## TITLE PAGE

**Title:** Improved de-identification of physician notes through integrative modeling of both identifying and non-identifying medical text

**Corresponding Author:** Andrew McMurry, MS
Countway Library of Medicine
Harvard Medical School
10 Shattuck St 4th Floor
Boston, Ma 02115

Email: McMurry.Andy@Gmail.com
Phone: 617-448-1288

**Co-Authors:**

1. Andrew J McMurry, MS *123
2. Britt Fitch *1
3. Guergana Savova, PhD 2
4. Isaac S Kohane, MD, PhD 124
5. Ben Y Reis, PhD 12

* Shared first authorship. Please acknowledge that the starred authors contributed equally to this research and would like to be acknowledged as dual first authors.
[1]Harvard Medical School Center for Biomedical Informatics, Boston, Massachusetts, USA
[2]Children's Hospital Informatics Program at the Harvard-MIT division of Health Sciences and Technology, Boston, Massachusetts, USA
[3]Boston University, Center for Advanced Genomic Technology, Boston, Massachusetts, USA
[4]Research Computing, Partners Healthcare System, Information Technology, Charlestown, Massachusetts, USA

**Keywords:**

1. Natural Language Processing (L01.224.065.580)
2. Confidentiality (I01.880.604.473.650.500)
3. Pattern Recognition, Automated (L01.725)
4. Electronic Health Records (E05.318.308.940.968.625.500)

## ABSTRACT

**Background:** Physician notes routinely recorded during patient care represent a vast and underutilized resource for human studies on a population scale. Their use in research is limited primarily by the need to remove all identifying information before they can be included in a study, a process that can be very resource-intensive when performed manually. While automated de-identification allows for greater numbers of physician notes to be included in a study, training such de-identification models requires access to large amounts of private information that in turn must also be annotated manually. This study seeks to create an automated method for de-identifying physician notes that does not require large amounts of private information: In addition to training a model to recognize Protected Health Information (PHI) from private physician notes, we reverse the problem and train a model to recognize non-PHI words that appear in public medical texts.

**Methods:** Multiple collections of public and private medical texts were analyzed to train a decision tree classification model. Publicly available medical vocabularies and journal publications were used to learn the negative examples -- the probability distributions of non-PHI words. Private physician notes were used to learn the positive examples -- probability distributions of PHI words. The positive examples were augmented with census names and word patterns published in other studies. In total, 26 word features were analyzed to model non-PHI words and each type of PHI.

**Results:** The model successfully de-identified texts from 220 discharge summaries with 98% sensitivity with 89% specificity. Furthermore, the model removed 99.4% of unique IDs that refer to the patient and not the hospital or treatment. The results exceed the previously approved criteria established by four Institutional Review Boards.

**Conclusions:** The results indicate that distributional differences between private and public medical text can be used to accurately classify PHI. The data and algorithms reported here are made freely available for evaluation and improvement.

## I. BACKGROUND

Physician's notes contain information that may never be recorded in a coded format in the patient health record[1-3], such as family history[4], smoking history[5-7], and descriptions of lab results[8, 9]. Nonetheless, the "uncoded" information buried in physician notes is so valuable that numerous attempts have been made towards indexing and sharing notes for research use. However, since physician notes can contain patient names, home addresses, social security numbers, and other types of Protected Health Information (PHI)[10], vast quantities of doctors' notes have gone largely unused for medical research studies. Methods to simultaneously protect patient privacy and increase research utility are needed – as the number of electronic health record systems increases and with it the opportunity to study larger numbers of patients[11, 12].

Existing methods for de-identifying medical texts range from simple rule-based systems to sophisticated machine learning algorithms [13, 14]. The majority of currently implemented methods are rule-based systems that match patterns and dictionaries of expressions that frequently contain PHI[15]. The advantage of rule-based systems is that experts can quickly define rules and iteratively fine-tune them to achieve higher accuracy. While rule-based systems have shown high sensitivity in some settings[15], they often have the disadvantage of hard coding rules to a specific note format or physician writing style, resulting in poor performance in other contexts. Adjusting existing rule systems for use at other medical centers is often too costly, limiting broad use across institutions. This problem is well recognized,[13, 14] and has prompted efforts using an alternative, machine learning approach. Rather than using the expert to author rules, the rules for PHI removal are "learned" by training an algorithm using human annotated examples (i.e. a supervised learning task). For example, competitors in the i2b2 de-identification challenge[13] were asked to train or tune their algorithms on one set of human annotated notes and then validate their best model on a separate set of annotated notes. Generally, the highest scoring algorithms used machine learning methods such as conditional random fields,[16] decision trees,[17] and support vector machines[18].

The work reported here was trained and validated on the same i2b2 challenge datasets, which allows for comparison to prior work. Our algorithm performed favorably with regards to sensitivity, albeit with lower specificity (see results). The primary difference between our method and other methods is the extensive use of publicly available medical texts. We show that publicly available medical texts provide an informative background distribution of sharable medical words, a property that is largely underutilized in patient privacy research.

## II. METHODS

Instead of trying to recognize PHI words in physician notes, we reversed the problem towards recognizing non-PHI words. We asked, “what are the chances that a word or phrase would appear in a medical journal or medical dictionary? What are the lexical properties of PHI words? To what extent can we use publicly available data to recognize data that is private and confidential?”

While human annotated datasets of PHI are few in number and difficult to obtain, examples of non-PHI medical text are broadly available and generally underutilized for de-identification. By definition, medical journal publications provide the distributional evidence for words that are not PHI. Of course, some medical words will end up being proper names but the public corpora provide a heuristic measure of likelihood that we exploit as described below. In this context, relatively fewer human annotated examples are treated as approximations of the distributional properties of PHI. Lexical comparisons between PHI words and non-PHI words reveal that all PHI words are nouns and numbers – whereas all verbs and adjectives are probably ok to share -- especially medically relevant verbs and adjectives that are of more relevant to research studies. Publicly available lists of suspicious words and expert rules are also incorporated into this algorithm, such as US census data and regular expressions found in or around PHI terms. We combine the discrimination power of these complementary perspectives to achieve improved de-identification performance. As an additional safeguard, notes can be indexed[19] and later searched using coded medical concepts, thereby reducing the number of full-text reports that need to be shared in early phases of research[20].

### Design Principles

The Scrubber was designed with the following general observations about physician notes and other types of medical text: (1) words that occur only in physician notes have increased risk for PHI, especially nouns and numbers which are the only types of PHI words; (2) words that occur in medical publications are not likely to refer to any specific patient; (3) words and phrases in medical vocabularies also do not refer to individually named patients; (4) words shared in many publically available medical text sources are very unlikely to contain PHI (FIGURE 1).

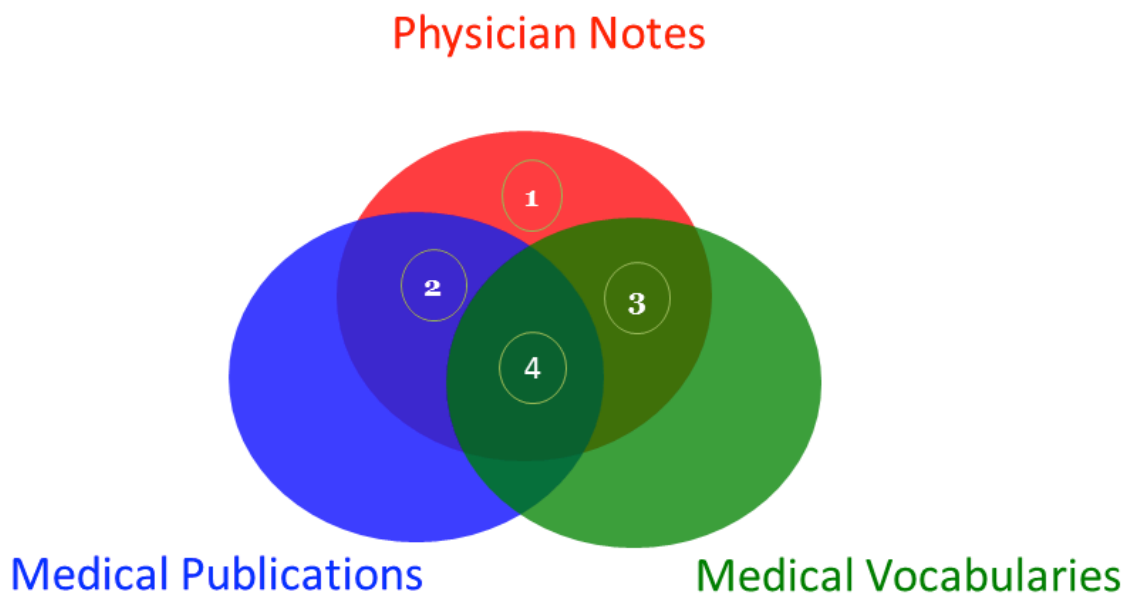


**Figure 1: Observations of physician notes with other types of medical text**

1. Words that occur only in physician notes have increased risk for PHI, especially nouns and numbers.
2. Words that occur frequently in medical publications are not likely to contain PHI.
3. Words and phrases that occur frequently in medical vocabularies are not likely to contain PHI.
4. Words shared in all three medical text sources are very unlikely to contain PHI.

## Different Types of PHI

The risk to patient confidentiality differs among the 8 major types of HIPAA defined PHI elements (TABLE 1). Accordingly, the primary goal of this study was to remove sensitive text that refers uniquely to a *single* patient including patient names, IDs such as medical record numbers, phone numbers, and home addresses. Fortunately, single patient identifiers are rarely necessary in patient research. As a secondary objective, this study sought to classify all types of PHI defined by HIPAA. This includes features that may refer to *many* patients, such as a hospital name, patient age, date of service, or doctor. These features are useful for studies of disease over time and should not necessarily be scrubbed if Limited Data Set[21] access is permitted by the hospital privacy board.

| PHI Type | Minimum Disclosure | Risk |
|---|---|---|
| Hospital | LDS | Minimal |
| Age | LDS | Minimal |
| Date | LDS | Minimal |
| Doctor | LDS | Minimal |
| Location | LDS | Varies |
| Patient | Identified | High |
| ID | Identified | High |
| Phone | Identified | High |

**Table 1: Types of PHI and their risk to patient confidentiality**
Minimum Disclosure refers to the level of permission typically required by an IRB. Limited Data Sets (LDS) contain features about a patient that refer to more than one patient. Identified level access almost always refers to a single patient, such as a patient name, medical record number, or phone number.

We anticipated that each type of PHI would have a unique set of association rules. For example, patient names are nouns whereas medical record numbers are numbers. Learning different association rules[22] for each type of PHI has the added benefit that additional weight can be placed on highest risk elements, such as the patient name or home address. All types of PHI types are generally represented as nouns and numbers with low term frequencies, low occurrence in medical controlled vocabularies, and non-zero regex matches of some type. Non-PHI words generally have higher term frequencies, higher occurrence in medical vocabularies and near zero matches in regular expressions of any type.

### Feature Set Construction

The Scrubber constructs a feature set in four steps: lexical, ontological, patterned, and statistical (FIGURE 2). First, the document instance is split into fragments and analyzed for part of speech and capitalization usage[23]. Second, each fragment is matched against dictionaries of controlled medical vocabularies[24] and US census data[25]. Third, regular expressions are applied for the eight categories of PHI. Lastly, term frequencies are assigned to each token in the document.

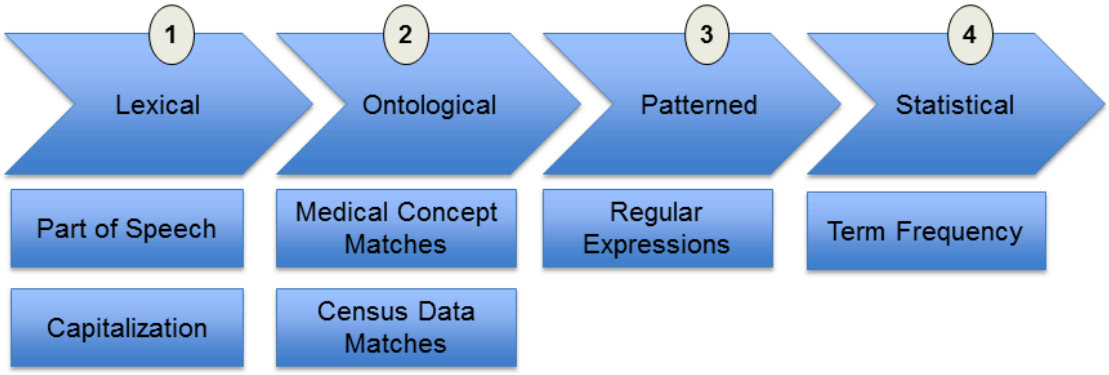


**Figure 2: Phases of the Scrubber Annotation Pipeline**

1. Lexical Phase: split document into sentences, determine part of speech for each word.
2. Ontological Phase: search for each word/phrase in the known medical dictionaries and census names list.
3. Patterned Phase: match regular expressions for each category of PHI
4. Statistical Phase: determine term frequency for each token both with and without regard to part of speech.

The Scrubber software leverages several other Open Source packages. The data processing pipeline is provided by Apache UIMA project[26], an engineering framework commonly used in Natural Language Processing[27]. Of note, UIMA does not provide any pre-built components for text processing, it provides the main "scaffolding" and flow between user developed components. In the lexical phase, cTAKES splits each document into sentences[28] and determines the part of speech for each token. cTAKES is especially appropriate because it has been extensively trained on medical documents[23]. In the ontological phase, each fragment is compared against phrases in publicly available sources, such as ICD9 diagnoses and US census names. In the term frequency phase, the count of each token is retrieved from a pre-processed corpus of open access medical publications.

The annotation pipeline produces a high dimensional feature set that is very sparse, making the classification step more difficult. There are a number of ways to reduce dimensionality and increase feature set density, such as clustering similar features[29-31], removing features with low information content[32], reducing the number of class labels[22], and aggregating feature counts. Aggregating feature counts provided adequate feature density and reduced the dimensionality without discarding features that could be informative. Specifically, features were aggregated by source and type with respect to processing phase: lexical, ontological, patterned, and statistical (TABLE 2).

| Lexical | Ontological | Patterned | Statistical |
|---|---|---|---|
| Part of Speech(POS) | # matches in HL7 2.5 | # matches of PHONE | TF (Token) |
| POS (Binned) | # matches in HL7 3.0 | # matches of DATE | TF (Token, POS) |
| Capitalization | # matches in ICD9 CM | # matches of AGE | |
| | # matches in ICD10 CM | # matches of ID | |
| | # matches in ICD10 PCS | # matches of PATIENT | |
| | # matches in LOINC | # matches of DOCTOR | |
| | # matches in MESH | # matches of LOCATION | |
| | # matches in RXNORM | # matches of HOSPITAL | |
| | # matches in SNOMED | # matches of Hospital File | |
| | # matches in COSTAR | # matches of Private File | |
| | # matches in U.S. Census Names | | |

**Table 2: Complete list of all 26 features annotated by the NLP pipeline.** In the lexical phase, part of speech and capitalization usage is annotated for each word token. In the ontological phase, each word is compared to a list of standard medical ontologies – dictionaries of medical conditions and services. In the patterned phase, word tokens are compared against suspicious patterns of HIPAA identifiers. In the statistical phase, each word token is annotated with the frequency of appearance in public and private medical texts.

### Classification

The feature set is then processed through Weka [33] using a J48 decision tree[34] classification algorithm, a popular open source implementation of the C4.5 decision tree algorithm. J48 was chosen for several reasons. First, decision trees do not require “binning” value ranges to be effective[35]. This was required because the correct value ranges were not known prior to classifier training. Second, decision trees can build a model for multiple class types. This is important because different types of PHI have different rules associated with them. For example, patient names are nouns whereas medical record numbers are numbers. A binary classifier would ignore these characteristic differences across PHI types and likely cause more errors.

### Training

The primary data used for training and testing was the I2B2 de-id challenge data.[13] This data consists of 669 training cases and 220 testing cases. The cases are a fully annotated gold standard set of discharge summaries. To calculate frequencies of word occurrences, we randomly selected 10,000 publicly

available peer reviewed medical publications. This was necessary as many valid word tokens appear only once or not at all in any random selection of physician notes. Using more than 10,000 publications for training did not alter performance, and was computationally feasible using inexpensive commodity hardware.

On average there were 520 words (tokens) per case, and an average of 39 PHI words per case. As expected, most word tokens were not patient identifiers (PHI) -- the ratio of PHI words to non-PHI words was 1:15. Training a classifier using all of the available training instances would highly favor non-PHI classifications[36]. To address this issue, the training set was compiled using all of the PHI words and an equally sized random selection of non-PHI words.

## III. RESULTS

### Summary

The training model was applied to an independent validation corpus of 220 discharge summaries from the i2b2 de-id challenge. Most importantly, this method was 99.4% sensitive to *unique identifiers* such as patient name and home address. Identifiers with lower risk – such as hospital name or dates of treatment– were removed 97.7% of the time. This method was 98% sensitive to removal of all PHI types and 89% specific (TABLE 3). In this context, sensitivity is the percentage of patient identifiers that were removed (Equation 1). Specificity is the percentage of sharable words that remain in the processed document (Equation 2). The automated performance matches or exceeds that of two human evaluators[37] and preserves the readability of the original text [37].

### Equation 1: Sensitivity

$$\text{sensitivity} = \frac{\text{number of PHI words removed}}{\text{number of PHI words removed} + \text{number of PHI words missed}}$$

### Equation 2: Specificity

$$\text{specificity} = \frac{\text{number of sharable words remaining}}{\text{number of sharable words remaining} + \text{number of sharable words removed}}$$

| | Sensitivity | | | | | | | | Sens | Spec |
|---|---|---|---|---|---|---|---|---|---|---|
| | ID | Hospital | Doctor | Date | Patient | Location | Phone | Age | Any PHI | Not PHI |
| **All** | 99.9% | 91.8% | 99.3% | 99.9% | 98.0% | 91.7% | 97.0% | 100.0% | 98.1% | 89.3% |
| **Part of Speech** | 95.3% | 90.8% | 90.3% | 99.4% | 91.5% | 98.2% | 100.0% | 0.0% | 95.0% | 55.7% |
| **UMLS** | 99.9% | 88.2% | 91.8% | 99.6% | 84.6% | 82.5% | 100.0% | 100.0% | 95.0% | 46.1% |
| **Regex** | 86.6% | 54.1% | 78.8% | 98.1% | 29.4% | 7.8% | 92.1% | 0.0% | 79.7% | 95.8% |
| **TF** | 99.9% | 97.7% | 93.0% | 100.0% | 84.6% | 91.7% | 100.0% | 100.0% | 97.2% | 73.9% |
| | **9 Way classification** | | | | | | | | **Binary Class** | |

**Table 3: Classifier results.** The Multi-class classifier shown on left achieves differing sensitivity depending on the PHI type considered and the group of features being evaluated. Specificity remains constant in both the multi-classifier case and binary classifier cases as there is no penalty for confusing *between* PHI categories. Green denotes excellent classifier performance >95%. Black denotes mediocre performance >85% <95%. Red denotes unacceptable performance, <85%.

## Misclassifications

The words “of”, “hospital”, and “clinic” were overwhelmingly the most commonly missed PHI words. These common words account for 122 of 173 partial misses, and pose little to no risk to patient privacy.

We performed a manual review of each misclassification (TABLE 4, supplemental material) and determined that no unique identifiers were left fully intact. Partial redactions – such as properly removing the patient last name but missing the patient first name were rare (13 word tokens in 12 cases). Lower risk identifiers such as hospital name and date of treatment were also rare. Only two dates and 2 hospital names were left fully intact.

| PHI Type | # Identified | # LDS | # Missed |
|---|---|---|---|
| Hospital | 0 | 134 | 134 (1633) |
| Age | 0 | 0 | 0 (3) |
| Date | 0 | 4 | 4 (3673) |
| Doctor | 0 | 15 | 15 (2097) |
| Location | 4 | 14 | 18 (217) |
| Patient | 8 | 0 | 8 (402) |
| ID | 1 | 1 | 2 (1455) |
| Phone | 0 | 5 | 5 (165) |
| Totals | 13 | 173 | 186 (9645) |

**Table 4: Misclassifications**. The majority of misclassifications would be allowed with a Limited Data Set (LDS) agreement. The minority of misclassifications refer to a single patient, which would require Identified level data access.

### Part of Speech

Every type of PHI is a noun or number. Interestingly, this fact alone yielded better than 90% sensitivity for 7 out of 8 PHI types (TABLE 3). However, many naturally occurring words and medically relevant concepts can also appear as nouns and numbers. To distinguish PHI from nouns and numbers that are naturally occurring, a term frequency calculation was applied. Similarly, nouns and numbers with medical relevance were distinguished by their presence in one or more medical vocabularies.

### Term Frequencies

Medical publications do not refer to individually named patients. Even in medical case studies, the patient name, home address, phone number, and medical record number must be withheld in accordance with law. This guarantees that all high-risk PHI elements in Table 1 will not be present in the publication dataset. It was therefore not surprising to find that patient specific identifiers were not frequently reported in the text of medical publications. As a result, classification of PHI using only term frequency and part of speech yielded high scrubbing performance, with 97% sensitivity and 74% specificity.

As expected, a first or last name would sometimes match an author name in the publication text. However, since author names and references list were removed during preprocessing, the overlap in names was minimized. There are other

examples where patient identifiers can overlap with text in publications, for example when a patient lives on a street with the same name as a medical facility used in a published study. Nevertheless, patient identifiers are much less likely to appear in journal publications. To test and quantify this assumption, term frequencies were calculated across all word tokens in publication, training, and test datasets. Training and test datasets were split into groups of words containing PHI and not containing PHI. Histograms were then created, where the x-axis is the number of times a word appeared in all medical publications and the y-axis is the number of distinct words. A small percentage of common words created a skewed distribution, which was log normalized for visualization clarity. Figure 3 shows that PHI words are less frequently used in journal publications than non-PHI words. This is true with or without considering the part of speech for both the training and test datasets.

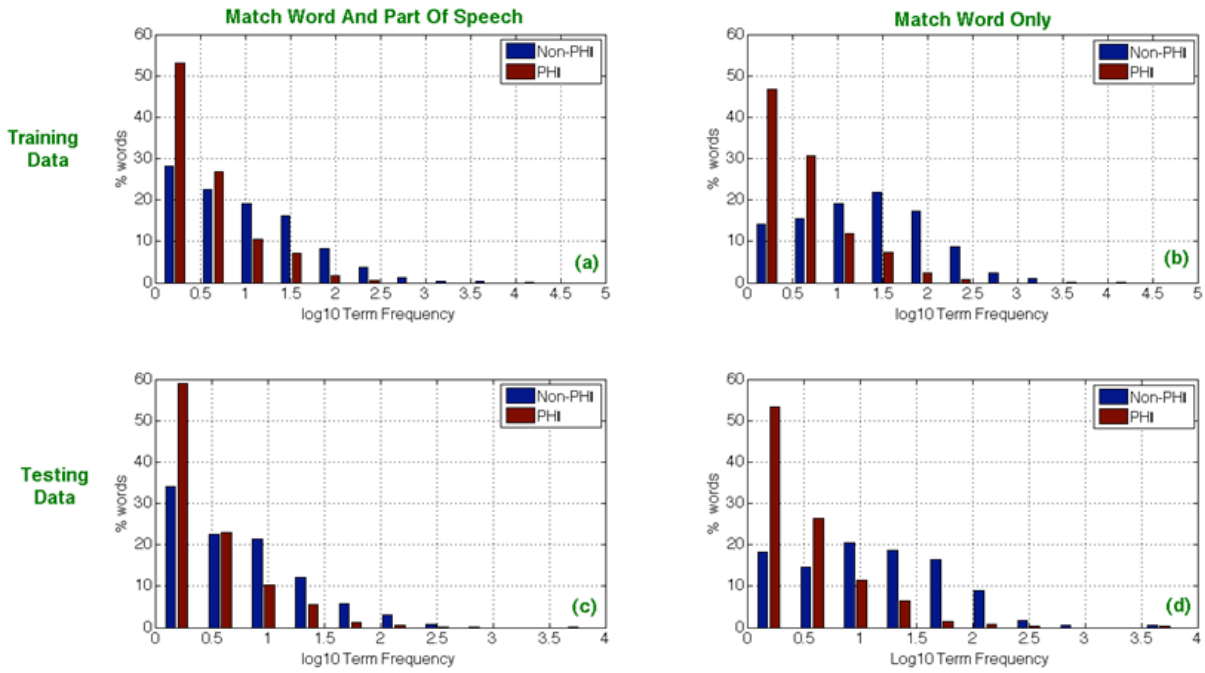


**Figure 3: Term Frequency Distributions in PHI and non-PHI word tokens**
In each of the four histograms, the log normalized term frequency (x-axis) is plotted against the percentage of word tokens. PHI words (red) are more common on the left hand side of each histogram, showing that PHI words tend to be rarer than non-phi words (blue). Top Figures (a) and (b) contain training data. Bottom Figures (c) and (d) contain testing data. Histograms for Training and Testing are characteristically similar. Term frequency histograms on the left (a) and (c) refer to words matched according to their part of speech. Term frequency histograms on the right (b) and (d) refer to raw word matches.

### Medical Vocabularies

Ten vocabularies in the Unified Medical Language System were selected in order to span a very wide range of demographic terms, diagnoses, lab tests, medication names, and procedures. Surprisingly, a decision tree trained only to distinguish PHI from medical concepts yielded very high sensitivity (95%), albeit with poor specificity (46%). This suggests that there is almost no overlap between medical concepts and patient identifiers. These findings provide evidence that automatic retrieval of coded medical concepts (autocoding) is also useful for de-identification. In this way, parallel autocoding and de-identification provides maximum research utility while minimizing the risk of patient disclosure.

### Regular Expressions

Regular Expressions yielded the highest specificity (96%) with the lowest sensitivity (80%) of any feature group tested in isolation (TABLE 3). This matches our experience using a previous version of the HMS Scrubber in new medical center settings without customization and without inspecting the pathology report header[15]. We expected the regular expressions to outperform all other feature groups with respect to dates, phone numbers, and ages but this was not the case. This either means that we used Beckwith's regular expression rules incorrectly or there are more ways to express these simple concepts than one might expect. Nevertheless, regular expressions slightly improved the overall classification specificity.

## IV. DISCUSSION

Can we use data that is publicly available to recognize and remove confidential information from physician notes? Can we accelerate the rate of sharing physician notes for research without compromising patient confidentiality? Can we achieve these goals while respecting the challenges and responsibilities among hospital privacy boards? These questions motivated the authors to compare public and private medical texts to learn the distributions and lexical properties of Protected Health Information. The results of this experiment show that publicly available medical texts are highly informative for PHI recognition, resulting in performance that is likely to be approved for research use among by hospital review boards. 99.4% of unique patient identifiers that pose highest confidentiality risk were removed by this method, and no complete patient names or ID numbers were missed. The remaining misclassifications were common words appearing in hospital names, which pose minimal risk to patient privacy. A useful byproduct of this de-identification process is that coded medical concepts[38] are also stored for later search[20] and retrieval[39]. This approach to de-identification both reduces unauthorized disclosures and increases authorized use, a position previously confirmed by numerous hospital privacy boards [15, 20, 39].

Comparing public and private text sources reveals interesting properties of PHI.

Words in physician notes that frequently appear in medical journal publications and concept dictionaries are highly unlikely to contain PHI. Conversely, words in physician notes that are nouns and numbers are more likely to contain PHI. It is interesting to speculate just how far publicly available text can be leveraged for de-identification tasks.

In a state of the art review of de-identification, Ozuner and Szolovits appropriately ask “how good is good enough?[13]” In this study, we sought to achieve performance levels that were already considered satisfactory by hospital privacy boards[15] with minimal investment. Numerous tradeoffs were made to achieve this goal. First, sensitivity was strongly favored over specificity, especially for patient names and ID numbers that have highest risk of disclosure. Second, we favored default configuration over hospital-specific human refinement. In our experience, site-specific modification of patient names lists and regular expressions can be laborious and can lead to “overscrubbing” information that is valuable for research. Third, we needed the algorithm to run on a single computer using commodity hardware, both to satisfy IRB concerns over data-duplication and reuse hardware already in place. Fourth, we wanted to make as few assumptions as possible about the training set to avoid unnecessary overfitting.

There were several limitations to this study. Term frequency calculations were performed for single word tokens. Increasing the term frequency to use two or more words might improve patient name recognition. For example, patients are more likely to have a first or last name in common with an author than a full name. Similarly, patient home addresses are highly unlikely to be found in published medical journals. However, ngram computation can quickly lead to exponential time complexity, requiring that unscrubbed notes be copied and shared on compute clusters for processing. This could create a paradox, as the very purpose of scrubbing is to increase sharing with minimal risk of disclosure. Simply, hospital privacy boards may disallow distributed processing for purposes of de-identification.

There is also the potential that we too have overfit our model to training examples and were fortunate enough to have the model validated in an independent sample. There are several cases where classifying PHI in new physician notes could be significantly less accurate. PHI words and phrases that frequently appear in medical publications and dictionaries are the most difficult to classify, although the number of times this occurs appears negligible. Incoherently written physician notes may be difficult to tag for part of speech, which would likely degrade classifier accuracy. Datasets that have different probability distributions and term frequency could also pose problems. In each of these potentially limiting examples, a new corpus would have to be characteristically different from the testing and training examples studied here.

We recommend that this de-identification method be used according to procedures that were previously acknowledged by four hospital IRBs[15, 20]. The recommended workflow is as follows. Physician notes are de-identified and autocoded such that the scrubbed report is saved in a secured database and searchable according to medical vocabularies. Search access is limited to authorized investigators affiliated with the institution hosting the data, and under no circumstances should the textual data be made available for public download. Searching for patient cohorts matching study criteria occurs in an anonymized manner, meaning that only counts are returned with the first level of access. After finding a cohort of interest, an investigator may apply for access to review the de-identified cases. By increasing the level of access commensurate with the needs of a study[20], the risk to patient disclosure is minimized while allowing many investigators the ability to query and browse the valuable collection medical notes.

The methods proposed here can be put to practical use today to help unlock the tremendous research potential of vast quantities of free-text physician notes accumulating in electronic medical record systems worldwide.

## V. ACKNOWLEDGEMENTS

The authors would like to acknowledge the National Cancer Institute for funding and evaluating this work for use in other medical centers. The authors would also like to thank Bruce Beckwith for insights and commentary in the authorship of this report. We would also like to thank the cTAKES team for providing extensible language processing tools.

Deidentified clinical records used in this research were provided by the i2b2 National Center for Biomedical Computing funded by U54LM008748 and were originally prepared for the Shared Tasks for Challenges in NLP for Clinical Data organized by Dr. Ozlem Uzuner, i2b2 and SUNY.
https://www.i2b2.org/NLP/DataSets/Main.php

Editors: please note that Britt Fitch and Andrew McMurry are sharing first authorship of this report.

### VI. REFERENCES

1. Uzuner, O., *Recognizing obesity and comorbidities in sparse data.* Journal of the American Medical Informatics Association : JAMIA, 2009. **16**(4): p. 561-70.

2. Liao, K.P., et al., *Electronic medical records for discovery research in rheumatoid arthritis.* Arthritis care & research, 2010. **62**(8): p. 1120-7.

3. Uzuner, O., I. Solti, and E. Cadag, *Extracting medication information from clinical text.* Journal of the American Medical Informatics Association : JAMIA, 2010. **17**(5): p. 514-8.

4. Goryachev, S., H. Kim, and Q. Zeng-Treitler, *Identification and extraction of family history information from clinical reports.* AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium, 2008: p. 247-51.

5. Heinze, D.T., et al., *Medical i2b2 NLP smoking challenge: the A-Life system architecture and methodology.* Journal of the American Medical Informatics Association : JAMIA, 2008. **15**(1): p. 40-3.

6. Zeng, Q.T., et al., *Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system.* BMC medical informatics and decision making, 2006. **6**: p. 30.

7. Savova, G.K., et al., *Mayo clinic NLP system for patient smoking status identification.* Journal of the American Medical Informatics Association : JAMIA, 2008. **15**(1): p. 25-8.

8. Patel, A.A., et al., *Availability and quality of paraffin blocks identified in pathology archives: a multi-institutional study by the Shared Pathology Informatics Network (SPIN).* BMC cancer, 2007. **7**: p. 37.

9. Hoshida, Y., et al., *Gene expression in fixed tissues and outcome in hepatocellular carcinoma.* The New England journal of medicine, 2008. **359**(19): p. 1995-2004.

10. Services, U.S.D.o.H.H. *Health Information Portability and Accountability act.* 1996; Available from: http://www.hhs.gov/ocr/privacy/.

11. Kohane, I.S., *Using electronic health records to drive discovery in disease genomics.* Nature reviews. Genetics, 2011. **12**(6): p. 417-28.

12. Kohane, I.S., S.E. Churchill, and S.N. Murphy, *A translational engine at the national scale: informatics for integrating biology and the bedside.* Journal of the American Medical Informatics Association : JAMIA, 2012. **19**(2): p. 181-5.

13. Uzuner, O., Y. Luo, and P. Szolovits, *Evaluating the state-of-the-art in automatic de-identification.* Journal of the American Medical Informatics Association : JAMIA, 2007. **14**(5): p. 550-63.

14. Meystre, S.M., et al., *Automatic de-identification of textual documents in the electronic health record: a review of recent research.* BMC medical research methodology, 2010. **10**: p. 70.

15. Beckwith, B.A., et al., *Development and evaluation of an open source software tool for deidentification of pathology reports.* BMC medical informatics and decision making, 2006. **6**: p. 12.

16. Aberdeen, J., et al., *The MITRE Identification Scrubber Toolkit: design, training, and assessment.* International journal of medical informatics, 2010. **79**(12): p. 849-59.

17. Szarvas, G., R. Farkas, and R. Busa-Fekete, *State-of-the-art anonymization of medical records using an iterative machine learning framework.* Journal of the American Medical Informatics Association : JAMIA, 2007. **14**(5): p. 574-80.

18. Uzuner, O., et al., *A de-identifier for medical discharge summaries.* Artificial intelligence in medicine, 2008. **42**(1): p. 13-35.

19. Berman, J.J., *Doublet method for very fast autocoding.* BMC medical informatics and decision making, 2004. **4**: p. 16.

20. McMurry, A.J., et al., *A self-scaling, distributed information architecture for public health, research, and clinical care.* Journal of the American Medical Informatics Association : JAMIA, 2007. **14**(4): p. 527-33.

21. U.S. Department of Health and Human Services, N. *De-identifying Protected Health Information Under the Privacy Rule*. 2007 2/2/2007 [cited 2012 4/3/2012]; Available from: http://privacyruleandresearch.nih.gov/pr_08.asp.

22. Allwein, E.L., R.E. Schapire, and Y. Singer, *Reducing multiclass to binary: a unifying approach for margin classifiers.* J. Mach. Learn. Res., 2001. **1**: p. 113-141.

23. Savova, G.K., et al., *Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications.* Journal of the American Medical Informatics Association : JAMIA, 2010. **17**(5): p. 507-13.

24. Bodenreider, O., *The Unified Medical Language System (UMLS): integrating biomedical terminology.* Nucleic acids research, 2004. **32**(Database issue): p. D267-70.

25. Bureau, U.S.C. *Frequently Occurring First Names and Surnames From the 1990 Census*. 1990; Available from: http://www.census.gov/genealogy/names/.

26. David Ferrucci, A.L., *UIMA: an architectural approach to unstructured information processing in the corporate research environment.* Natural Language Engineering, 2004. **10**(3-4).

27. Nadkarni, P.M., L. Ohno-Machado, and W.W. Chapman, *Natural language processing: an introduction.* Journal of the American Medical Informatics Association : JAMIA, 2011. **18**(5): p. 544-51.

28. Zhang, T., *Updating an NLP System to Fit New Domains: an empirical study on the sentence segmentation problem*, IBM T.J. Watson Research Center.

29. Reshef, D.N., et al., *Detecting novel associations in large data sets.* Science, 2011. **334**(6062): p. 1518-24.

30. Frey, B.J. and D. Dueck, *Clustering by passing messages between data points.* Science, 2007. **315**(5814): p. 972-6.

31. Lin, F. and W.W. Cohen, *A Very Fast Method for Clustering Big Text Datasets*, in *Proceedings of the 2010 conference on ECAI 2010: 19th European Conference on Artificial Intelligence*2010, IOS Press. p. 303-308.

32. Dhillon, I.S. and Y. Guan, *Information Theoretic Clustering of Sparse Co-Occurrence Data*, in *Proceedings of the Third IEEE International Conference on Data Mining*2003, IEEE Computer Society. p. 517.

33. Mark Hall, E.F., Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten, *The WEKA Data Mining Software: An Update.* SIGKDD Explorations, 2009. **11**(1).

34. Quinlan, J.R., *C4.5: programs for machine learning*1993: Morgan Kaufmann.

35. Ying Yang, G.W., *Proportional k-Interval Discretization for Naive-Bayes Classifiers.* ECML01: 12th European Conference on Machine Learning 2001: p. 564-575.

36. Chen, Y., *Learning Classifiers from Imbalanced, Only Positive and Unlabeled Data Sets.* CS573 Project, (2009), 2009.

37. Neamatullah, I., et al., *Automated de-identification of free-text medical records.* BMC medical informatics and decision making, 2008. **8**: p. 32.

38. Wu, S.T., et al., *Unified Medical Language System term occurrences in clinical notes: a large-scale corpus analysis.* Journal of the American Medical Informatics Association : JAMIA, 2012. **19**(e1): p. e149-e156.

39. Drake, T.A., et al., *A system for sharing routine surgical pathology specimens across institutions: the Shared Pathology Informatics Network.* Human pathology, 2007. **38**(8): p. 1212-25.