Pseudonymisation of Swedish Electronic Patient Records Using a Rule-based Approach

Hercules Dalianis
Department of Computer
and Systems Sciences
Stockholm University
hercules@dsv.su.se

Abstract

This study describes a rule-based pseudonymisation system for Swedish clinical text and its evaluation. The pseudonymisation system replaces already tagged Protected Health Information (PHI) with realistic surrogates. There are eight types of manually annotated PHIs in the electronic patient records; personal first and last names, phone numbers, locations, dates, ages and healthcare units.

Two evaluators, both computer scientists, one junior and one senior, evaluated whether a set of 98 electronic patients records where pseudonymised or not. Only 3.5 percent of the records were correctly judged as pseudonymised and 1.5 percent of the real ones were wrongly judged as pseudo, giving that in average 91 percent of the pseudonymised records were judged as real.

1 Introduction

Electronic patient records also called clinical text contain valuable information that may be extracted and used for improving healthcare, see Chapter 10 in (Dalianis, 2018).

The records are becoming more and more accessible for the research community, but under strict confidential restrictions since they contain sensitive information about patients. Before being accessible for research the electronic patient records are required to be anonymised in the way that they do not contain any information (or data tables) that may identify any patient. However in the unstructured part of the records, that is the free text fields, there is information such as personal names, phone numbers that may identify the patient. In the structured part there might also be sensitive information, but that is easily identifiable, since that column can be called social security number, temperature, or ICD-10-code.

Therefore, there is a significant research area in clinical text mining called automatic de-identification (DEID) of electronic patient records (Meystre et al., 2010; Uzuner et al., 2007). These DEID systems are either rule-based, machine learning-based or hybrid approaches where the best systems obtain up to 0.97 F-score, (Uzuner et al., 2007).

Of course, one requirement on these DEID systems is high recall over high precision since it is more important to find all instances of sensitive information than to risk predicting false positives.

A DEID system works in such a way that it tags the identified sensitive information or what more precisely is called Protected Health Information (PHI) and removes the sensitive information inside the PHI tag. The tag is left telling what type of PHI it contained. When the system misses identifying an entity in the text, such as a personal name or a phone number, it is visible and obvious. These un-identified PHIs are also called residual identifiers. One method to increase security is described by Carrell et al. (2013) and is called Hiding In Plain Sight (HIPS) and consists of replacing all the identified PHI tags with surrogates or pseudonymised information, such that the un-identified residual PHI will be perceived by the reader to be already replaced by a surrogate and hence pseudonymised. Pseudonymisation is the method where an identified PHI, for example, a personal first name is replaced with a fake first name, that obviously need to be a common name to diminish the risk of identification.

An example of a pseudonymised electronic patient record is presented in Figure 1.

A hypothesis is that when reading the patient records, the reader should not be annoyed by strange names or places or tags and focus on the

Proceedings of the Workshop on NLP and Pseudonymisation, pages 16–23,
Turku, Finland, 30 September 2019.

medical content.

The research question is whether a reader of an electronic patient record could reveal if one record as pseudonymised or not. The reader/evaluator should judge whether a clinical text describing a patient's personal name, family relation and mentioned addresses, phone numbers, locations, health care units and dates look real or not.

2 Related research

One of the first attempt in creating surrogates after the DEID process was presented by Sweeney (1996). Dates were replaced with a similar date nearby. Personal names were replaced with a fictitious unique name that sounded reasonable. The article does not mention how the system processed locations and phone numbers and other PHI.

In (Douglass et al., 2004), a similar approach is described where dates were shifted by the same random number of weeks or years, but keeping the days of the week. Personal names were replaced with names from a publicly available list from the Boston area in the US, but randomly mixing first and last names. Locations were replaced with randomly selected small towns. Hospital and clinical units were given fictitious names.

One of the first studies on Swedish was presented in (Alfalahi et al., 2012) and were carried out on the Stockholm EPR PHI Corpus (Velupillai et al., 2009), where personal names were replaced in a context-sensitive way. Female first names where replaced by common female first names, and similar for male names and for last names that also were replaced with common names. Gender-neutral first names are replaced with other randomly chosen gender-neutral name. Addresses and phone numbers are replaced, and dates are shifted, ages changed slightly. Locations and healthcare units are replaced with only one location and healthcare unit respectively.

Another study on the same Swedish corpus was carried out by Antfolk and Branting (2016). However the study focused only on locations. The system replaced locations such as *places, cities*, and *countries* with locations that were situated closely geographically. One problem was that many locations were misspellt or abbreviated and hence challenging to replace with a proper surrogate location. Prepositions in front of countries could pose problems since countries written in Swedish need the Swedish preposition *i* (English: in) while countries on islands require *på* (English: on). Complete addresses with street and number or cities were not replaced since they were not in the scope of the study.

Björkegren (2016) carried out a similar study on the same Swedish corpus and with focus on locations. The annotated class *Location* is a broad concept covering everything from a street, a place, a city, a municipality (Swe: kommun), a county, a country or a continent and sometimes an organisation, a company or a product name and thus difficult to process unless identifying what type of concept it is. The reason for this broad coverage is that the corpus was used for machine learning training and there were very few concepts for location in the corpus, hence several classes describing locations were collapsed into one class *Location* in the gold standard (Dalianis and Velupillai, 2010).

In the study by Björkegren (2016), an evaluation was carried out were three respondents had to evaluate which of the 17 patient records were pseudonymised and which contained real PHI. Half of the records were identified as pseudonymised thus indicating that the pseudonymisation program was not good enough. Many errors occurred in the geographical context where one street was mentioned in the wrong part of the city.

In another approach for English by Deleger et al. (2014), both the American English clinical corpus from Physionet, i2b2 and the Cincinnati Children's Hospital Medical Center (CCHMC) corpus were used. One important feature was when replacing one PHI in the data set it should not resemble any other replaced PHI. Personal names are replaced from a list of real names from US Census Bureau having a frequency above 144, meaning 0.004 percent of the data. Gender of personal names are replaced consistently. Combinations of street and street numbers are not reoccurring as in the original corpus. Email addresses are replaced with a set of a random set of characters as the length in the original email address.

Meystre et al. (2014) carried out a study where 86 patient records in English were de-identified and where none of the five treating physicians could recognise their patients after de-identification.

Grouin et al. (2015) carried out an experiment where they de-identified a group of patient records in French and they asked physicians to identify

the patient. However, they could not succeed in this, unless they had access to the whole hospital system and found other documents of the same (de-identified) patient, so they could group and regroup patients and consequently identify them.

3 Methods and data

The method chosen for this study is rule-based since training data is scarce in the clinical domain that can be used for replacing sensitive PHI with realistic surrogates.

The implementation is carried out in the programming language Python. Three personal name lists are used representing Swedish female first names, male first names and last names, also lists with candidate surrogates in form of the 100 most common (frequent) Swedish personal names were prepared: female first and male first names and last names (gender-neutral), moreover, a list of all streets and place names in the city of Stockholm and locations in Sweden.

A list of all postal codes in Sweden, (where the 461 most common are used for candidate surrogates), a list of all area codes for phone numbers as well as a list of all prefixes for mobile phone numbers in Sweden and finally a list of generic candidate healthcare units were also used. These lists were also used by Velupillai et al. (2009) to build a de-identification system and re-used in this study for the rule-based psuedonymisation system.

The authors of the article (Antfolk and Brantingh, 2016) kindly provided a hierarchical list with locations. Locations divided into the different Swedish counties and finally all locations in the world including islands in all continents of the world and capital cities except the locations in Sweden.

Special list for town squares and parks were provided by Andreas Amsenius that had used them for earlier work on de-identifications of emails in Swedish.

For an overview of all lists and number of entities, see Table 1.

Swedish entities	**Number**	**Most common**
Female first names	122,622	100
Male first names	120,167	100
Gender-neutral names	22	22
Last names	34,894	100
Health care units	430	20
Postal codes	9,724	461
Streets in Stockholm	1,719	-
Parks in Stockholm	131	-
Provinces (Landskap)	21	-
Places in provinces	27,883	-
Squares and places	6,910	-
Area code phone numbers	265	-
Prefix mobile phone numbers	17	-
Outside Sweden		
Continents	5	-
Countries	203	-
Capitals	206	-
Cities	1,386	-
Provinces	131	-

Table 1: Overview of all the types of lists used both to find PHIs and also to generate surrogates (Column most common). Observe that Provinces and Continents are hierarchical in one level; hence Provinces and Continents are not replaced, just the content within each group.

The used data consists of the Stockholm EPR PHI Corpus[1] that contains 100 electronic patient records written in Swedish from five different clinical units: *neurology, orthopaedia, infection, dental surgery* and *nutrition* (Velupillai et al., 2009; Dalianis and Velupillai, 2010), see Table 2 for the distribution of the eight types of PHI entities. The Stockholm EPR PHI Corpus is part of Health Bank - Swedish Health Record Research Bank.

The pseudonymisation program work in such a way that it matches the found tagged personal first name and last name. The first name is checked whether the name is in the list of male first names or female first names. If it is a male name, it is replaced with another common male name and if it is a female name, it is replaced with another com-

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

mon female name. First names are also replaced in such a way that if it is repeated several times in a paragraph, then the same generated pseudonym is kept. If the first name is not in any female or male name list it is spellchecked using a Levenshtein spell checking module implemented by Nick Sweeting in 2014[2] based on Peter Norvig's spell checker, but here in this study adapted for personal name correction, first using a male personal name dictionary and then a female name dictionary. Gender-neutral names are replaced with a gender neutral name, and also genitive "s" in names is always taken care of. For last names they are replaced with one common last name, no spell checking is used.

Street addresses are replaced with another random street address in Stockholm with a random street number, ditto postal number, and postal location. Locations outside of Stockholm are replaced according to a system that if a location in a specific county is found, it is replaced with another random location on the same county for geographical proximity, and when the county name itself is found it is not replaced since counties are considered as large geographical areas. The same goes for continents, a country in one continent is replaced with another random country but not the continent name itself.

Dates are shifted plus seven days upwards or downwards. Weekdays and weekends are kept intact, since the activities on clinical units are different on weekdays compared to weekends.

Ages are shifted a couple of years upwards or downwards.

Regarding phone numbers, the area code for fixed phone numbers are randomly shifted to another area code for fixed phone numbers as well as the whole number were randomly changed. For mobile phone numbers, the same procedure is carried out but where the prefix for mobile phone numbers is randomly shifted to another prefix for mobile phone numbers.

For healthcare units, they are randomly changed to some few generic healthcare units that cannot identify a specific healthcare unit, see also Table 1.

Any found social security number if found is simply removed.

The data to be evaluated were prepared as follows: The texts were randomly pseudonymised so half of the 98 record text were pseudonymised and the other half were real non-pseudonymised patient records. The random distribution became 58 true records and 40 pseudo records, hence 59 percent true records and 41 percent pseudo records, which as the result of the *random.choice()* function of Python.

To avoid forcing the evaluators to read texts with few or no PHIs the data were prepared in the following way: First by ordering the 98 records with the highest amount of PHIs first and with falling density and then by extracting a section of maximum 20 consecutive lines with the highest density of PHI. Before presenting the records to the evaluators, the tags marking up the PHI were removed.

The reason to give the evaluators only 20 consecutive lines with the highest density of PHI was to make the evaluation practical otherwise the evaluators had to read plenty of clinical text (with no PHIs) and the evaluation would take long time be tedious and tiresome for the evaluator, and risk that the will not be concentrated on their work.

4 Results

The pseudonymisation program was executed on the whole Stockholm EPR PHI Corpus that contained the tags described in Table 2. In the figure, the number of replacements by the pseudonymisation program is also presented.

In Figure 1, a original but pseudonymised record can be seen, where first and last personal names have been replaced as well as healthcare units and phone numbers.

5 Evaluation

The evaluation of the system was carried out by two evaluators, both computer scientists one senior and one junior with knowledge in clinical data mining. The senior computer scientist is a second language Swedish speaker, and the junior computer scientist is a native speaker of Swedish. None of the evaluators had seen the electronic patient records beforehand. Both evaluators had also signed confidentiality agreements, the same as the author of this article had signed.

The two evaluators, the senior and the junior, could only correctly judge that 4 and 6 records respectively were pseudonymised of the total 98 records where 40 were pseudo records.

[2]Nick Sweeting 2014 implementation of Peter Norvig's spell checker, https://github.com/pirate/spellchecker/

Epikris Huddinge
Ansv. specialist-/ överläkare Caroline Berg
Journalförare Marianne Lindgren
Utskriftsdatum 20120325
Vårdtid 20120311-20120318
Huvuddiagnos enl. ICD-10
Anamnes 52-årig kvinna, välkänd på kliniken. Går hos Karin Lundgren samt på smärtmottagningen.
Har en kronisk huvudvärk utan säker genes. Insatt på Metadon, Actiqe och Stesolid.
Sökte den 22/5 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 Torbjörn Andreasson.
Pat kommer till NIVA och kräver att få läggas in på IVA och hotar att sluta med samtliga mediciner.
Pat har haft flera samtal med PAL på Löwet, Sandra Månsson. Hänvisar till tidigare anteckningar.

In Eng:
Discharge letter Huddinge
Responsible. specialist / chief physician Caroline Berg
Medical secretary Marianne Lindgren
Print Date 20120325
Care episode 20120311-20120318
Main diagnosis according to ICD-10
History of 52-year-old woman, well known in the clinic. Treated by Karin Lundgren and at the pain clinic.
Has a chronic headache without a known origin. Given Methadone, Actiqe and Stesolid.
Came to clinic on the 22/5 due to unsustainable situation with inadequate pain control.
Pat. is frustated over the long waiting time for the discontinuation of opiates which was to be done via IVA and planned by Dr. Torbjörn Andreasson.
Pat comes to NIVA and demands to be admitted to IVA and threatens to stop taking all drugs.
Pat had several conversations with PAL at Löwet, Sandra Månsson. Refers to previous notes.

Figure 1: An example of a pseudonymised electronic patient record written in Swedish (and its translation to English). It was judged as real by both evaluators. The text is relatively coherent, Mentioning places such as *Huddinge*, dates and date periods *20120311-20120318*, several personal names healthcare units such as *IVA*, Intensiv VårdsAvdelning (in Eng: Intensive Care Unit), *NIVA*, Neurologisk Intensiv VårdsAvdelning (in Eng: Neurological Intensive Care Unit) and *Löwet* a colloquial for Löwenströmska sjukhuset, (In Eng: Löwenströmska hospital) all of them pseudonymised. The underlinings have been added in this figure to show the PHI.

Entity	PHI-instance	Pseudo-nymised
First Name	923	-
Female First Name	-	364
Male First Name	-	555
Gender-neutral First Name	-	2
Last Name	929	929
Age	56	56
Phone Number	125	137
Location	148	148
Full Date	551	551
Date Part	711	711
Health Care Unit	1,025	1,026
Sum	**4,468**	**4,479**

Table 2: Types and numbers of all annotated PHI tokens in the Stockholm EPR PHI Corpus, replaced by the pseudonymisation program, the spell checker was used for nine males first names and ten females first names. (No social security number was found). The numbers are not matching completely. Location corresponds to all locations in Table 1.

Ny kontakt med smärtmottagning måndag.
Närstående Mamma Madeleine tfn: 0652 7256 , Bror Madeleine tfn 078 1295067
Upplysning Får lämnas

In Eng:
New contact with pain clinic Monday.
Related Mother Madeleine ph: 0652 7256 ,
Brother Madeleine ph 078 1295067
Inquiry May be given

Figure 2: An example of a pseudonymised electronic patient record written in Swedish that was correctly judged as pseudonymised by the senior evaluator by the agreement brother *Bror* and gender of the name of the brother *Madeleine* which is a female name. Also the mother *Mamma* has the same name *Madeleine* as the brother, which makes it confusing.

The two evaluators also judged incorrectly that two and one records respectively were pseudonymised, while they were non-pseudonymised records. In average only 3.5 percent of the psuedonymised records were correctly judged as psuedonymised and 1.5 percent of the non-pseudonymised records were wrongly judged as psuedonymised.

Concluding that 91 percent of the pseudonymised records were judged as original (or non-pseudonymised) records in average.

The senior evaluator obtained a precision of 67 percent and a recall of 10 percent. The junior evaluator obtained a precision of 75 percent and a recall of 8 percent. None of the pseudonymised records was judged as pseudonymised by both of the evaluators.

Some of the comments for reasons for revealing the records as pseudonymised were that the physician's name was strange, a brother with a female name, and also wrong gender, see Figure 2.

6 Discussion and conclusion

Many of the identified pseudonymised patient records were revealed because of strange combinations of first names and last names, for example, a Christian male first name and a Muslim last name such as *Peter Mohamed*, both separately common Swedish names. Another case was family relationships such as husband, wife, daughter or son combined with the wrong gender of the name. Another exemple was two dates that were too inconsistent in time, where the cause happens after the effect, also wrong type of healthcare unit mentioned for the specific disease the patient has. For example "bor permanent på Löwet", (In Eng: "Lives permanent at Löwet"), where *Löwet* is colloquial for "Löwenströmska sjukhuset", (In Eng: "Löwenströmska hospital"), and a patient does not usually live at a hospital but on a Geriatric unit or Residential home.

Also, the wrong preposition used for a location reveals that the records are pseudonymised, for example, prepositions in front of places and countries could pose problems since countries written in Swedish need the Swedish preposition *i* (English: in) while streets and islands require *på* (English: on).

The PHI tag *Location* is very general and covers everything from different locations (street, place, city, muncipality, country...), and in some cases organisations, companies or product names. The annotations for locations and healthcare units are basic and when pseudonymising them, strange combinations can occur. The generated phone numbers could in some case generate strange grouping of numbers.

The research question was whether a reader of an electronic patient record could reveal if one record was pseudonymised or not. The answer is no, hence a layperson reader probably cannot reveal to 91 percent that a record is pseudonymised or not. Probably a clinically trained person might reveal more, and adding external databases will ofcourse assist in revealing if the record is pseudonymised or not.

The results in this study are much more promising than the study by Björkegren (2016) where half of the randomly generated records that was pseudonymised were revealed as pseudonymised by the evaluators, in their case, three different evaluators evaluating the same 17 records were eight records were pseudonymised. One of the evaluator was a physician at Karolinska University Hospital, the two others were computer scientists.

The results are also in line with the results in (Meystre et al., 2014) where five physicians were asked if they could recognise their patient in 86 de-identified patient records and where none succeeded in this task.

The purpose of Hiding In Plain Sight (HIPS) to not reveal the un-annotated PHI can also be considered fulfilled.

Future research will be to improve the pseudonymisation program to diminish the number of bugs, but also to create common combinations of first and last names, checking the agreement for family relationship and the right gender of first names, for example, brother and the use of a male name and not a female name, etc. Solve the problem with the broad class *Location* by dividing it into the classes "real' locations and organisations.

The use of more general healthcare unit names and also improve the date shifting mechanisms, but also to make more significant variations of the produced vocabulary to make the pseudonymised data useful as training data for machine learning algorithms.

There is also research going on for generating synthetical patient records from real patient records, that does not contain any sensitive information. This method could be an entirely different way to go.

Acknowledgments

Great thanks to Panos Papapetrou and Isak Samsten for carrying out the evaluation of the electronic patient records texts. Thanks also to Andreas Amsenius that provided the special list for town squares and parks from earlier work on de-identifications of emails. and thanks also to André Antfolk and Rikard Branting that provided with list on locations that were organised hierarchically.

References

Alyaa Alfalahi, Sara Brissman, and Hercules Dalianis. 2012. Pseudonymisation of personal names and other PHIs in an annotated clinical Swedish corpus. In *Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012) held in conjunction with LREC 2012, May 26, Istanbul*, pages 49–54.

Andreé Antfolk and Rikard Branting. 2016. Pseudonymisering av platser i patient-journaltexter (In Swedish). Bachelor's thesis, Department of Computer and Systems Sciences, Stockholm University.

Andreas Björkegren. 2016. Pseudonymisering av digitala patientjournaler (In Swedish). Bachelor's thesis, Department of Computer and Systems Sciences, Stockholm University.

David Carrell, Bradley Malin, John Aberdeen, Samuel Bayer, Cheryl Clark, Ben Wellner, and Lynette Hirschman. 2013. Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. *Journal of the American Medical Informatics Association*, 20(2):342–348.

Hercules Dalianis. 2018. *Clinical Text Mining: Secondary Use of Electronic Patient Records*. Springer.

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*, 1:6.

Louise Deleger, Todd Lingren, Yizhao Ni, Megan Kaiser, Laura Stoutenborough, Keith Marsolo, Michal Kouril, Katalin Molnar, and Imre Solti. 2014. Preparing an annotated gold standard corpus to share with extramural investigators for de-identification research. *Journal of Biomedical Informatics*, 50:173–183.

Margaret Douglass, Gari D. Clifford, Andrew Reisner, George B. Moody, and Roger G. Mark. 2004. Computer-assisted de-identification of free text in the MIMIC II database. In *Computers in Cardiology, 2004*, pages 341–344. IEEE.

Cyril Grouin, Nicolas Griffon, and Aurélie Névéol. 2015. Is it possible to recover personal health information from an automatically de-identified corpus of french ehrs? In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, pages 31–39.

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

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.

Latanya Sweeney. 1996. Replacing personally-identifying information in medical records, the scrub system. In *Proceedings of the AMIA annual fall symposium*, page 333. American Medical Informatics Association.

Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Evaluating the State-of-the-Art in Automatic De-identification. *Journal of the American Medical Informatics Association*, 14(5):550–563.

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.