doi:10.3233/978-1-61499-753-5-216

Prevalence Estimation of Protected Health Information in Swedish Clinical Text

Aron HENRIKSSON[a,1], Maria KVIST[a,b] and Hercules DALIANIS[a]
[a] *Department of Computer and Systems Sciences, Stockholm University, Sweden*
[b] *Department of Laboratory Medicine, Karolinska Institutet, Sweden*

Abstract. Obscuring protected health information (PHI) in the clinical text of health records facilitates the secondary use of healthcare data in a privacy-preserving manner. Although automatic de-identification of clinical text using machine learning holds much promise, little is known about the relative prevalence of PHI in different types of clinical text and whether there is a need for domain adaptation when learning predictive models from one particular domain and applying it to another. In this study, we address these questions by training a predictive model and using it to estimate the prevalence of PHI in clinical text written (1) in different clinical specialties, (2) in different types of notes (i.e., under different headings), and (3) by persons in different professional roles. It is demonstrated that the overall PHI density is 1.57%; however, substantial differences exist across domains.

Keywords. electronic health records, protected health information, de-identification, natural language processing, predictive modeling

1. Introduction

Healthcare produces an abundance of data that is stored in electronic health record (EHR) systems. The secondary use of EHR data, which describes the health conditions and treatments of patients over time, holds much promise in facilitating medical research and epidemiological activities; EHR data can also be exploited for providing clinical decision support at the point of care. However, this requires that privacy-preserving measures, such as de-identification, are taken. Automatic de-identification of EHR data includes the detection and obscuring of sensitive information in clinical notes. The US Health Insurance Portability and Accountability Act (HIPAA) defines 18 types of protected health information (PHI) that should be obscured for EHR data to be considered de-identified [1]. In recent years, there has been a surge in research efforts to construct automatic de-identification tools [2,3], many of which rely on machine learning and manually annotated corpora for identifying PHI in clinical notes.

In this study, we seek to estimate the prevalence of PHI in Swedish clinical text. In particular, we want to uncover if differences exist in the distribution of PHI – both generally and with respect to specific PHI classes – across different types of notes. That this may, in fact, be the case is substantiated by the knowledge that one writes differently in different clinical specialties and professional roles [4]. The findings from this study are also intended to inform future development of automatic de-identification systems.

[1] Corresponding author: aron.henriksson@dsv.su.se

Previous research has attempted to estimate PHI prevalence based on small samples of annotated data and covered only a few types of clinical notes. An early study based on a sample of nursing notes in MIMIC-II – an EHR database which has indeed been de-identified and made publicly available for research – revealed that around 0.5% of all tokens were instances of PHI [5]. In another study, PHI density in a diverse set of clinical domains was found to be 2.9% and name density 1% [6]. In discharge summaries, the PHI density amounted to around 3.6% [7]. A similar PHI density was found in the 2014 i2b2/UTHealth corpus, comprising health records of diabetic patients [8]. In one study, the distribution of PHI classes across different types of notes, i.e. written under different headings, was described: the most prevalent PHI types were dates, names and phone numbers, while the note types with the highest PHI density were *Discharge Summary*, *Outpatient Consult*, and *Admission History and Physical* [9]. A few similar studies have been performed on non-English languages. In a French corpus comprising clinical notes of various types from a range of specialties, PHI density was as high as 11% [10], while a study of Danish clinical text revealed a PHI density of around 1.8% [11].

2. Methods & Materials

This study seeks to estimate the prevalence of PHI in Swedish clinical text and to learn if differences therein exist between types of clinical text. As manual annotation is cumbersome, we estimate PHI prevalence by applying a predictive model that has been trained on an existing PHI corpus to a larger unannotated corpus of clinical text. Both the annotated corpus [12] and the data extracted for this study are from the Stockholm EPR Corpus[2] [13], which contains health records from Karolinska University Hospital.

The annotated PHI corpus comprises 100 health records from five different clinics (Neurology, Orthopedics, Infection, Dental Surgery, and Nutrition) in 2008. The corpus contains a total of 198,466 tokens (0.1 types/token) and 4,220 annotated PHI instances. The PHI density is 2.13% and the class distribution is as follows: Health Care Unit (23.9%), First Name (21.7%), Last Name (21.5%), Date Part (16.6%), Full Date (8.7%), Location (3.3%), Phone Number (3.2%), Age (1.2%). The average sentence length is 8.9 (± 6.4) tokens. Approximately 13.7% of all sentences include a least one PHI mention, while, on average, there are 0.19 (± 0.55) PHI mentions per sentence.

This manually annotated PHI corpus was used for training a predictive model. To that end, a linear-chain CRF [14] was used that, in addition to being dependent on the input features, is also dependent on the previous and subsequent output variable. The following features were used: (a) token, (b) lemma, (c) part of speech, (d) capitalization, (e) digit, (f) compounds, (g) dictionary matching against SNOMED CT, MeSH etc. The same features were used in previous studies on named entity recognition in Swedish clinical text [15,16] (see [15] for more details). As is common for sequence labeling tasks, IOB-encoding of class labels was used, which indicates whether a token is at the beginning (B), inside (I) or outside (O) a given named entity mention. 10-fold cross-validation was carried out when tuning the CRF hyperparameters: two forms of regularization (L1/L2), the c-value governing the balancing between underfitting and overfitting, and the window size, which determines to what extent dependencies should be modeled between input features and output variables. Considered c-values were 2^x,

[2] This research has been approved by the Regional Ethical Review Board in Stockholm (2012/834-31/5).

where x ∈ {-2,-1,0,1,2,3,4,5}, while the following symmetric window sizes were explored: 1+1, 2+2, 3+3, 4+4.

A predictive model was then trained on the entire annotated PHI corpus using the best-observed set of hyperparameters and subsequently applied on various subsets of the unannotated corpus, comprising all clinical notes from a single year: 2009. PHI prevalence was estimated and compared along three dimensions: (1) specialty, i.e., notes from units belonging to different clinical practices (geriatrics, oncology, orthopedics), (2) note types, i.e., notes written under different headings (admission, day, discharge), and (3) professions, i.e., notes written by persons in different professional roles (physicians, nurses, physiotherapists). Since these subcorpora are not equal in size, we use a normalized metric, PHI density, to quantify prevalence: this is defined as the number of PHI mentions divided by the total number of tokens in the corpus.

3. Results

The results of the parameter optimization are shown in Figure 1. The trend was the same irrespective of whether F_1-scores were micro- or macro-averaged over classes: L2 regularization led almost invariably to higher performance, with higher c values favored in comparison to L1 regularization. The best results (precision: 92.65%, recall: 81.29%, F_1: 0.87) were obtained using a narrow context window (1+1) and a c value of 16.

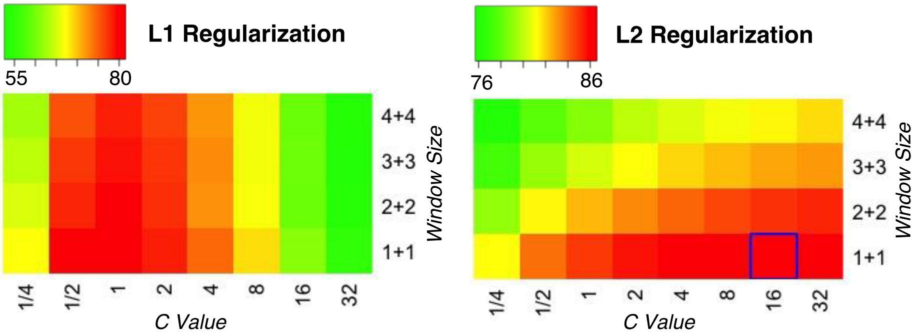

Figure 1. Parameter optimization results with 10-fold cross-validation on the training set

A predictive model was then trained and applied to various subsets of clinical notes from a different year. Descriptive statistics of the sub-corpora along with PHI density estimates are shown in Table 1. The average sentence length is 9.1 tokens, with no domain standing out from the others. There are, however, differences in type-token ratios, with larger lexical variation observed in notes written by physicians (0.007) and nurses (0.008) compared to physiotherapists (0.015); there is less lexical variation in geriatrics (0.016) than in oncology (0.011) and orthopedics (0.010). Interesting to note is that the overall type-token ratio is as low as 0.004. On average across domains, the PHI density is 1.57%. There is an average of 0.14 PHI tokens per sentence and around 10% of sentences contain at least one PHI instance. With respect to different specialties, there is a fairly substantial difference in PHI density between orthopedics (1.50%), on the one hand, and geriatrics (2.12%) and oncology (2.05%), on the other. Looking at specific PHI classes, there are relatively fewer dates in the notes produced in orthopedics;

geriatrics mentions relatively more first names, ages and phone numbers; oncology generally stands out less, but there does seem to be a propensity for writing dates. When it comes to different types of clinical notes, the observed PHI distribution is even more skewed: admission notes comprise the least amount of PHI (1.00%), discharge notes the most (2.94%), and day notes somewhere in the middle (1.64%). Names and health care units are particularly prevalent in discharge notes in comparison to the other note types. Regarding notes written by persons in different professions, differences in PHI density are somewhat smaller, with physiotherapist notes exhibiting the least amount of PHI (0.97%), followed by physician notes (1.43%) and nurse notes (1.66%). Nurses seem to mention names, dates and locations to a greater extent than do the other professions.

Table 1. PHI density estimates across various types of clinical text

	Specialty			**Note**			**Profession**		
	Geriatrics	Oncology	Orthopedics	Admission	Day	Discharge	Dr	Nurse	Physio
Sentence Length	9.9 ± 8.5	9.5 ± 7.0	8.7 ± 6.5	8.4 ± 7.1	9.5 ± 6.9	9.6 ± 8.8	9.5 ± 7.6	8.1 ± 5.8	10.4 ± 8.6
Type:Token	0.016	0.010	0.011	0.015	0.014	0.017	0.007	0.008	0.015
PHI Density (%)	2.12	2.05	1.50	1.00	1.61	2.94	1.43	1.66	0.97
First Name	0.55	0.32	0.33	0.11	0.29	0.58	0.29	0.37	0.31
Last Name	0.44	0.39	0.45	0.17	0.35	0.68	0.36	0.44	0.22
Age	0.04	0.01	0.02	0.05	0.01	0.06	0.03	0.00	0.01
Health Care Unit	0.32	0.30	0.24	0.25	0.20	0.47	0.26	0.16	0.11
Location	0.07	0.07	0.06	0.10	0.06	0.08	0.09	0.58	0.04
Full Date	0.30	0.27	0.17	0.10	0.12	0.59	0.17	0.14	0.09
Date Part	0.32	0.68	0.20	0.22	0.56	0.43	0.22	0.44	0.16
Phone Number	0.08	0.02	0.03	0.00	0.03	0.05	0.01	0.04	0.04

4. Discussion

This study sought to estimate the prevalence of PHI in Swedish clinical text and investigate differences across various types of notes. The amount of sensitive information (1.57%) is comparable to previous reports on other languages, although numbers range from 0.5% to 11%. It is, however, problematic to compare these directly as different PHI classes and definitions are used. In contrast to previous work, we looked specifically at different types of notes written in different specialties and professions. This revealed some notable differences in PHI density, primarily when comparing admission, day and discharge notes. The highest PHI density was observed in discharge notes: almost 20% of the sentences contained at least one PHI instance. Many plausible explanations can be found for the observed differences in PHI density: names, especially surnames are, e.g., prevalent in discharge summaries in part because physicians involved in the healthcare process are typically mentioned; that physicians mention healthcare units and nurses mention locations, respectively, can be attributed to the fact that physicians tend to write about the healthcare process both within the hospital and with general practitioners, while nurses need also to coordinate with the outside world.

Knowing about differences in PHI density is useful for development of de-identification systems, e.g. when creating training data for machine learning. In future work, we plan to evaluate the predictive performance on different types of notes in order to assess to what extent domain adaptation may be necessary. In comparison to previous studies, in which small annotated corpora were used, we proposed an alternative way of estimating PHI prevalence by using predictive modeling. While highly efficient, there are limitations in terms of reliability as the estimates are dependent on the performance of the predictive model. Here, we obtained an F_1-score of 0.87, cross-validated on the training set, outperforming previous models trained on the same corpus [12,17,18]; however, in future work, we also need to determine the performance on the target domains.

References

[1] HIPAA 2003. Health Insurance Portability and Accountability (HIPAA), Privacy Rule and Public Health Guidance, 2003, From CDC and the U.S. Department of Health and Human Services, 2016. Accessed 2016-10-18.

[2] Özlem Uzuner, Yuan Luo, and Peter Szolovits. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 14(5):550–563, 2007.

[3] Stephane Meystre, Jeffrey Friedlin, Brett South, Shuying Shen, and Matthew Samore. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC medical research methodology*, 10(1):70, 2010.

[4] Kelly Smith, Beata Megyesi, Sumithra Velupillai, and Maria Kvist. Professional language in Swedish clinical text: Linguistic characterization and comparative studies. *Nordic Journal of Linguistics*, 37(02):297–323, 2014.

[5] Margaret Douglass, Gari D Clifford, Andrew Reisner, George B Moody, and Roger G Mark. Computer-assisted de-identification of free text in the mimic ii database. *Computers in Cardiology*, pages 341–344, 2004.

[6] David A Dorr, WF Phillips, Shobha Phansalkar, Shannon A Sims, and John Franklin Hurdle. Assessing the difficulty and time cost of de-identification in clinical narratives. *Methods of information in medicine*, 45(3):246–252, 2006.

[7] Özlem Uzuner, Tawanda C Sibanda, Yuan Luo, and Peter Szolovits. A de-identifier for medical discharge summaries. *Artificial intelligence in medicine*, 42(1):13–35, 2008.

[8] Amber Stubbs and Özlem Uzuner. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/uthealth corpus. *Journal of biomedical informatics*, 58:S20–S29, 2015.

[9] David Hanauer, John Aberdeen, Samuel Bayer, Benjamin Wellner, Cheryl Clark, Kai Zheng, and Lynette Hirschman. Bootstrapping a de-identification system for narrative patient records: cost-performance tradeoffs. *International journal of medical informatics*, 82(9):821–831, 2013.

[10] Cyril Grouin and Aurélie Névéol. De-identification of clinical notes in French: towards a protocol for reference corpus development. *Journal of biomedical informatics*, 50:151–161, 2014.

[11] Kostas Pantazos, Soren Lauesen, and Soren Lippert. De-identifying an EHR database - anonymity, correctness and readability of the medical record. In *Medical Informatics Europe*, pages 862-866, 2011.

[12] Hercules Dalianis and Sumithra Velupillai. De-identifying Swedish clinical text-refinement of a gold standard and experiments with Conditional random fields. *J. Biomedical Semantics*, 1:6, 2010.

[13] Hercules Dalianis, Martin Hassel, Aron Henriksson, and Maria Skeppstedt. Stockholm EPR Corpus: a clinical database used to improve health care. In *Swedish Language Technology Conference*, 2012.

[14] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[15] Maria Skeppstedt, Maria Kvist, Gunnar H Nilsson, and Hercules Dalianis. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: an annotation and machine learning study. *Journal of biomedical informatics*, 49:148–158, 2014.

[16] Aron Henriksson, Maria Kvist, Hercules Dalianis, and Martin Duneld. Identifying adverse drug event information in clinical notes with distributional semantic representations of context. *Journal of biomedical informatics*, 57:333–349, 2015.

[17] Aron Henriksson, Hercules Dalianis, and Stewart Kowalski. Generating features for named entity recognition by learning prototypes in semantic space: The case of de-identifying health records. In *International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 450–457, 2014.

[18] Aron Henriksson. Learning multiple distributed prototypes of semantic categories for named entity recognition. *International journal of data mining and bioinformatics*, 13(4):395–411, 2015.