Location Bias of Identifiers in Clinical Narratives

David A Hanauer[1], MD, Qiaozhu Mei[2], PhD, Bradley Malin[4,5], PhD, Kai Zheng[2,3], PhD

[1]Dept. of Pediatrics, Medical School; [2]School of Information; [3]School of Public Health, University of Michigan, Ann Arbor, MI. [4]Dept. of Biomedical Informatics; [5]Dept. of Electrical Engineering & Computer Science, Vanderbilt University, Nashville, TN

Abstract

Scrubbing identifying information from narrative clinical documents is a critical first step to preparing the data for secondary use purposes, such as translational research. Evidence suggests that the differential distribution of protected health information (PHI) in clinical documents could be used as additional features to improve the performance of automated de-identification algorithms or toolkits. However, there has been little investigation into the extent to which such phenomena transpires in practice. To empirically assess this issue, we identified the location of PHI in 140,000 clinical notes from an electronic health record system and characterized the distribution as a function of location in a document. In addition, we calculated the 'word proximity' of nearby PHI elements to determine their co-occurrence rates. The PHI elements were found to have non-random distribution patterns. Location within a document and proximity between PHI elements might therefore be used to help de-identification systems better label PHI.

Introduction

Electronic health records (EHRs) have enabled the accumulation of millions of clinical narratives that can be readily used for a variety of secondary use purposes including research,[1] which is an important goal of the proposed national 'learning health system'.[2] Many documents stored in EHRs are unstructured (i.e., free text) rather than structured (i.e., coded) due to the greater flexibility in which clinicians can express complex ideas.[3] One potential limitation to the broader use of these free text documents for research remains the difficulty in accurately removing identifiers in order to preserve privacy.[4] In the United States, various regulations are in place to preserve patient and research subject confidentiality, including both the Health Insurance Portability and Accountability Act (HIPAA) and the Common Rule.[5, 6] The HIPAA Privacy Rule defines 18 types of identifiers (e.g., *Patient Name* and *Date*), or protected health information (PHI), which must be removed for a dataset to be considered de-identified under the Safe Harbor standard. To achieve this goal with a large corpus of documents, automated approaches are necessary.

Much work has been done to build de-identification systems that can be used to remove PHI in a wide variety of contexts and documents. Several recent publications provide good summaries of these efforts including those related to rule-based and machine learning systems.[7-10] Machine learning systems construct statistical models to predict whether an element in a document is PHI. The underlying models incorporate multiple features that are thought to be important in discriminating PHI from non-PHI, and in classifying among PHI types. While there is no consensus as to which features are most important, many systems include (1) morphological features such as capitalization, neighboring words (often within a window of a predefined, relatively small size), and punctuations; (2) syntactic features such as parts of speech; and (3) semantic features such as dictionary terms (e.g., names, cities, hospitals) which are often incorporated as additional, extrinsic resources to improve the performance.[11-21]

Most features are based on information found *locally* near the target word(s) of interest, and these features generally do not consider the context in which the targets appear within the *global* document. One exception was a de-identification system developed by Aramaki *et al.* which used a machine learning approach that incorporated 'sentence features' including the sentence position in the record.[12] For instance, one sentence feature was divided into three categories: (1) the top ten lines; (2) the bottom five lines; and (3) all lines in-between. They also included a 'last sentence' feature that specifically targeted the last three words in the last sentence of the document. These features were incorporated because it was noted that many documents have PHI near the beginning and end of the narrative, and this additional knowledge was leveraged to boost the performance of their de-identification system. The authors found that inclusion of these sentence features increased the performance of their system, especially for *ID* (medical record number), *Date*, and *Patient Name* all of which were often found near the top of the documents.

Indeed, many clinical documents have a general structure. The top/header section often contains patient demographic information and the bottom/footer section often contains information identifying the signing clinician. In between these two regions, many clinical notes follow an overall flow of information modeled after the problem-oriented medical record initially proposed by Dr. Lawrence Weed nearly 50 years ago.[22, 23] Such notes are commonly referred to as SOAP notes, for the four main sections that are usually written in the order of Subjective, Objective, Assessment, and Plan.[24] While this structure was initially developed for paper records, clinicians have continued to follow this overall plan with electronic notes.[25]

The common document structure used by clinicians, and the success with which sentence location was used in a prior de-identification system, suggests that richer features about the structure and the layout of the documents could be leveraged to improve the performance of de-identification systems. In this paper we present an analysis that describes the distribution of PHI among a large corpus of documents with a focus on the location of the PHI elements within a document, and with respect to each other. Non-random distributions of PHI elements could imply that using additional information, such as location within a document or proximity to other PHI, might result in better detection of PHI.

Methods

A. Empiric Dataset

We randomly selected 140,000 documents from the University of Michigan's (UM) locally developed electronic health record (EHR) from decedent hematology/oncology patients. Their vital status was confirmed using two sources, the UM EHR and the UM cancer registry. Because of their decedent status this study was determined by the institutional review board to be exempt as nonhuman subjects research. A subset of these documents, as well as further details on their curation, was reported on in prior studies.[26, 27] Clinicians could create notes by dictation or by the use of customized templates if typing. The use of section headers was at the discretion of each clinician and no standardization for header types or labels existed. However, all notes had a 'footer' appended by the EHR with a clinician 'signature'. Dictated notes often had patient 'header' data added automatically by the transcription service.

B. De-identification

We used the MITRE Identification Scrubber Toolkit (MIST),[11] which uses a probabilistic model in the form of conditional random fields (CRF), to de-identify the clinical notes in our corpus. All identifiers were replaced with general placeholders, such as [AGE], [DATE], and [PATIENT NAME]. Our de-identification model was trained on 600 hand-annotated documents in the UM EHR, of which 360 had been used to build a prior model that had achieved an F-score of 0.964.[28] The distinct PHI types our model used were based on categories that the UM Health System's compliance office had defined in accordance with the standard HIPAA Safe Harbor identifiers. These included (1) *Address*; (2) *Age*; (3) *Clinician Name*; (4) *Date*; (5) *E-mail*; (6) *Health Plan Number*; (7) *Healthcare Facility*; (8) *Medical Record Number*; (9) *Patient Name*; (10) *Phone*; (11) *Place*; and (12) *URL*. Healthcare facilities included hospitals, treatment centers and nursing homes. Clinician names included any care provider names including physicians, nurses, therapists, and other healthcare workers. Patient names included all non-healthcare workers such as patients, family members, and friends of the patient. Address was specific to actual addresses such as '1500 East Medical Center Drive', whereas Place was more generic, such as 'Ann Arbor, Michigan'.

C. Characterization of PHI Distribution

We identified the location of the PHI placeholders in each document with respect to the length of the entire document, and normalized the measurements by length to allow for comparisons between documents of different lengths. Overall data were plotted as relative distributions to visualize the most likely location of each PHI element type. We compared the differences between 30,000 dictated versus 30,000 typed notes because prior work illustrated there are differences in the nature of these two document categories which could have an impact on the performance of natural language processing (NLP) systems (and many de-identification algorithms are an application of NLP).[27] We also characterized the differences in PHI distribution among the four most common document types in our corpus, using 10,000 notes from each type (see Table 1). For the overall set of 140,000 documents we used the one-sample Kolmogorov-Smirnov (KS) test (using R version 2.15.3 for OS X) to determine whether the distributions of PHI deviated significantly from a uniform probability distribution. We used the two-sample KS test to compare

differences between distributions among the 10 PHI types. For the dictated versus typed notes we used the two-sample KS test to determine differences in PHI distributions among the same note types across both document categories. KS 'D statistics' and p-values were calculated for all tests. With KS tests, smaller D statistics result when two distributions are more similar and larger D statistics are observed when the distributions are more divergent.

D. Determination of PHI Proximity

To study the correlated usage of the PHI elements, we determined the distances between consecutive appearances of PHI elements in the clinical notes. A consecutive appearance of two PHI elements (*X*, *Y*) is defined as one PHI element *X* followed by another PHI element *Y* in the same document, with at most 1,000 characters in between and without any intervening PHI. This distance is related to the concept of 'burstiness' in classical NLP literature.[29] Intuitively, when the average distance between two types of PHI elements is lower than the average distance between any two PHI elements, there is a burstiness to the elements. In other words, the usage of the two types of PHI elements is more correlated than random. We also included three special entities to use in the proximity analysis, which were: (1) *START*, the start of a clinical note; (2) *END*, the end of a clinical note; and (3) *UNDEFINED*, which recorded the cases in which there was no other PHI element before or after a given entity within 1,000 characters.

Results

A. Overall metrics

As an example of how the PHI elements were distributed throughout the documents, a visual representation of 5 documents and the location of PHI is shown in Figure 1.

The 140,000 documents in the final corpus had an average of 92 lines and 457 words. The documents contained a total of 2,553,890 PHI elements. However, 'E-mail' only appeared once in the entire corpus and 'Health Plan Number' only appeared 77 times, so both types of elements were removed from the analysis. The remaining ten PHI types are shown in Figure 2.

A total of 122 distinct document types were represented in the corpus, with the most common being 'Progress Note' (n = 25,401). There were eight distinct document types that each only appeared once including 'Ph Probe Note' and 'Thyroid Biopsy Note'. The top 20 most frequently occurring notes types are listed in Table 1, with the top 4 document types encompassing half (49.7%) of the entire corpus.

Table 1. The 20 most frequent document types of the UM EHR corpus.

Document Type	Total	% of Total
Progress Note	25,401	18.1
Letter/Note – Return Visit	22,150	15.8
Inpatient Consult Follow-up (F/U)	11,788	8.4
Phone Note	10,266	7.3
New Inpatient Consult	5,041	3.6
Nutrition Note	4,475	3.2
Emergency Department Note	4,004	2.9
Admission History & Physical	3,920	2.8
Nursing Note	3,808	2.7
Discharge Summary	3,558	2.5
Physical Therapy Inpatient Note	3,438	2.5
Social Work note	3,167	2.3
Final Plan	2,994	2.1
Letter/Note – New Patient	2,670	1.9
Nursing Progress Note	2,591	1.9
Initial Evaluation	2,458	1.8
Procedure Note	2,449	1.7
Chemotherapy Administration Note	2,322	1.7
Results Management Note	1,914	1.4
Occupational Therapy Inpatient Note	1,912	1.4

B. PHI Distribution

The distributions of PHI elements are shown in Figures 2 through 4. In each panel the distribution of PHI is shown from the start of a document (top of rectangle) to the end (bottom of rectangle). Figure 2 specifically shows the distribution for the ten PHI types across all 140,000 documents in the corpus. D statistics from the KS test are reported in Figures 2 and 4. All p-values were $< 2.2 \times 10^{-16}$ and are therefore not individually reported. The PHI type *Medical Record Number* most often appears at the start of the documents whereas *Clinician Name* and *Date* are usually found at the end. Table 3 reports KS test D statistics for the pair-wise comparisons of PHI distributions among the entire document corpus. *Age* and *URL* are most similar based on the D statistic whereas *Clinician Name* and *Medical Record Number* are most dissimilar.

In Figure 3, the distributions of five types of PHI are shown for the four most frequent document types. Caution must be used in interpreting some of the panels due to the low number of elements found in some of the document types (e.g., *Address* in Progress Notes). Nevertheless, some patterns are easy to spot visually. *Age* and *Clinician Name*, for example, each appears to have similar distributions for the two types of inpatient notes (i.e., Progress Note and Inpatient Consult Follow-up notes), but are different for the other two note types (i.e., Letter/Note – Return Visit and Phone Note). By contrast, *Name* appears to be distributed across documents types in a more consistent manner. Figure 4 compares the distribution for five PHI types across 30,000 typed notes and 30,000 dictated/transcribed notes. Visually, the distribution of *Age* as well as *Phone* appears to be most divergent across these two classes of documents, whereas based on the D statistic both *Address* and *Phone* are most divergent.

C. PHI Proximity

The average distance between any two PHI elements based on chance alone was 197.8 characters. Thus, any average distance less than approximately 200 characters would suggest a non-random distribution of words or, in this case, PHI elements. The top 20 pairs of PHI elements (including the 3 special entities *START*, *END*, *UNDEFINED*) that co-occurred in the dataset at least 10,000 times each and had an average distance of less than 200 characters are shown in Table 2. The pair *Date* → *END* can be interpreted to mean that *Date* preceded the end of a document nearly 138,000 times with an average distance of 12.3 characters before the end of the document. Similarly, the pair *START* → *Patient Name* appeared on average 42.5 characters from the beginning of the document nearly 60,000 times. Other pairs also occurred frequently in the documents but with a larger distance. These include *Date* → *Clinician Name* (n=96,228; distance 214.4), *Age* → *Date* (n=51,06; distance 233.3), *Date* → *Undefined* (n=40,276; distance 1,627.8), and *Patient Name* → *Patient Name* (n=38,275; distance 265.8).

Figure 1. Illustrations of five history and physical notes obtained from the University of Michigan EHR with the relative location of PHI highlighted in the text. The EHR automatically appends a 'signature' with *Clinician Name* and *Date* to the end of all documents in the system which is evident in all of the documents shown below.

Discussion

There is increasing awareness that documents and the language contained within them can demonstrate non-local features[30] that exist beyond the typical small word windows that are often used in natural language processing tasks, of which de-identification is a subset.[31] As mentioned in the introduction, a de-identification system that did use similar features (e.g., the location of the sentence in the document) resulted in improved performance.[12] Yet these

non-local features are not used in readily available de-identification software, such as MIST[11] or Health Information De-identification (HIDE),[15] which both leverage a probabilistic framework of CRFs.

Our analyses demonstrate that PHI is not distributed randomly throughout documents, but rather follows potentially predictable patterns. Additionally, the burstiness, or degree in which PHI elements co-occur, is also non-random. This information could be used as additional features with which to accurately label PHI in clinical narratives. Adding other metadata elements such as document type or a document's method of creation (e.g., dictated versus typed) could also provide useful features to help discriminate PHI from non-PHI. Prior evidence has illustrated that de-identification models trained and tested on the same document type tend to yield better performance than those trained and tested on a heterogeneous mix of document types,[11] but the type itself has not been leveraged in de-identification model building.

Based on our findings, we believe there are several possibilities for how this information could be translated into de-identification algorithms and software tools. First, for the overall distributions, the location within a document could be used as a standalone feature in much the same way that characteristics such as part-of-speech or word capitalization might be used. Second, the PHI distributions could be pre-computed and be used to prime a prior probability of a statistical learning method for detecting such information. Using extrinsic sources to improve the performance of de-identification tasks has been done before, such as using an external dictionary of names[32] including those derived from census data,[33] the Internet,[19] or from names in the EHR itself.[34] Third, the numerical distance between PHI elements can be considered as a new category of proximity features, which provides a more general treatment than using a fixed-length window.

Table 2. Distance between PHI entities and the number of co-occurrences. Based on the document metrics, average distance of less than 200 characters suggests a non-random distribution.

PHI Pair (initial → subsequent)	Total Consecutive Co-occurrences	Average Distance (characters)
Address → Place	17,003	0.6
Patient Name → Med Rec Num	53,347	10.2
Med Record Number → Date	54,370	11.6
Date → END	137,853	12.3
Clinician Name → Address	10,793	13.5
Clinician Name → Date	194,121	27.0
Patient Name → Age	45,351	30.1
START → Patient Name	59,494	42.5
Phone → Phone	26,782	51.9
Place → Clinician Name	17,977	52.7
Date → Date	965,024	55.2
Clinician Name → Phone	18,263	56.7
Clinician Name → Clinician	273,381	64.1
Patient Name → Phone	12,160	80.1
Phone → Clinician Name	42,850	97.4
Healthcare Facility → Date	14,812	126.6
START → Date	37,804	127.9
Clinician Name → Pat Name	16,214	129.2
Date → Phone	22,136	134.8
Phone → Date	13,872	138.2

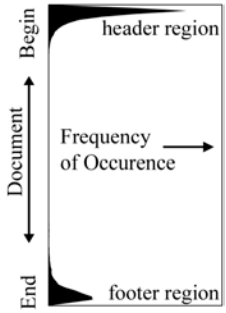
We wish to highlight that there are similarities between our work and that of others. First, a process for removing names in pathology reports leveraged the observation that names often occurred in pairs, such as forename followed by surname or 'Mrs' followed by surname.[35] In our analysis of burstiness, it was found that *Patient Name → Patient Name* appeared on average 265.8 characters apart rather than being adjacent like the prior study. However, the likely explanation for the difference of our findings from the previous work is that we labeled an entire name (first and last) as a single PHI entity (e.g., 'Mary Smith' became simply [PATIENT NAME]), and thus they would not have been counted as directly adjacent with our approach.

Second, it should be noted that work has been done to utilize non-local features to improve NLP algorithms, albeit not specific to de-identification tasks. These have included using non-local dependencies for improving named entity recognition algorithms such as using a 'majority feature' for labeling entities consistently.[36] A majority feature will consider identical entities labeled differently and assign all of them to the most common label. For example, if 'University of Michigan' was labeled as *Healthcare Facility* twice and as *Location* once, all three would ultimately be assigned to *Healthcare Facility*. Other investigations found that including long-distance dependencies or global (e.g., sentence-level as opposed to token-level) features in their models improved natural language

processing tasks.[37, 38] More specifically, location features[39-41] as well as word distance (in terms of proximity) have been used to improve the performance of CRF models.[42, 43]

Beyond the conflation of forename and surname, there are several limitations to our study. First, while we included a large number of documents and document types in our corpus, they were derived from a single institution's EHR system. Additionally, while some PHI elements were well represented, others were rare. However, it may be the case that some PHI elements would not need to rely on additional features to boost accuracy since simple patterns using regular expressions may already work well enough for highly structured elements such as phone or Social Security numbers.[33] Second, some of the characteristics of our documents (e.g., number of words) may have been slightly changed by removing the identifiers and replacing them with common labels. Third, normalizing PHI location as a function of document length may introduce biases and add unnecessary variation that could negatively impact performance. Lastly, the identifiers in our document corpus were automatically labeled using MIST,[11] which derived a model that was built from 600 documents. It is likely that some labeling errors were introduced and some of these may have been non-random. However, we believe the impact on our analysis would be small.

Figure 2. Normalized distribution for 10 types of PHI, derived from 140,000 documents. The top of each rectangle represents the beginning of a document, and the bottom represents the end. Clinician name (i.e., *Clinician*), for example, mostly appears at the end of documents. Thus, an element near the bottom of the document is more likely to be a clinician name compared to the middle. Numbers in parentheses display how many elements were used in making the figure. One-sample Kolmogorov-Smirnov D statistics are reported.



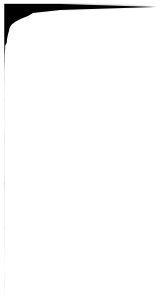
Address
(n=17,830)
D statistic: 0.545

Age
(n=117,912)
D statistic: 0.330

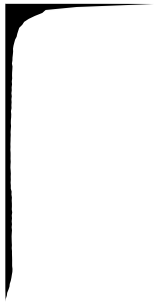
Clinician Name
(n=536,303)
D statistic: 0.483

Date
(n=1,394,046)
D statistic: 0.447

Healthcare Facility
(n=46,314)
D statistic: 0.270



Med Record Number
(n=58,658)
D statistic: 858



Patient Name
(n=229,204)
D statistic: 0.347

Phone
(n=96,634)
D statistic: 0.153

Place
(n=40,806)
D statistic: 0.259

URL
(n=15,095)
D statistic: 0.351

Conclusion

In this analysis we found that 1) PHI is distributed throughout clinical narratives in a non-random manner and 2) the co-occurrence of PHI elements often appear together in the document in a non-random manner. We believe this information may be useful as additional features are leveraged to further improve the performance of de-identification systems. Incorporating additional document features, such as section headings,[44] could also help

identify the location in which potential PHI elements appear (and may obviate the need for more detailed location information) as this has helped with other NLP applications.[45-47] Future work should test the merits of these measures by incorporating them as additional features into de-identification systems.

Figure 3. Relative distribution of 5 PHI types, comparing the top four most frequently occurring note types.

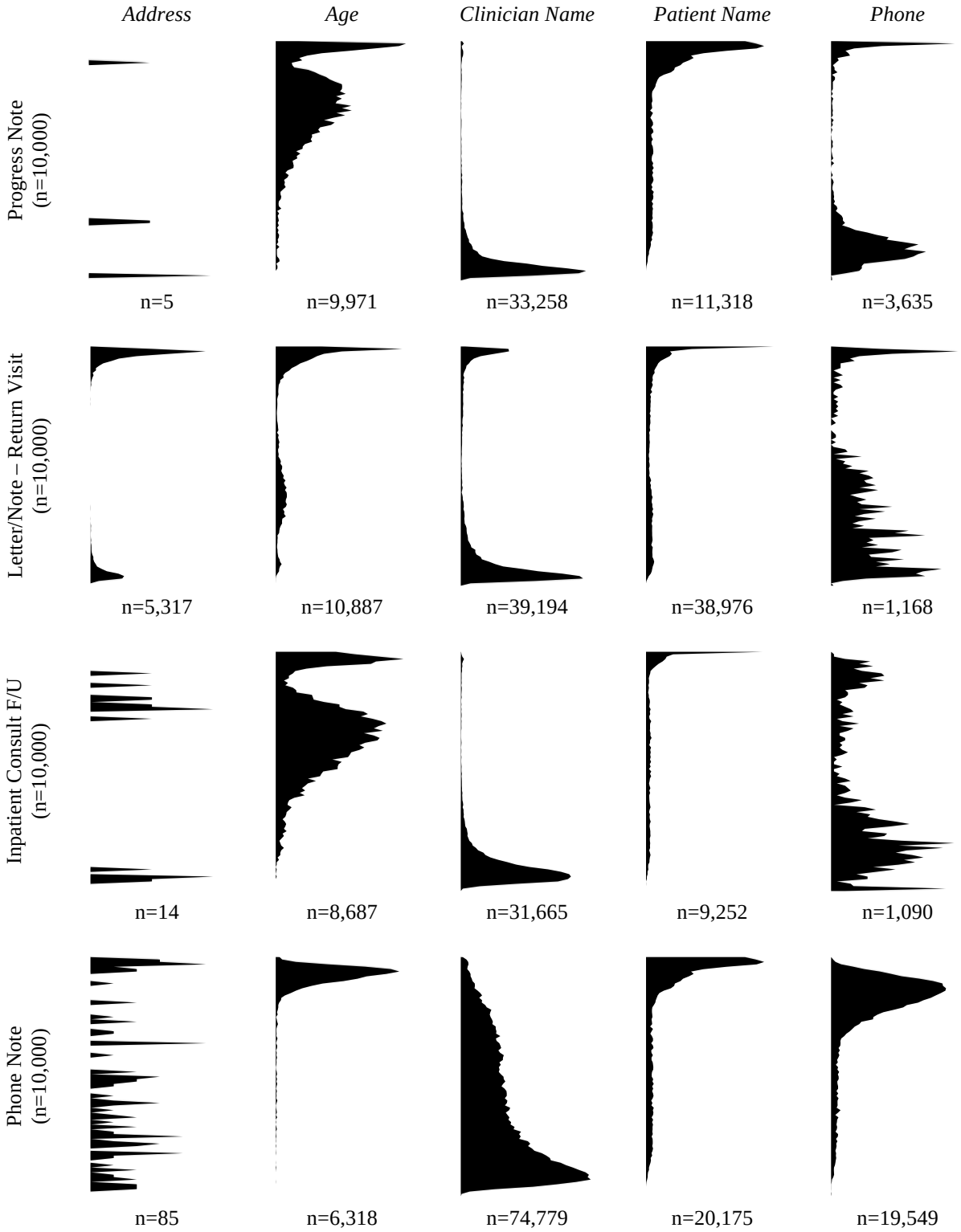


Figure 4. Relative distribution of 5 PHI types, comparing dictated versus typed notes. Two-sample Kolmogorov-Smirnov D statistics are also reported.

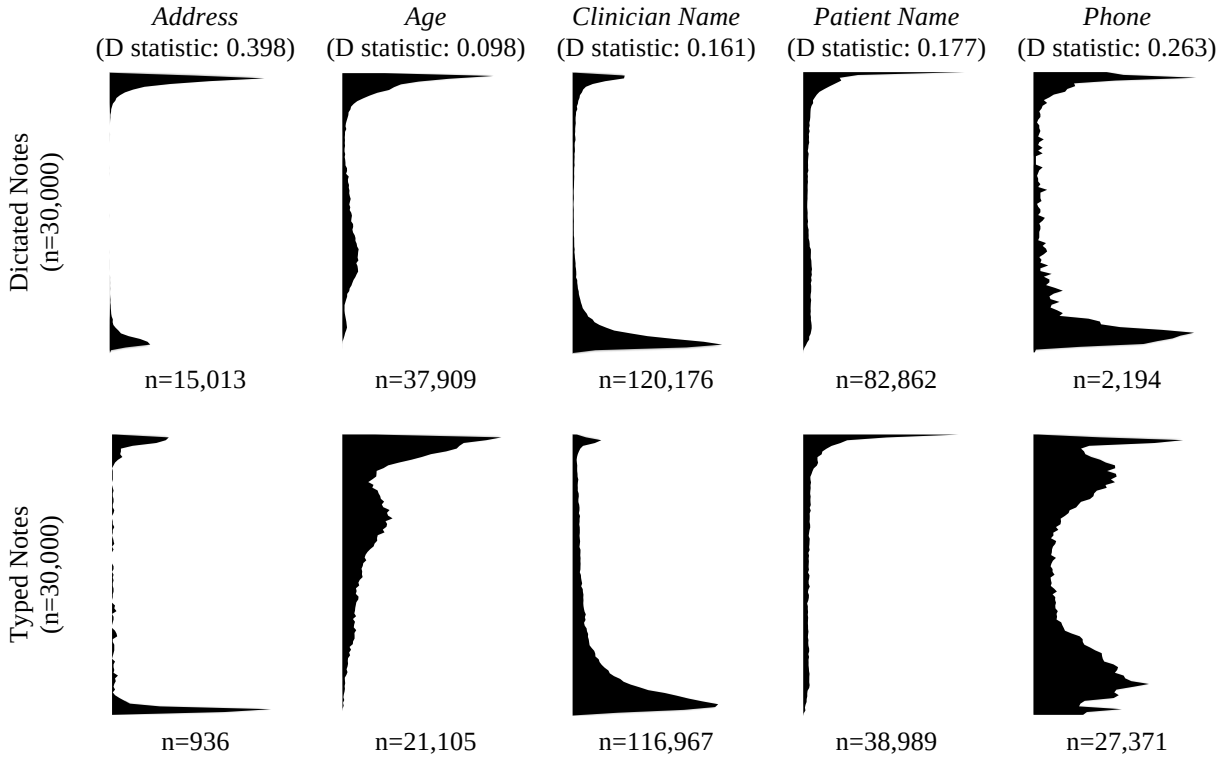


Table 3. Two-sample Kolmogorov-Smirnov D statistics for the pair-wise comparisons of PHI elements from the corpus of 140,000 documents. Lower numbers are shaded darker and represent greater similarities between the two distributions being compared.

	URL	*Place*	*Phone*	*Patient Name*	*Med Rec Num*	*Healthcare Facility*	*Date*	*Clinician Name*	*Age*
Address	0.279	0.288	0.500	0.269	0.430	0.352	0.510	0.568	0.308
Age	0.079	0.199	0.365	0.139	0.575	0.150	0.545	0.676	
Clinician Name	0.651	0.500	0.369	0.580	0.896	0.526	0.275		
Date	0.491	0.358	0.422	0.498	0.829	0.461			
Healthcare Facility	0.136	0.104	0.225	0.197	0.666				
Med Rec Num	0.592	0.600	0.810	0.517					
Patient Name	0.191	0.185	0.295						
Phone	0.361	0.211							
Place	0.154								

Acknowledgements

The research reported in this publication was supported in part by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR000433, the National Library of Medicine under Award Number 1R01LM011366, the National Human Genome Research Institute under Award Number 1U01HG006385, and the National Science Foundation under grant numbers CCF-0424422 and CCF-1048168. The content is solely the responsibility of the authors and does not necessarily represent the official views of the supporting agencies. We would also like to thank John Aberdeen from the MITRE Corp. for his input.

References

1. **Ohno-Machado L**. Realizing the full potential of electronic health records: the role of natural language processing. J Am Med Inform Assoc. 2011 Sep-Oct;**18**(5):539.
2. **Friedman CP**, Wong AK, Blumenthal D. Achieving a nationwide learning health system. Sci Transl Med. 2010 Nov 10;**2**(57):57cm29.
3. **Rosenbloom ST**, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. J Am Med Inform Assoc. 2011 Mar-Apr;**18**(2):181-6.
4. **Malin BA**, Emam KE, O'Keefe CM. Biomedical data privacy: problems, perspectives, and recent advances. J Am Med Inform Assoc. 2013 Jan 1;**20**(1):2-6.
5. Summary of the HIPAA Security Rule. [cited February 10, 2013]; Available from: http://www.hhs.gov/ocr/privacy/hipaa/understanding/srsummary.html
6. Human Subjects Research (45 CFR 46). [cited February 10, 2013]; Available from: http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html
7. **Ferrandez O**, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. Evaluating current automatic de-identification methods with Veteran's health administration clinical documents. BMC Med Res Methodol. 2012;**12**:109.
8. **Kushida CA**, Nichols DA, Jadrnicek R, Miller R, Walsh JK, Griffin K. Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. Med Care. 2012 Jul;**50 Suppl**:S82-101.
9. **Meystre SM**, Friedlin FJ, South BR, Shen S, Samore MH. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Med Res Methodol. 2010;**10**:70.
10. **Uzuner O**, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc. 2007 Sep-Oct;**14**(5):550-63.
11. **Aberdeen J**, Bayer S, Yeniterzi R, et al. The MITRE Identification Scrubber Toolkit: design, training, and assessment. Int J Med Inform. 2010 Dec;**79**(12):849-59.
12. **Aramaki E**, Imai T, Miyo K, Ohe K. Automatic Deidentification by using Sentence Features and Label Consistency. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, Washington, DC; 2006.
13. **Deleger L**, Molnar K, Savova G, et al. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. J Am Med Inform Assoc. 2013 Jan 1;**20**(1):84-94.
14. **Ferrandez 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 Assoc. 2013 Jan 1;**20**(1):77-83.
15. **Gardner J**, Xiong L. HIDE: An Integrated System for Health Information DE-identification. 2008:254-9.
16. **Gardner J**, Xiong L, Wang F, Post A, Saltz J, Grandison T. An evaluation of feature sets and sampling techniques for de-identification of medical records. 2010:183.
17. **Guo Y**, Gaizauskas T, Roberts I, Demetriou G, Hepple M. Identifying Personal Health Information Using Support Vector Machines. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, Washington, DC; 2006.
18. **Hara K**. Applying a SVM Based Chunker and a Text Classifier to the Deid Challenge. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, Washington, DC; 2006.
19. **Szarvas G**, Farkas R, Busa-Fekete R. State-of-the-art anonymization of medical records using an iterative machine learning framework. J Am Med Inform Assoc. 2007 Sep-Oct;**14**(5):574-80.
20. **Taira RK**, Bui AA, Kangarloo H. Identification of patient name references within medical documents using semantic selectional restrictions. Proc AMIA Symp. 2002:757-61.

21. **Uzuner O**, Sibanda TC, Luo Y, Szolovits P. A de-identifier for medical discharge summaries. Artif Intell Med. 2008 Jan;**42**(1):13-35.
22. **Weed LL**. Medical records that guide and teach. N Engl J Med. 1968 Mar 14;**278**(11):593-600.
23. **Weed LL**. Medical records that guide and teach. N Engl J Med. 1968 Mar 21;**278**(12):652-7 concl.
24. **Zierler-Brown S**, Brown TR, Chen D, Blackburn RW. Clinical documentation for patient care: models, concepts, and liability considerations for pharmacists. Am J Health Syst Pharm. 2007 Sep 1;**64**(17):1851-8.
25. **Rosenbloom ST**, Grande J, Geissbuhler A, Miller RA. Experience in implementing inpatient clinical note capture via a provider order entry system. J Am Med Inform Assoc. 2004 Jul-Aug;**11**(4):310-5.
26. **Hanauer DA**, Liu Y, Mei Q, Manion FJ, Balis UJ, Zheng K. Hedging their mets: the use of uncertainty terms in clinical documents and its potential implications when sharing the documents with patients. AMIA Annu Symp Proc. 2012;**2012**:321-30.
27. **Zheng K**, Mei Q, Yang L, Manion FJ, Balis UJ, Hanauer DA. Voice-dictated versus typed-in clinician notes: linguistic properties and the potential implications on natural language processing. AMIA Annu Symp Proc. 2011;**2011**:1630-8.
28. **Hanauer D**, Aberdeen J, Bayer S, et al. Bootstrapping a de-identification system for narrative patient records: Cost-performance tradeoffs. Int J Med Inform. 2013 Sep;**82**(9):821-31.
29. **Church KW**, Gale WA. Poisson mixtures. Natural Language Engineering. 2008;**1**(02).
30. **Altmann EG**, Pierrehumbert JB, Motter AE. Beyond word frequency: bursts, lulls, and scaling in the temporal distributions of words. PLoS One. 2009;**4**(11):e7678.
31. **Meystre SM**, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform. 2008:128-44.
32. **Douglass M**, Clifford GD, Reisner A, Moody GB, Mark RG. Computer-assisted de-identification of free text in the MIMIC II database. 2004:341-4.
33. **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.
34. **Roden DM**, Pulley JM, Basford MA, et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther. 2008 Sep;**84**(3):362-9.
35. **Thomas SM**, Mamlin B, Schadow G, McDonald C. A successful technique for removing names in pathology reports using an augmented search and replace method. Proc AMIA Symp. 2002:777-81.
36. **Krishnan V**, Manning CD. An effective two-stage model for exploiting non-local dependencies in named entity recognition. 2006:1121-8.
37. **Sutton C**, McCallum A. Collective Segmentation and Labeling of Distant Entities in Information Extraction; 2004.
38. **Nakagawa T**. Multilingual Dependency Parsing Using Global Features. Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007,; 2007; Prague; 2007. p. 952-6.
39. **Feng H**, Guo-hui L. Mining Chinese comparative sentences by semantic role labeling. Machine Learning and Cybernetics, 2008 International Conference on; 2008 12-15 July 2008; 2008. p. 2563-8.
40. **Li J**, Wang R, Wang W, Gu B, Li G. Automatic Labeling of Semantic Role on Chinese FrameNet Using Conditional Random Fields. Web Intelligence and Intelligent Agent Technologies, 2009 WI-IAT '09 IEEE/WIC/ACM International Joint Conferences on; 2009 15-18 Sept. 2009; 2009. p. 259-62.
41. **Nishiyama R**, Tsuboi Y, Unno Y, Takeuchi H. Feature-Rich Information Extraction for the Technical Trend-Map Creation. Proceedings of the 8th NTCIR Workshop Meeting; 2010; 2010. p. 318-24.
42. **Yang Z**, Lin H, Li Y. BioPPISVMExtractor: a protein-protein interaction extractor for biomedical literature using SVM and rich feature sets. J Biomed Inform. 2010 Feb;**43**(1):88-96.
43. **Hui Z**, Liu J, L. O. Question Classification Based on an Extended Class Sequential Rule Model. Proceedings of the 5th International Joint Conference on Natural Language Processing; 2011; Chiang Mai, Thailand; 2011. p. 938–46.
44. **Denny JC**, Spickard A, 3rd, Johnson KB, Peterson NB, Peterson JF, Miller RA. Evaluation of a method to identify and categorize section headers in clinical documents. J Am Med Inform Assoc. 2009 Nov-Dec;**16**(6):806-15.
45. **Carroll RJ**, Thompson WK, Eyler AE, et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. J Am Med Inform Assn. 2012;**19**(e1):e162-e9.
46. **Denny JC**, Bastarache L, Sastre EA, Spickard A, 3rd. Tracking medical students' clinical experiences using natural language processing. J Biomed Inform. 2009 Oct;**42**(5):781-9.
47. **Xu H**, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. J Am Med Inform Assoc. 2010 Jan-Feb;**17**(1):19-24.