Utility Preservation of Clinical Text After De-Identification

Thomas Vakili and Hercules Dalianis
Department of Computer and Systems Sciences (DSV)
Stockholm University
Kista, Sweden
{thomas.vakili, hercules}@dsv.su.se

Abstract

Electronic health records contain valuable information about symptoms, diagnosis, treatment and outcomes of the treatments of individual patients. However, the records may also contain information that can reveal the identity of the patients. Removing these identifiers - the Protected Health Information (PHI) - can protect the identity of the patient. Automatic de-identification is a process which employs machine learning techniques to detect and remove PHI. However, automatic techniques are imperfect in their precision and introduce noise into the data. This study examines the impact of this noise on the utility of Swedish de-identified clinical data by using human evaluators and by training and testing BERT models. Our results indicate that de-identification does not harm the utility for clinical NLP and that human evaluators are less sensitive to noise from de-identification than expected.

1 Introduction

The training data for clinical NLP models are sensitive because they often contain information that can reveal the identity of real patient. This makes sharing data and models difficult.

One way of decreasing the privacy risks of using clinical data is to de-identify them. A popular way of doing this is by automatically detecting and removing Protected Health Information (PHI). This is often done using Named Entity Recognition models that can find these sensitive data in clinical texts. The de-identified clinical data can be used both for clinical research but also as training data for machine learning algorithms. These approaches will be described in the next section.

One concern is that de-identification will deteriorate the quality of clinical texts. The risk is that this will not only harm down-stream NLP tasks, but also make the data less useful for other research purposes. We have also identified a hesitancy from lawyers who fear that de-identification can harm the safety of the data.

In this paper, we evaluate the extent to which de-identification harms perceived utility using human evaluators. We then evaluate the impact of de-identification on the utility of the datasets for building clinical NLP models.

2 Related Research

This related research section presents studies regarding the quality of the de-identification system both in terms of safety and privacy but also for down-stream tasks as for medical research.

In a study by Meystre et al. (2014) 86 patient records in English were de-identified. Eight physicians and 11 medical students that have treated and written these records 1-3 months earlier could not recognise their patients. Some of physician suspected that they could recognize their patient on some clinical details but after a control it was found that the wrong patient had been identified.

Sánchez et al. (2014) propose a sanitation process that removes information that might make the patient record sensitive. This is not done by replacing typical PHI like names, but instead by replacing sensitive diseases such as *Clamydia, AIDS,* or *HIV* with less sensitive terms such as *virus*. The ideas is to aggregate information but this limits the utility of the patient records.

Dalton-Locke et al. (2020) used de-identified patient records from mental health clinics to perform research regarding the housing service of patients suffering from mental illnesses. Structured data had previously been used for this research. The researchers compared the two approaches and found it feasible to use de-identified patient records and de-identified structured data jointly for this research. The system called CRIS is a combination of a de-identification algorithm and a security model has been approved for use in mental health research by the ethics board. This allows researchers to ex-

Proceedings of the BioNLP 2022 workshop, Dublin, Ireland, pages 383–388
May 26, 2022. ©2022 Association for Computational Linguistics

tract data from the the patient record system without requiring individuals' informed consent (Fernandes et al., 2013).

In an other study, Dalianis (2019) constructed a pseudonymization system for Swedish clinical text. The pseudonymization system replaced PHI with pseudonyms or surrogates. All tags that could identify a PHI were removed so the records looked realistic and neutral. The system was evaluated by two computer scientist that had worked with clinical text mining, specifically with this type of text. The text they evaluated had not been seen by the scientists before. They read 98 patient records where half were pseudonymized and the other half were not. On average, 91 percent of the pseudonymized records were judged as original.

Pantazos et al. (2017) carried out de-identification and pseudonymization of over 323,000 Danish patient records and then carried out a manual review of 369 de-identified and pseudonymized patient records with a total length of over 71,000 words, this revealed seven words where quasi-identifiers[1] had not been de-identified and it revealed 109 words where it was incorrectly de-identified, this reduced the medical correctness and readability according to the authors. A finding by the authors was if they use abbreviation lists and also medical lists the number of false positives would probably been diminished.

Berg et al. (2020) evaluated the performance of the Conditional Random Field (CRF) algorithms on down-stream tasks based on clinical training text that have been de-identified with increasing degrees of recall. The authors used four different de-identification strategies: pseudonymization (replace with surrogats), masking, keeping the class name and removing the entire sentence containing the PHI. Pseudonymization was the most effective strategy for preserving down-stream utility. Masking and replacing the PHI with the class name had a larger negative impact. The most severe impact was seen when employing the sentence removal strategy. However, a balanced recall (not high recall) on all four strategies did not affect the down-stream performance significantly.

3 Data

The clinical data used in this study originates from the Karolinska University Hospital and is stored in the research infrastructure called Health Bank – Swedish Health Record Research Bank[2] (Dalianis et al., 2015). The data encompasses 2 million patient records[3].

Three clinical data set have been de-identified using the BERT model created by Lamproudis. et al. (2022). The experiments fine-tune models using both the unaltered data and the de-identified data. The following datasets were used:

Stockholm EPR Gastro ICD-10 Corpus A corpus of 795,839 tokens in 6,062 discharge summaries encompassing 4,985 unique patients with gastrointestinal diseases. Each discharge summary is associated with multiple ICD-10 codes which have been divided into blocks. The dataset is a described in Remmer et al. (2021). The task is to predict the correct ICD-10 block for each discharge summary.

Stockholm EPR Diagnosis Factuality Corpus A corpus of 240,000 tokens in 3,710 clinical notes and their diagnosis. Each note has been annotated with the factuality of the diagnosis. There are six levels of factuality for each diagnosis: *Certainly Positive*, *Probably Positive*, *Possibly Positive*, *Possibly Negative*, *Probably Negative*, and *Certainly Negative*. The dataset is a described in (Velupillai, 2011) and (Velupillai et al., 2011). major is to process each clinical note and predict the degree of factuality for the diagnosis.

Stockholm EPR Clinical Entity Corpus A corpus consisting of 70,852 tokens in which 7,946 entities have been annotated. The annotations are for four clinical entity classes: *Diagnosis*, *Drugs*, *Body parts*, and *Findings*. The dataset is a described in (Skeppstedt et al., 2014). The task is an NER problem which requires the model to locate the entities within each sample.

These three datasets can be divided into two categories. The *Stockholm EPR Gastro ICD-10 Corpus* and *Stockholm EPR Diagnosis Factuality Corpus* are sequence classification problems with labels on the sample level. On the other hand, the *Stockholm EPR Clinical Entity Corpus* is a token classification problem which requires the model to assign

[1] A quasi-identifier is a identifier that indirectly can identify a patient such an street name or a zip code.

[2] Health Bank, `http://www.dsv.su.se/healthbank`

[3] This research has been approved by the Swedish Ethical Review Authority under permission no. 2019-05679.

the correct label to each token in a sequence. This dataset was also used by Berg et al. (2020) which makes it possible to assess whether switching to a BERT-based approach affects the results.

A fourth dataset was used for the qualitative experiments:

Gastro Pseudo Clinical Dataset Based on the Stockholm EPR Gastro ICD-10 Corpus (Remmer et al., 2021), this dataset contains 6,062 de-identified discharge summaries in the medical speciality of Gastrointestinal diseases. The data have been de-identified and pseudonymized using the HB Deid CRF.

4 Methods & Experiments

4.1 De-Identification

This study uses two versions of the de-identification system HB Deid described in Berg et al. (2019). The first version uses Conditional Random Fields (CRF) to locate sensitive entities.

The second version instead uses a clinical BERT model fine-tuned for Named Entity Recognition (NER). This model is pre-trained on Swedish general-domain data (Malmsten et al., 2020) and adapted to the clinical domain. This adaptation involved changing the vocabulary and continuing the pre-training using sensitive clinical data (Lamproudis. et al., 2022).

PHI Class	Recall	Precision
Age	100%	100%
First Name	100%	100%
Last Name	98%	98%
Partial Date	99%	97%
Full Date	90%	91%
Phone Number	81%	68%
Health Care Unit	85%	94%
Location	100%	100%
Organization	71%	100%

Table 1: The NER model's recall and precision for each PHI type are displayed and were calculated on the gold standard called *Stockholm EPR PHI Corpus*, (Dalianis and Velupillai, 2010). For details on the annotation process see (Velupillai et al., 2009).

The BERT model from Lamproudis. et al. (2022) selected as it was the best model available for Swedish clinical NER. We can call it. This model was fine-tuned for NER using the Stockholm EPR PHI Corpus (Velupillai et al., 2009). The precision and recall for each PHI class, estimated using a held-out dataset, is shown in Table 1. Both the precision and the recall values are high for many of the classes. It correctly identifies most names, but struggles with detecting organizations. We call this fine-tuned model *SweClin-BERT NER*.

4.2 Comparing Fine-Tuned BERT Models

This experiment quantitatively evaluated the performance on down-stream tasks when using de-identified or unaltered training data. First, each dataset was processed using the BERT-based de-identifier to detect all sensitive PHI entities. These entities were replaced with realistic surrogates.

The resulting collection of datasets was used to create two different classes of models. One was trained only using pseudonymized datasets and the other was trained using unaltered datasets. The models are trained using 10-fold cross validation and are compared by studying their F_1 scores.

Table 2 shows the results on the three tasks described in section 3. The performance of models trained using pseudonymized data is indistinguishable from the performance of models trained using real data.

This lack of difference was confirmed by performing Wilcoxon rank-sum tests (Mann and Whitney, 1947) on the folds of each task. None of the tests found any statistically significant differences between models trained on real or pseudonymized data.

4.3 Qualitative study I

This first qualitative study involved two human evaluators: One coordinating officer (a) that decides on the exportation of clinical data for research as well as a chief physician (b) who also is responsible for deciding on the exportation of clinical data for research. Both evaluators work at Region Stockholm county council in Sweden.

The requirements of the de-identification of the Gastro dataset was preceded by a discussion with the two human evaluators as well as a lawyer also working with exportation of clinical data. They decided that entities classed as *First Name, Last Name, Location, Phone Number, Organization* and *Social Security Number* are sensitive and should be removed from the patient record. They also decided that entities classified as *Age, Health Care Unit, Full Date* and *Date Part* were *not* sensitive

Data version	**ICD-10**	**Factuality**	**Clinical Entity**
	Classification	**Classification**	**NER**
Unaltered	0.86	0.74	0.85
Pseudonymized	0.86	0.75	0.86

Table 2: Models were trained on both pseudonymized and unaltered versions of each dataset. The average F_1 score of each model class on the held-out dataset is shown for each of the tasks.

and should be retained. Hence only six classes were used for de-identification.

The three experts also decided that they did not want to pseudonymize the text with surrogates. Instead, the sensitive entities were replaced with their class names. A control data set was also created using all ten classes for de-identification. Both de-identifications were done using HB Deid CRF.

Evaluator (a) read 100 pages each of the two sets of de-identified files. They found the word "Inga" as "No" in English was tagged as First Name. A similar pattern was found for the personal names "Per", "Tages" which can also mean, "Per day" or "Take" respectively. The evaluator also noticed that locations such as country names were removed, but not a patient's nationality. They also noticed that the span of the predicted entity did not always cover the whole PHI expression, especially when they were multi-word expression. *Evaluator (a)* found also that the set with 10 types of class tags was easier to read that the one with 6 types of class tags.

Evaluator (b) read 100 pages of the original file (that did not contain any class tags) by mistake instead of reading the de-identified version and they commented that they did not find much sensitive information. They said that the de-identification system managed to effectively replace the sensitive information. *Evaluator (b)* thought they found some problematic cases where, for example, names and locations were incorrectly assumed to be de-identified. All these cases were double checked and we confirmed that these were correctly tagged in the de-identified data set.

4.4 Qualitative study II

A second qualitative study was carried out where the version of HB Deid described in (Berg and Dalianis, 2021) was used. 100 patient patient records from an emergency unit at Skåne University Hospital were used for the de-identification experiment.

The evaluator was an computer scientist at Lund University at the Faculty of Engineering. Snippets of 50-200 words were extracted from each patient record and processed using HB Deid CRF. The de-identification system found a large number PHI tokens and classified most of them correctly. However, many abbreviations were incorrectly classified as *Organization*s. Generally, there where a few false positives and some false negatives, but these misclassifications alone did not provide enough information to reveal the identity of any patients.

5 Discussion

The human evaluators in this study disliked the idea of replacing PHI with surrogate values. Instead, they preferred replacing sensitive entities with their PHI class. However, the resulting dataset not only fails to use the protecting effects of HIPS (Hiding In Plain Site) (Carrell et al., 2019). It also makes it obvious to an adversary that any PHI they encounter is in fact a real PHI that the system failed to replace.

Human evaluations also uncover the problem of dealing with abbreviations, since they can be mistaken for organizations. This can be dealt with by adding word lists to deal specifically with abbreviations.

The human evaluators did not find the pseudonymized text difficult to read. On the contrary, one evaluator had difficulties distinguishing between real and pseudonymized data. This indicates that pseudonymized health records retain much of their utility for non-NLP research.

Comparing the utility of pseudonymized and real datasets, we find no harmful effects from de-identification on down-stream performance. This confirms and builds upon previous results in (Berg et al., 2020) that also showed that utility was retained after de-identification.

6 Conclusion

De-identification works best when the underlying NER classifier has both high precision and high recall. A high recall is crucial to ensure that as many PHI as possible are detected. At the same time, having a low precision may introduce noise into the data which can harm its utility.

In this study, we show that existing de-identification systems can effectively be used to make datasets safer. We also show that the noise introduced in this process does not harm downstream performance in clinical NLP tasks.

The qualitative evaluations also show that humans have trouble distinguishing between pseudonymized and real data. We also uncover a discrepancy between the NLP community and our human evaluators regarding the perceived value of hiding sensitive data using HIPS.

Acknowledgements

We would like to thank Sonja Remmer for introducing us to the Wilcoxon rank-sums test as a tool for comparing machine learning models. Thanks also to Anders Björkelund for evaluating the results from HB Deid at Lunds university.

We are also grateful to the DataLEASH project for funding this research work.

References

Hanna Berg, Taridzo Chomutare, and Hercules Dalianis. 2019. Building a De-identification System for Real Swedish Clinical Text Using Pseudonymised Clinical Text. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 118–125, Hong Kong. Association for Computational Linguistics.

Hanna Berg and Hercules Dalianis. 2021. HB Deid-HB De-identification tool demonstrator. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 467–471.

Hanna Berg, Aron Henriksson, and Hercules Dalianis. 2020. The Impact of De-identification on Downstream Named Entity Recognition in Clinical Text. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, Louhi 2020, in conjunction with EMNLP 2020*, pages 1–11.

David S Carrell, David J Cronkite, Muqun (Rachel) Li, Steve Nyemba, Bradley A Malin, John S Aberdeen, and Lynette Hirschman. 2019. The machine giveth and the machine taketh away: a parrot attack on clinical text deidentified with hiding in plain sight. *Journal of the American Medical Informatics Association*, 26(12):1536–1544.

Hercules Dalianis. 2019. Pseudonymisation of Swedish electronic patient records using a rule-based approach. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 16–23, Turku, Finland. Linköping Electronic Press.

Hercules Dalianis, Aron Henriksson, Maria Kvist, Sumithra Velupillai, and Rebecka Weegar. 2015. HEALTH BANK- A Workbench for Data Science Applications in Healthcare. *CEUR Workshop Proceedings Industry Track Workshop*, pages 1–18.

Hercules Dalianis and Sumithra Velupillai. 2010. De-identifying swedish clinical text-refinement of a gold standard and experiments with conditional random fields. *Journal of biomedical semantics*, 1(1):1–10.

Christian Dalton-Locke, Johan H Thygesen, Nomi Werbeloff, David Osborn, and Helen Killaspy. 2020. Using de-identified electronic health records to research mental health supported housing services: A feasibility study. *PloS one*, 15(8):e0237664.

Andrea C Fernandes, Danielle Cloete, Matthew Broadbent, Richard D Hayes, Chin-Kuo Chang, Richard G Jackson, Angus Roberts, Jason Tsang, Murat Soncul, Jennifer Liebscher, et al. 2013. Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records. *BMC medical informatics and decision making*, 13(1):1–14.

Anastasios Lamproudis., Aron Henriksson., and Hercules Dalianis. 2022. Vocabulary modifications for domain-adaptive pretraining of clinical language models. In *Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies - HEALTHINF,*, pages 180–188. INSTICC, SciTePress.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with Words at the National Library of Sweden – Making a Swedish BERT. *arXiv:2007.01658 [cs]*. ArXiv: 2007.01658.

Henry B. Mann and D. Ransom Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50–60. Publisher: Institute of Mathematical Statistics.

Stéphane M. Meystre, Shuying Shen, Deborah Hofmann, and Adi V. Gundlapalli. 2014. Can physicians recognize their own patients in de-identified notes? In *MIE-Medical Informatics Europe*, pages 778–782.

Kostas Pantazos, Soren Lauesen, and Soren Lippert. 2017. Preserving medical correctness, readability and consistency in de-identified health records. *Health informatics journal*, 23(4):291–303.

Sonja Remmer, Anastasios Lamproudis, and Hercules Dalianis. 2021. Multi-label Diagnosis Classification of Swedish Discharge Summaries – ICD-10 Code Assignment Using KB-BERT. In *Proceedings of RANLP 2021: Recent Advances in Natural Language*

Processing, RANLP 2021, 1-3 Sept 2021, Varna, Bulgaria, pages 1158–1166.

David Sánchez, Montserrat Batet, and Alexandre Viejo. 2014. Utility-preserving privacy protection of textual healthcare documents. *Journal of biomedical informatics*, 52:189–198.

Maria Skeppstedt, Maria Kvist, Gunnar H Nilsson, and Hercules Dalianis. 2014. 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.

Sumithra Velupillai. 2011. Automatic classification of factuality levels: A case study on Swedish diagnoses and the impact of local context. In *Fourth International Symposium on Languages in Biology and Medicine, LBM 2011*.

Sumithra Velupillai, Hercules Dalianis, Martin Hassel, and Gunnar H Nilsson. 2009. Developing a standard for de-identifying electronic patient records written in Swedish: precision, recall and F-measure in a manual and computerized annotation trial. *International Journal of Medical Informatics*, 78(12):e19–e26.

Sumithra Velupillai, Hercules Dalianis, and Maria Kvist. 2011. Factuality levels of diagnoses in Swedish clinical text. In *User Centred Networked Health Care*, pages 559–563. IOS Press.