Pseudonymisation of Personal Names and other PHIs in an Annotated Clinical Swedish Corpus

Alyaa Alfalahi, Sara Brissman, Hercules Dalianis
Department of Computer and Systems Sciences (DSV)
Stockholm University
Forum 100, 164 40 Kista, Sweden
E-mail: alyalfa@dsv.su.se, sarabri@dsv.su.se, hercules@dsv.su.se

Abstract

Today a large number of patient records are produced and these records contain valuable information, often in free text, about the medical treatment of patients. Since these records contain information that can reveal the identity of patients, known as protected health information (PHI), the records cannot easily be made available for the research community. In this research we have used a PHI annotated clinical corpora, written in Swedish, that we have pseudonymised. Pseudonymisation means to replace the sensitive information with fictive information for example real personal names are replaced with fictive personal names based on the gender of the real names and family relations. We have evaluated our results and our five respondents of who three were clinicians found that the clinical text looks real and is readable. We have also added pseudonymisation for telephone numbers, locations, health care units, dates and ages. In this paper we also present the entire de-identification and pseudonymisation process of a sample clinical text.

Keywords: Protected Health Information PHI, Electronic Patient Records EPRs, De-identification, Pseudonym, Swedish.

1. Introduction

Electronic patient records, EPRs, include valuable information about the treatment of the patient. Patient records also often contain sensitive information regarding the situation of the patient which may disclose private information about the patient; i.e. protected health information (PHI) such as names, locations, health care units, phone numbers, etc. (HIPAA 2003). Patient records include information about the patient such as their social situation, health history, symptoms, and previous diagnoses and planned treatment. Most of this information is presented in the unstructured free text (Dalianis et al, 2009) which can be extremely useful for the clinical researcher, for hospital management and for educational purposes. In Sweden, all research that deals with patients and data about patients requires permission for use from ethics committees, specifically regional Ethics Committees (Lag, 2003:460). The assumption is therefore that sensitive information must be removed from the records before EPRs can be made freely available for research, so the task of the de-identifying PHI instances is both important and difficult.

In Meystre et al. (2010) there is an overview of the different de-identification approaches for EPRs written in English. Most of the researchers have applied de-identification methods by identifying PHI instances and annotating them with PHI classes, e.g *First_Name, Last_Name,Health_Care_Unit,Location, Phone_Number,* etc. However, the records produced with these annotation classes are not easy to read as plain text. Similarly, the output text will be less readable if the PHI instances, such as personal names are replaced with ID numbers. As well as first and last names, there also phone numbers, locations, health care units, dates and ages that need to be pseudonymised and made readable. Some questions that arise involve how to replace the personal names and make the text coherent with respect to, for example, gender or
family relations, phone numbers that are realistic, locations that are geographically correct but not real, replacing ages without making patients too old or too young and finally changing times so they are realistic to weekends, seasons and public holidays. In this paper we will focus on first and last names.

2. Related Research

Meystre et al. (2010) reviewed different de-identification approaches for EPRs written in English, but they did not mention pseudonymisation. Pseudonymisation algorithms have, however, been mentioned in Sweeney (1996), Douglass et al. (2004), Pestian et al. (2005) and Neamatullah et al. (2008), who have all worked with EPRs written in English. Furthermore, Pantazos et al. (2011) focused on a de-identification algorithm for a Danish database and generated a new version by replacing real data with other new data. However, their algorithm has not been developed to handle unknown, misspelled names, or names that can be used for both genders.

Generally speaking there are few clinical corpora available for research in English and Finnish respectively.

- I2B2[1] corpus, the Informatics for integrating biology and the bedside (i2b2 2008) centre has created a clinical English corpus which consists of approximately 1,000 notes that is available for researchers after signing an agreement.
- The CMC[2] is one of the clinical corpora which has been analysed by Pestian et al. (2007) and

[1] http://www.i2b2.org
[2] http://www.inf.u-szeged.hu/rgai/bioscope

has been made public for study purposes.

- The De-id[3] corpus, Neamatullah et al. (2008), have published a clinical corpus in English. The De-id corpus consists of 412,509 nursing notes and 1,934 discharge summaries and is publicly available.
- A clinical Finnish corpus contains 2800 sentences (17,000 tokens) of nursing notes which have been manually anonymised by removing or changing all name. Furthermore, this corpus has been developed and published by Haverinen et al. (2010).

In Velupillai et al. (2009) an annotation process for de-identification is described. The annotation was carried out by three annotators, (one junior, one senior computer scientist and one senior physician) and gave rise to three sets of annotated data. Dalianis & Velupillai (2010) created a consensus of the annotated data in Velupillai et al. (2009) described with the de-identification system for Swedish clinical texts. The consensus is called the Stockholm EPR PHI Corpus, and contains 100 patient records with an equal distribution of male and female patients. These records were compiled from five clinics: neurology, orthopaedic, oral and maxillofacial surgery, an infection clinic and a dietetics clinic. However the Stockholm EPR PHI Corpus has not yet been pseudonymised. Another completely different approach is described by Dalianis & Boström (2012), who suggest releasing the Stockholm EPR PHI Corpus for research into different de-identification methods, augmenting the corpora with a large feature set but without giving access to the actual words. A non-worded corpora could be used for a set of machine learning experiments but unfortunately the feature set needs to be further developed.

3. Method

Our approach is to replace the PHI instances with new realistic instances. We call this process pseudonymisation. This is replacing real PHI instances in the clinical text with pseudonyms, surrogates or what we call fictive names. An automated algorithm (Pseudonoma) will be developed to replace the annotated first/last names in clinical free text notes with fictive names depending on the gender of the patient and the referential structure. Moreover, the algorithm has also been expanded to handle unknown and misspelled annotated personal names by continually checking the name lists. The algorithm will be executed on training (development) data and test data respectively. Finally we will also add pseudonymisation for *Phone_Number, Location, Health_Care_Unit, Full_Date, Part_Date* and *Age*.

The constructed pseudonymisation system (Pseudonoma) is a rule-based system with name lists. Pseudonoma is implemented partly in the Perl programming language, and consists of two algorithms, first name and last name algorithms, to replace real names with other fictive names. 'Name' could refer to patient's name, names of patient's relatives and health staff names. The other part of the Pseudonoma, for phone numbers, locations, health care units, dates and ages, was developed in Python and Excel script language. Pseudonoma has been tested on the Stockholm EPR PHI Corpus[4] (380,000 tokens) which is a subset of the Stockholm EPR Corpus (Dalianis et al, 2009).

The Stockholm EPR PHI Corpus is our input file to the first name algorithm, which includes annotated names with tags, such as <*First_Name*> <*/First_Name*> and <*Last_Name*> <*/Last_Name*>. The name lists have been created by retrieving personal names from Swedish names lists (Swedish names 2009) and have been manually checked regarding the gender of names. These name lists consist of a list of female first names (173 names), a list of male first names (114 names), and a list of last names (368 names). The programme contains several hash tables that store the processed names, inspired by Sweeney (1996). The output file from the first name algorithm is the input file to the last name algorithm. The final output file is our pseudonymised text.

We observed that annotated personal names in genitive form, misspelled or unknown personal names could cause problemd for our pseudonymisation algorithm and therefore negatively influence the readability and consistency of the pseudonymised text. Therefore, the algorithm (Pseudonoma) was improved to solve these problems. The algorithm was developed to handle misspelled names through the usage of edit distance (Levenshtein Distance[5]). The algorithm was also adjusted to deal with the genitive form and with typical Swedish name combinations such as *Anna-Lena, Eva-Britt, etc.* Also, unknown and gender- neutral names have been replaced with other gender- neutral names such as *Kim* and *Denis,* for details see (Alfalahi 2011).

4. Evaluation and Result

To evaluate Pseudonoma we have used two corpora. Firstly, the Stockholm EPR PHI Corpus, which consists of 100 patient records, has been used as training (or development) data for Pseudonoma. The 100 patient records have previously been manually annotated (Velupillai et al. 2009). Secondly, a new extract from the Stockholm EPR Corpus that also includes 100 patient records has been used as test data which has been annotated automatically by applying the Stanford CRF NER programme (Dalianis & Velupillai 2010). We have executed the Pseudomona on both above mentioned corpora. In Figure 1 we can see the complete automated

[3] http://www.physionet.org/physiotools/deid

[4] The study was carried out after approval from the Regional Ethical Review Board in Stockholm, permission number 2009/1742-31/5.

[5] http://en.wikibooks.org/wiki/Algorithm_implementation/Strings/Levenshtein_distance#Perl

de-identification and pseudonymisation process: firstly the clinical text annotated with respect to PHI by the Stanford CRF NER, trained on the Stockholm EPR PHI Corpus, and then the annotated text pseudonymised by Pseudonoma.

The outcome is pseudonymised patient records that contain fictive names. 14 records were chosen from the Stockholm EPR PHI Corpus and 12 records were selected from a new extract from the Stockholm EPR Corpus for manual evaluation by five respondents, three of whom were clinicians. In Figure 2 we can see the distribution of names and tokens in our data.

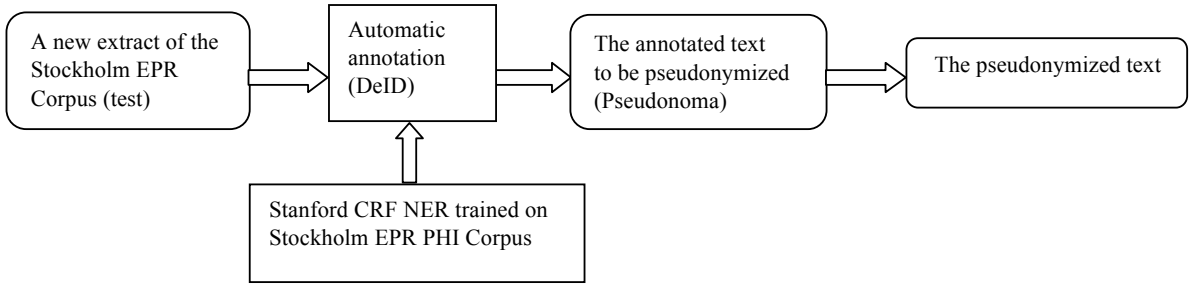

Figure 1: The complete de-identification and pseudonymisation process using Stanford NER CRF and Pseudonoma

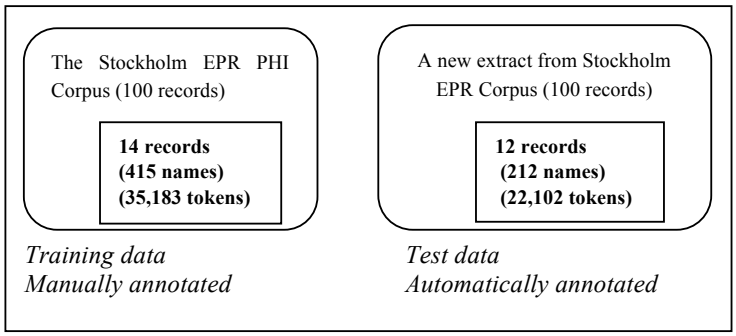

Figure 2: Comparison of the two evaluation sets and their distribution of names

Questions	Total names	Yes (%)	No (%)
Gender determination[6]	97	95 (98)	2 (2)
Gender agreement[7]	97	97 (100)	0 (0)
Difference between first and last name	212	212 (100)	0 (0)
Pseudonymisation	212	212 (100)	0 (0)
Repetition of the same fictive name	212	212 (100)	0 (0)
First/ last name tags are replaced	212	212 (100)	0 (0)

Table 3: Answers to the questionnaire (questions 1-6) on test data (212 first and last names, 97 first names)

[6] The fictive names have the right gender according to the text.

[7] The fictive name has the same gender as the real name, i.e. if the real name has a female gender so a female name must be chosen as a fictive name.

We created a questionnaire with 11 questions to be answered by each respondent. Two types of results were obtained by applying the questionnaire to the patient records (the original and the pseudonymised records): the results according to the number of records in both corpora (training and test data) and the results according to the number of annotated names in both corpora. Table 3 illustrates the responses to six questions depending on the number of annotated names in the records.

The aim of our questionnaire was to evaluate six different values that reflect the quality of the text. The *gender determination* of the fictive names means that the fictive names have a clear gender so it is easy to distinguish between female and male fictive names in the output text. G*ender agreement* means that if the real name is female in gender (Eva) so a female name must be chosen as the fictive name (Sara). The *difference between first and last name* means that the respondent can specify the first and the last names from the correct selection of fictive names such as *Johan* for first name and *Johansson* for last name. Another evaluation point that has been added to the questionnaire concerns the replacement (*pseudonymisation*) of the real name with the fictive and whether all real names have been replaced with fictive names. A further evaluation point in the questionnaire relates to the *repetition of the same fictive name* in the text, specifically, whether the real name has been replaced with the same fictive name each time the real name appears in the text. For example, the real name *Erik* should have the same fictive name *Tomas* whenever *Erik* is repeated in the text. The last question concerns left over tags *(<First_Name> </First_Name>,<Last_Name> </Last_Name>*) in the patient records.

Our goal was to develop an automated algorithm which can correctly replace all annotated real names with other real names (100 per cent) in patient records. The algorithm developed was tested on the Stockholm EPR PHI Corpus (the development or training data) and on a new extract from Stockholm EPR Corpus as test data.

The questionnaire analysis shows that the main goal is achieved by correctly replacing all annotated names with other realistic names. The automated algorithm depends exclusively on the annotation process. There were two un-annotated names in the chosen records (14 records, 415 names) from the Stockholm EPR PHI Corpus (training data) and 16 un-annotated names in the Stockholm EPR Corpus (test data) (12 records, 212 names), and so these un-annotated names were not replaced by the algorithm. This made the training text slightly more readable and coherent than the test text.

The questionnaire analysis illustrates that selection of the right gender during the replacement process is achieved to a high percentage in training and in the test corpora, 99 and 98 per cent respectively. Furthermore, the genitive form, and misspellings have also obtained a high percentage of accuracy in both corpora. The question about genitive form includes a check of the genitive *s* in the fictive name if the real name takes the form of genitive *s*. For example, if the real name is in the genitive form i.e. *Eva's* mother (Evas mor), then the fictive name for *Eva*, i.e. (*Karin*) must definitively have the genitive *s*, *Karin's* mother (Karins mor). Another question tests whether the misspelling of real names has been handled by the correction technique, which has been improved in the algorithm. Additionally, the processing of Swedish characters (*ääö ÄÅÖ*) is not standard in Perl language so the algorithm has been developed to handle these types of characters which can occur in names such as *Märta*, *Håkan*, *Göran*, Åsa, etc. This handling of language-specific problems obtained a high percentage of accuracy in both corpora.

We continued the pseudonymisation work by pseudonymising Phone*_Number, Location, Health_Care_Unit, Full_Date, Part_Date* and *Age*. Altogether we pseudonymised 4421 instances in our corpus distributed over the classes described in Table 4.

The second part of Pseudonoma was developed in Python and Excel scripts except for one section regarding ages, which was manual as the number of ages were few. All dates were shifted by an unknown, arbitrary number of days and months respectively. *Phone_Number* was pseudonymised except for the area code and finally *Location* and *Health_Care_Unit* was assigned to the default location Stockholm and default health care unit Solvillan respectively, which was a naive approach. Age was manually shifted by an unknown arbitrary number of years except for *Age over 89* years which was shifted to Age over 89. Please see Figure 5 for an example of this.

Annotation class	**Instances**
Age	56
Full_Date	710
Date_Part	500
First_Name	923
Last_Name	928
Location	1 021
Health_Care_Unit	148
Phone_Number	135
Sum	4 421

Table 4: The distribution of annotation classes and instances of pseudonymised annotation

<Age>53-årig</Age> kvinna, välkänd på kliniken. Går hos <First_Name> Åsa</First_Name> <Last_Name>Lindqvist</Last_Name> samt på smärtmottagningen. Har en kronisk huvudvärk utan säker genes. Insatt på Metadon, Actiqe och Stesolid. Sökte den <Date_Part>8/8</Date_Part> pga ohållbar situation med bristfällig smärtkontroll. Pat är frusterad över lång väntetid på inneliggande utsättning av opiater som skulle göras via IVA och planerats av dr <First_Name>Emil</First_Name> <Last_Name>Engström</Last_Name>. Pat kommer till <Health_Care_Unit>Solvillan </Health_Care_Unit> och kräver att få läggas in på IVA och hotar att sluta med samtliga mediciner. Pat har haft flera samtal med PAL på <Health_Care_Unit> Solvillan</Health_Care_Unit>, <First_Name>Åsa</First_Name> <Last_Name>Lindqvist </Last_Name>. Hänvisar till tidigare anteckningar.

Figure 5. Example of pseudonymised Swedish clinical text by Pseudonoma, where all annotated instances have been replaced by pseudonymised instances. (In the real output text the annotation tags are, of course, removed).

All our data which could reveal a patient's identity, such as corpora, name lists and hash tables, is stored encrypted.

5. Conclusions and Future Work

The main contribution of this paper is that pseudonymisation, i.e. replacement of real names with fictive names in electronic patient records written in Swedish with an automated algorithm, is possible to a high quality (100%). Maintenance of the patient's gender and family relationships makes the text readable and coherent to a high quality (100%). The process of pseudonymisation is exclusively dependent on the quality of annotation of the text, whether manual or automatic. To the best of our knowledge this is the first algorithm that has been developed to automatically replace all real names with other real names in Swedish clinical text, as well as for phone numbers, location, health care units, dates and ages. The pseudonymised text may additionally be used for medical educational purposes or the development of new tools, and it is therefore important that the text maintains the readability. We believe the algorithm for pseudonymisation can be easily adapted to other languages, one only need change the name lists to the local language.

We have also in this paper showed the complete de-identification process of a sample clinical text (a new extract) using a machine learning system Stanford NER CRF trained on the manually annotated Stockholm EPR PHI Corpora to de-identify (annotate the PHI) the sample text, followed by the pseudonymisation of the sample clinical text using Pseudonoma, (see Figure 1).

In the future we would like to extend the location and health care unit pseudonymisation to select similar general locations and health care units to those written in the patient record. One issue that arises is that if one replaces health care units one may miss important information about diseases. Date shifts are also sensitive, cannot be completely randomized and have to be consistent; one needs to consider that weekends, public holidays and the seasons are different in respect of health care, as during weekends and public holidays there are fewer health care personnel on duty and some diseases are seasonally dependent.

In the future we plan to add automatic age replacement, and to evaluate the performance and quality of the pseudonymisation of phone numbers, locations, health care units and ages. We also plan to manually re-read the entire corpus one more time to be sure that we have not missed annotations for any PHI or pseudonymising any PHI. We have also previously tested Stanford NER CRF trained on the same corpus and found another 49 false positives that have been annotated (Dalianis & Velupillai 2010).

We are currently in the process of applying for ethical permission to release this pseudonymised variant of the Stockholm EPR PHI Corpus, which we call the Stockholm EPR PHI Pseudo Corpus.

6. Acknowledgements

This work was partly supported by the project High-Performance Data Mining for Drug Effect Detection at Stockholm University, funded by the Swedish Foundation for Strategic Research under grant IIS11-0053.

7. References

Alfalahi, A., (2011). Pseudonymization of person names in an annotated clinical Swedish corpus, Master thesis, Department of Computer and Systems Sciences (DSV) KTH/Stockholm University. Internet: https://daisy.dsv.su.se/fil/visa?id=62734

Dalianis H. and Boström, H. (2012). Releasing a Swedish clinical corpus after removing all words - de-identification experiments with conditional random fields and random forests, in Proceedings of The Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012) held in conjunction with LREC 2012, May 26, Istanbul.

Dalianis, H., Hassel, M. and Velupillai, S. (2009). The Stockholm EPR Corpus - Characteristics and some initial findings, Proceedings of ISHIMR (2009), Evaluation and implementation of e-health and health information initiatives: International perspectives, 14th International Symposium for Health Information Management Research. Kalmar, Sweden, 14-16 October, pp. 243-249.

Dalianis, H. and Velupillai, S. (2010). De-identifying Swedish Clinical Text - Refinement of a Gold Standard and Experiments with Conditional Random Fields, Journal of Biomedical Semantics, 1:6 (12 April 2010).

Douglass, M., Clifford, G., Reisner, A., Moody, G. and Mark, R. (2004). Computer-assisted de-identification of free text in the MIMIC II database, Computers in Cardiology 31: 341–344. Internet: http://mimic.mit.edu/Archive/Publications/Douglass04.pdf.

Haverinen, K., Ginter, F., Laippala, V., Viljanen, T. and Salakoski, T., (2010). Dependency- based PropBanking of clinical Finnish, In Proceedings of The Fourth Linguistic Annotation Workshop (LAW IV) held at ACL2010, Uppsala, Sweden. Internet: http://bionlp.utu.fi/clinicalcorpus.html

HIPAA (2003). Health insurance portability and accountability (HIPAA), privacy rule and public health guidance, From CDC and the U.S. Department of Health and Human Services, http://www.cdc.gov/mmwr/preview/mmwrhtml/m2e411a1.htm.

I2b2, (2008) Informatics for integrating biology and the bedside, Internet: http://www.i2b2.org.

Lag (2003:460) om etikprövning av forskning som avser människor (SFS). Stockholm: Utbildningsdepartementet. (In Swedish, Law (2003:460) Ethical review regarding research considering humans), Internet: http://www.notisum.se/rnp/sls/lag/20030460.HTM

Meystre, S., Friedlin, F., South, B., Shen, S. and Samore, M. (2010). Automatic de-identification of textual documents in the electronic health record: a review of recent research, BMC Medical Research Methodology, 10:70.

Neamatullah, I., Douglass, M., Lehman, L., Reisner, A., Villarroel, M., Long, W., Szolovits, P., Moody, G., Mark, R. and. Clifford, G. (2008). Automated de-identification of free text medical records, BMC Medical Informatics and Decision Making, 8: 32. Internet: http://www.physionet.org/physiotools/deid

Pantazos, K., Lauesen. S. and Lippert, S. (2011). De-identifying an EHR Database – Anonymity, Correctness and Readability of the Medical Record, European Federation for Medical Informatics.

Pestian, J. P., Itert L., Andersen C. L. and Duch W. (2005). Preparing clinical text for use in biomedical research, Journal of Database Management, 17(2):1-12.

Pestian, J., Brew, C., Matykiewicz, P., Hovermale, D., Johnson, N., Cohen, K. and Duch, W. (2007). A shared task involving multi-label classification of clinical rree text, BioNLP : Biological, translational, and clinical language processing, pages 113–120. Prague, June, Association for Computational Linguistics. Internet: http://www.inf.u-szeged.hu/rgai/bioscope

Swedish names, in Swedish (2009). Internet http://svenskanamn.se

Sweeney, L. (1996). Replacing personally-identifying information in medical records, the Scrub system, Proceedings, Journal of the American Medical Informatics Association, pp. 333-337.

Velupillai, S., Dalianis, H., Hassel, M. and Nilsson, G. (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, e19-e26.