitates tagging by allowing the human tagger to select an area of text by smearing the cursor over it and then choose a tag listed in from a drop down menu. It also provides a visual display of the tagged document representing each tag in a distinct visual format. VTT stores these annotated documents in a pure-ASCII machine readable format. In Figure 1 below, a mock-up version of the sort of records that make up our corpus is shown as displayed in VTT.

At the beginning of the process the human annotators tagged overlapping sets of records and came to a consensus on the results. The annotators conferred on specific questions as they worked through different sets of records. Then different sets of records were assigned to each annotator. Organized by patient, all the records of a particular patient would be completed by the same annotator in succession.

We tagged personal names, dates, addresses, ages and alphanumeric identifiers.

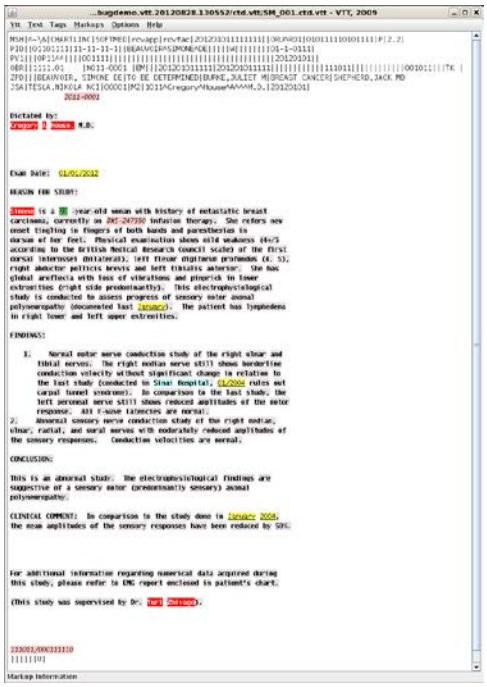

Figure 1

Names

HIPAA specifies that names must be redacted when they refer to patients and their relatives, employers and household members. Names of providers are not considered personally identifying, but almost all de-identifying programs scrub all personal names (care providers as well as patients), and it is probably the safest course because it can be difficult to distinguish between the two. We assigned the tag 'Name' to all personal names in our corpus. VTT allows tags to have subtags further refining the meaning of the tag. This distinction between identifiers related to patients and those not related to patients carries through a number of different tags. A 'Patient' subtag added to the 'Name' tag was used to indicate the names of patients as opposed to the names of hospital personnel and other people mentioned in the records. Names of relatives and members of the patient's household also received the 'Patient' subtag. Names were tagged as whole; that is the entire string "John Q. Public" would be given a single 'Name' tag. Suffixes like "III" or "Jr." were considered part of the personal name and tagged accordingly, but titles like "Mr.", "Dr.", or "Col.", were not unless their occurrences were rare. Initials seen in a longer name string were included in the name. Single first-names and last-names standing alone, e.g. 'John' by itself and "Public" in "Mr. Public", also received the 'Name' tag. Initials standing alone in the text, e.g. 'JQP' or 'JP', were marked 'PNinit' and separated into patient and non-patient using the 'Patient' subtag. While OCR guidelines states that initials are considered names,[6] we separated them because we felt initials were different and perhaps much less identifying than spelled out names. The vast majority of initials that we encountered denoted providers or transcriptionists who used them to sign the record. Names that were neither providers nor patients, for example names appearing as citations to the literature like "Greulich and Pyle", influential authors of articles regarding bone age, or the name of the current president when evaluating a patient's orientation were not tagged at all. These decisions introduced a difficulty in evaluation in those names like 'Greulich' and were counted as false positives when the system labeled them as names.

Addresses

Whole address strings such as "905 Maple Street, Apartment 2, Littletown, Minnesota, 55021" were tagged 'Address'. Cities standing alone like "Baltimore" were tagged as addresses but states and countries, for example "Maryland" and "Argentina", were not in keeping with guidelines of the Privacy

Rule which say that units smaller than a state should be redacted, although Baltimore has a population of well over 20,000, the size limit for Zip-Codes. D.C. was considered a state for this purpose. State-like subdivisions of countries other than the U.S., like "Alberta", were treated as states. Specific locations within the hospital received the 'Location' tag. For example, "OP9" or "3-Northeast" or "Day Hospital" were considered locations.

In a later iteration, we divided the Address tag into 8 tags identifying the street address, unit number, city, state, country, ZIP-code, county, and kept the old address tag as a catch-all. This revision reflected the realization that not all errors in redacting addresses were alike in seriousness. By tagging Address strings whole we made it difficult at the time of evaluation to recognize that redacting "Maryland" and not "Baltimore" from the string "Baltimore, Maryland" would result a violation of HIPAA but redacting "Baltimore" and not "Maryland" would not.

Street address includes strings like "904 South Madison Street" as well as building names and numbers "The Dakota" and "Building 9". The unit number captures apartment and suite numbers. States include U.S. states and their equivalents in other countries, e.g. "British Columbia". ZIP-codes have not been observed to stand alone in our documents. They are individually tagged as part of a larger address. To clarify, a lengthy address like "907 Madison Street, suite #5, Silver Spring, Maryland, 20190, U.S.A" represents 6 entities. A street address "907 Madison Street", a unit number, "Suite #5", a city, "Silver Spring", a ZIP-code "20910", a state "Maryland" and a country "U.S.A". Only the states and the countries need not be redacted under HIPAA.

This breakdown of larger address tags allows us to improve our evaluation by facilitating a more granular understanding of how identifying partially redacted addresses might be. It opens up a better method of counting errors than either a binary question of whether the address was (completely) redacted or a simple count of how many tokens or what percent of the address was redacted. The best evaluation might count the number of identifying parts that are redacted or better yet we should consider an evaluation of how identifying the unredacted parts are. For example "907", "Suite #5" or even "Madison" alone without the rest of the address could hardly be considered identifying.

Alphanumeric Identifiers

Alphanumeric identifiers were defined as strings of letters and/or numbers used as identifiers, excluding those identifiers that are part of a personal name, address, date, and age. We divided them into three different types: communication identifiers, protocol numbers and other Alphanumeric Identifiers, receiving 'Comm', 'Prot', or 'Alphanum' tags respectively. HIPAA calls for all of these identifiers to be redacted. A 'Comm' tag was assigned to Communication identifiers included numbers such as telephone numbers, email addresses, URLs and the like. Protocol numbers were common in our corpus and have a fairly typical form. The remaining numbers comprised the Other Alphanumeric identifiers and came from a range of types, including sample numbers, blood unit numbers, radiologic ids and lab test numbers. Communication number may or may not pertain to a patient so 'Comm' tags can take the patient subtag depending on whether they pertain to the patient. The telephone number of a referring physician would be marked 'Comm' but without a 'Patient' subtag. Protocol numbers and other alphanumeric identifiers were all considered patient related and subtagged 'Patient'.

Ages

We used three tags 'Age-PII' and 'Age-NPII' and 'Age-fract' to mark ages found in the corpus. Age-PII identified ages 90 years and over, since HIPAA specifically requires that ages over 89 be redacted. Age-NPII was used to mark ages in years, less than 90 which are not PII. Ages less than a year e.g. "3 days", "2 ½", "fifth week of life" belong to a special case, because they were not singled out by HIPPA as PII but they certainly could be much more identifying than an age in whole years. In the case of 'Age-PII' and 'Age-NPII' only the numeric part of the age was tagged, e.g. in "patient is 56 years old", we only tagged "56". In the case of age ranges like "3-5 years" only "3" and "5" were tagged. Ages given as decades, e.g. "in his 60's", were not tagged because they represented so large an age span that they were not considered identifying. In the case of fractional ages, both the number and the unit of measure were included in the tag, for example "thirteen months" and 'three weeks' would be tagged 'Age-fract'. By

keeping the unit of measure we allowed the de-identification system to round these ages to a year or redact them completely. All three age tags were used to mark only the ages of patients, their relatives or household members. No provider ages were seen in our corpus. Gestational ages and bone ages were not tagged 'Age'.

Dates

Date strings were initially tagged whole in our first iteration. "Wednesday, June 14, 1996" was tagged as a date as were free standing date parts. Months like "February", days like "Thursday" and years like '1998' when standing alone in text were tagged 'Date'. Decades like "The 60's" and plural days, "Fridays" were not tagged because such long or repetitive dates were not considered identifying. By default all dates were considered to be relevant to patients. In date ranges like "June 3 – July 15" both the beginning and end points are tagged as dates, separately. An exception appears when part of the range is unable to stand on its own, e.g. "2005-6". In this case the whole string was tagged. We considered special days like "Christmas" or "Mother's Day" to be dates and tagged them as such since they are equivalent to a date like 'December 25". Although HIPAA requires redaction of all elements of Dates except the year, strings like "September 23rd" standing alone don't indicate a specific date unless they are in the context of a year. But since HIPAA allows years not to be redacted we should assume that there might be a year in context depending on how the de-identification run is configured. Days of the month standing alone are another matter since "the third" or "the third of the month" does not specify a particular date and would only do so in conjunction with a month name and a year. Since month names must be redacted under HIPAA, weekday names would not be particularly identifying.

Similar to our treatment of Address, in a later iteration of our tagging effort we broke Dates into their parts again facilitating a more granular evaluation. By tagging months, days and years separately in a long date string like "February 27th, 1991" we can not only take account of the fact that the substring "1991" need not be redacted at all, but we can also consider that "27th" without the month does little to identify that actual date even in the presence of the year.

This more granular approach to tagging also facilitates treatment of conjoined and otherwise obscured items. For example the date string "the 5th, 6th, and 18th of May" presents several difficulties for evaluation. The tokens "the", "and" and "of", though parts of the date string, are not really parts of the three dates represented here and might be ignored during evaluation. By tagging "May" as a month and "6th" as a day we will be able to recognize that "May" alone is identifying in a way that "6th" alone is not, even in the context of a year.

Revision of other tags

Although we have not yet moved to a more granular treatment of other identifiers, it would clearly help our evaluation to do so. In the course of evaluation we decided not to count single initials that are part of a name like the "H" in "William H. Macy" as a name token. That is, we did not consider it a false negative if it was left unredacted when the rest of the name was redacted. Although this situation seldom arose imprecise because the scrubber was generally able to recognize middle initials from their position between two name tokens. That decision points to a future re-tagging of the corpus to reflect the parts of full names. Similarly, telephone numbers might be sub-divided into the area code prefix and number. Area codes and prefixes have a geographical association and might be considered more identifying than the 4 digit number itself. HIPAA already contemplates the internal structure of ZIP-codes specifying that the initial three digits of the ZIP code need not be redacted if "The geographic unit formed by combining all the ZIP codes with the same three initial digits contains more than 20,000 or fewer people." Something similar might be applied to telephone area codes or prefixes.

Results

We annotated a total of 21,849 records, representing 7,571 patients. Of those records, 3093 were used for evaluation in our de-identification study[2] and 1,140 were used for training. In addition to the two iterations of tagging, errors found in the course of evaluation were fed back into the gold standard after review by the taggers.

Discussion

The creation of our gold standard represented a number of challenges as described above including the lack of clear definitions of redactable items. One of the main lessons to come out of the effort was the realization that a finer grained analysis of strings representing PII facilitates a better understanding of evaluation results and points to a better method of evaluation. Counting whole strings as either properly redacted or not does not take into account which parts of the string might be left unredacted. Using token counts in the calculation of sensitivity and specificity also has inherent drawbacks, especially when a singly revealed token is a part of a multi-token identifier such as “September 11, 2001.”The potential of a particular token to identify the patient is less clear than the potential of properly tagged parts of the whole string.

Another consideration not explored above is the possibility of tagging more than the PII in a file. We found our evaluation hindered by a lack of knowledge about which tokens could be removed without loss of clinically pertinent information. Examination of actual redacted records shows that precision based on the number of false positive tokens overestimates the loss of information in actual records. Natural language including the sublanguage of clinical records is sufficiently redundant so that the loss of tokens often does not result in a significant loss of readability. Some sections of the record will inevitably contain information not relevant to a particular medical task and loss of that information would not damage the usefulness of the record. We are exploring a methodology to identify and subsequently tag clinical information so as to rationalize future evaluation of precision in our de-identification effort. The challenge in this task would be similar; that is, how should we categorize clinical information and label it with finer granularity so that we can fairly measure the loss of clinical information?

Funding

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

Competing Interests

The second 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. 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.
2. Kayaalp M, Browne AC, Dodd ZA, Sagan P, McDonald CJ. De-identification of Address, Date, and Alphanumeric Identifiers in Narrative Clinical Reports. AMIA Fall Symposium, 2014.
3. Lu, Chris J; Divita Guy; Browne, Allen C. “Development of Visual Tagging Tool”. AMIA 2010 Annual Symposium, Washington, DC, November 13-17, 2010, p. 1156
4. National Library of Medicine. Visual tagging tool, 2010. URL: http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/vtt/current/web/index.html. Accessed in 8/20/2013
5. Kayaalp, M. Separation of Data, Interpreters and Likelihood. Report number: LHNCB-TR-2007-001, Affiliation: Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, 2007.
6. 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.