Chapter 9
Ethics and Privacy of Patient Records for Clinical Text Mining Research

There is an abundance of electronic patient records produced today within healthcare. These records contain valuable information on symptoms and disorders, reasoning to determine on the diagnosis and the treatment of the patient, but also on adverse events the patient might have experienced. The whole process of obtaining access to electronic patient records for research is complicated and requires certain steps. This chapter will describe the process of applying for of an ethical permission, data extraction and safe storage of sensitive data. Sensitive data in that sense it is personal data that can identify individuals. Clinical free text extraction require often to identify the sensitive data, so-called de-identification, followed by replacing the identified sensitive data with fake data, so-called pseudonymisation. Privacy preserving methods will also be discussed when connecting the data to other sources and databases there is a risk of re-identification and to avoid this privacy preserving data linkage has to be carried out.

9.1 Ethical Permission

In Sweden and in many other countries ethical permission is needed to perform research on health related data that involves humans. Therefore, first of all an *ethical permission* is needed from an *ethical review board* before any research involving patient records can be carried out. The ethical review board is used by medical researchers applying for permission to perform research involving humans or animals. In the case of clinical text mining, no experiments on humans or animals are carried out but a large amount of data is studied. Data in the form of patient records that contain information about people who have not been asked if they want to participate in any experiment. Usually each individual is asked to agree to participate in a research project, however, in our case we are studying information that is de-identified and no individual can be singled out. The data is also on an

H. Dalianis, *Clinical Text Mining*, https://doi.org/10.1007/978-3-319-78503-5_9

aggregated level and not at an individual level, hence we do not study specific people.

The planned research has to be described in an *application for ethical review* (in Swedish: Etikprövningsansökan) by the researcher, which is addressed to the ethical review board. In the ethical review application, the experiments, the possible outcome and also how the data will be stored are described. The research shall of course not harm any individual and aims to be a benefit for healthcare and humanity. The review board then decides if the research is approved and can be carried out.

As previously mentioned the law in Sweden says that consent must be given by the patient before his or her data can be used for research; however, the patient records used in this research are de-identified meaning that the patient's identity is unknown. According to the Swedish Personal Data Act, (in Swedish) Personuppgiftslagen (1998:204), also abbreviated to PUL, personal information is all type of information that can be linked to a physical person, if it cannot be linked then it is not personal information.

However, in the free text of the patient records there is sometimes mentioned information that may identify the patient, as for example telephone numbers of relatives, (for example the *patient's wife Mary's phone number is 081-590 29 38*), in these cases the patient can be considered to be identifiable and hence the patient has to either be asked for consent to be part of the research project, or the identifiable parts removed.

A overview of the process of obtaining clinical data Internationally, in Australia and in Finland is described in Suominen (2012) in Australia, Austria, Finnish, Swiss and the USA is described in Suominen et al. (2017).

9.2 Social Security Number

After the review board has approved the application for ethical review the research can be carried out. The hospital is contacted (again) with the approval from the ethical review board and then the hospital management decides if the hospital can hand out data for research. This decision can also be taken at a clinical unit level for each clinical unit, but it is better to have the hospital management as the decision maker, since the research project can get access to more data from *all* clinical units. In some cases extra sensitive data for patients treated for psychiatric reasons or sexually transmitted diseases has to be left out of the study.

Before obtaining the data from the hospital the personal names and the social security numbers must be removed. The social security number is replaced by a serial number. At the hospital a key is generated and stored to link the serial number to the social security number so the patient can be tracked later, or the data linked to another register.

Observe also the Scandinavian countries, have a unique personal identity number (in Swedish: *personnummer*) used for all contact with the authorities. (each Scandinavian country has its own system of personal identity numbers).

In Sweden a personal identity number is given to each individual at birth and is kept for the whole life span. Immigrants are given a number on arrival to the country, a number which is also kept for their whole life. The advantage of this number is that a person or a patient can be followed when admitted to the hospital but also when discharged from or re-admitted to the hospital. The same number is also used in other registries, such as the prescribed drugs registry, the cause of death registry, or a cancer registry and business registries. However, to connect different registries permission is needed from different authorities. When linking registries, a serial number is used instead of the personal identity number to make the transaction anonymous.

After obtaining ethical permission for the research, the researcher needs to get access to the anonymised patient records to perform the research. The access to the patient records may be cumbersome technically if there is no easy way to extract the data from the electronic patient record system. However, if the data is extracted it is necessary to store it at a safe place at the research unit or at the hospital. A safe place that cannot be accessed by any unauthorised personnel.

9.3 Safe Storage

If the data is stored at the academic institution and not at the hospital certain steps have to be carried out. In the case of Stockholm University with HEALTH BANK the following applies (Dalianis et al. 2015): The de-identified patient records for research are stored on a server that is encrypted and password protected, without any Internet connection. The server is locked to the server rack. The server rack is in a server room in that in turn is locked and alarmed, and has no Internet connection.

The server room can only be accessed by people who are authorised to work with the data or the people that take care of the servers, backups etc. Backups are stored in a safe place such as on an encrypted hard disc, that is in turn stored in a safe place.

The researchers work with the data have all signed the confidentiality agreement, which is the same type that the health personnel working at hospitals have signed.

Possible encryption systems to use are TrueCrypt or Veracrypt[1] both can be executed in Windows, Linux and Mac OS X; however, TrueCrypt is not maintained any more.

9.4 Automatic De-Identification of Patient Records

At Stockholm University for the infrastructure HEALTH BANK—Swedish Health Record Research Bank, the patient records are anonymised in the sense that the tables in the database containing personal names and personal identity numbers

[1] https://veracrypt.codeplex.com. Accessed 2018-01-11.

have been replaced with serial numbers so each individual can be tracked (Dalianis et al. 2015). However this is not really enough for de-identification purposes since plenty of sensitive information resides in the free text in the form of personal names, addresses, telephone numbers etc.

This sensitive information in patient record text is also called Protected Health Information (PHI). In the USA, for example, and many other countries there are requirements that all information in a patient record should be de-identified before the research can be carried out.

The secondary use of these records demands privacy preserving measures. Specifically, it is valuable to know the number, type and nature of fields the sensitive information where present in the electronic patient record.

The American Health Insurance Portability and Accountability Act (HIPAA) (2003), stated which protected health information (PHI) should be removed or obscured from an electronic patient record before the patient record can be used for research. It specifies 18 different PHI classes, such as *personal names, addresses, phone numbers, email addresses, dates, ages over 89, social security numbers*, etc. Note that HIPAA does not require changes to ages under 90 years, institutions or initials.

There have been many attempts to construct de-identification tools, see Uzuner et al. (2007) and Meystre et al. (2010) for nice review articles of different systems. The de-identification systems use both rule-based approaches and machine learning approaches to perform the de-identification, in some cases they use a hybrid approach.

The performance of the systems presented in Meystre et al. (2010) and Uzuner et al. (2007), had F-scores ranging from 0.80 up to, in some cases, 0.97. High recall is preferred over high precision since it is important to identify all sensitive data. Dates and phone numbers obtain the highest scores in de-identification (using rule-based systems or regular expressions), while personal names and in some cases locations obtain lower scores.

For Swedish, Kokkinakis and Thurin (2007) constructed a rule-based system and obtained 96.7% precision, 89.35% recall and obtained an F-score of 0.93. A machine learning-based approach by Dalianis and Velupillai (2010b) utilising Stanford NER CRF and the manually annotated Stockholm EPR PHI Corpus as training and evaluation data yielded an F-score of 0.80. Henriksson et al. (2017b) improved the results on the same data. The authors used extracted features jointly with the CRF to build a predictive model applied on a larger clinical corpus. IOB-encoding of class labels was used, indicates whether a token is inside (I), outside (O) or at the beginning (B), a given named entity in the text. For the feature optimisation L1 and L2 regularisation were used. The best results obtained were 92.65% for precision, 81.29% for recall and an F-score of 0.87.

In production systems both an automatic approach for de-identification and a manual inspection is carried out before the corpus is released for research.

In a recent study by Meystre et al. (2017) the authors have made an extensive review of the status of clinical data reuse and secondary use.

9.4.1 Density of PHI in Electronic Patient Record Text

Regarding the density of PHI in the free text of patient records we can observe the studies such as Douglass et al. (2004). They used MIMIC II, which is a well-known database of patient records and found 1776 instances of PHI in 339,150 tokens giving 0.5% sensitive information in the clinical text.

Dorr et al. (2006) found 1074 instances of PHI in 70,552 tokens, giving 2.9% sensitive information including 1% personal names. The also looked at the different professions and their contribution to the density of PHI, see Table 9.1. One interesting observation is that one human annotator averaged 13,100 tokens and 326 PHI elements per hour.

One well cited work is by Neamatullah et al. (2008) using part of the MIMIC II database. The clinical text consists of 2434 nursing notes containing 334,000 words and 1779 instances of PHI, giving 0.5% PHI. Their manual annotation rate was 18,000 words and 90 PHI terms per hour. The source code and data is available for use.

In Uzuner et al. (2008) is described 889 de-identified discharge summaries in the so-called challenge corpus containing in total 472,315 tokens and 28,188 PHI or 6% of the total data.

The studies mentioned were all for English clinical text, for other languages such as Swedish, Danish and French we have the following studies:

For Swedish clinical text Kokkinakis and Thurin (2007) extracted 14,000 tokens from 200 hospital discharge letters containing 1450 instances of PHI, or 10% of the total amount of text. Velupillai et al. (2009) annotated 174,000 tokens in a Swedish clinical text and found one third was personal names. In total 4423 instances of PHI were found equating 2.5% PHI of the total amount of tokens.

This corpus is also called the Stockholm EPR PHI Corpus and contains 100 patient records from five different clinical units: *neurology, orthopaedia, infection, dental surgery* and *nutrition* (Velupillai et al. 2009), see Table 9.2 for the distribution of PHI entities.

A large amount of the information in an electronic patient record system is unstructured in form of free text. The Stockholm EPR PHI corpus contained

Table 9.1 Authorship of PHI distributed over professions, part of Table 3 in Dorr et al. (2006) for some reason the results add up to over 100%. The authors do not report if these results are normalised according to text size or not

Profession	Percentage PHI
Physician or extender	67.6
Nurse	20.6
Pharmacist	8.0
Social worker	3.4
Other	3.8
Sum	**103.4**

Table 9.2 Types and numbers of annotated tokens in the Stockholm EPR PHI Corpus

Entity	Number of PHI-instances
First name	923
Last name	929
Age	56
Health care unit	1021
Location	148
Full date	500
Date part	710
Phone number	136
Sum	**4423**

380,000 tokens, of these 174,000 tokens were free text giving 46% unstructured information and 54% structured information. The structured information did not include X-rays, images etc. The term, token and word are here used interchangeable. Tokens may also include interpunctuations (Velupillai et al. 2009). The amount of text in the patient records varies depending on the domain studied, for example, psychiatric patient records tend to contain much more text than emergency or general practitioner's records.

For Danish clinical text a study by Pantazos et al. (2016)[2] involved the annotation of 369 full patient records which contained 73,150 words and 1320 instances of PHI, in total 1.8% PHI.

For French clinical text one study by Grouin and Névéol (2014) was performed, annotating 29,437 tokens containing 3964 PHI or 13.4% PHI of which 41% were personal names.

In Hanauer et al. (2013) there is an illustrative heat map visualising and comparing the prevalence of PHI in different document types. In total 116,000 clinical documents written in American English were studied. The study included bootstrapping and pre-annotation to make annotation faster. The highest amount of PHI are in *discharge summaries, outpatient consult, admission history and physical* and *social work notes* overall; dates, clinician and patient personal names were found to have the highest density. In one subset with 20,500 tokens, 752 PHI instances were annotated (in 1 h) giving 3.7% PHI.

Carrell et al. (2013) studied oncology reports found 343 PHI in 22,525 words giving 1.5% PHI.

In another study described in Henriksson et al. (2017b) Swedish patient records were studied, and the highest density of PHI was found in discharge notes written by physicians, and specifically in oncology. The type of PHI that was most prevalent were dates, and first and last names (surnames) of the patient. On average 1.57% PHI was found in the corpus, see Fig. 9.1. In Henriksson et al. (2017a) the trained model were used on other sub corpora, but the performance dropped, probably due

[2]This book is written in memory of Kostas Pantazos who passed away at a young age in October 2015.

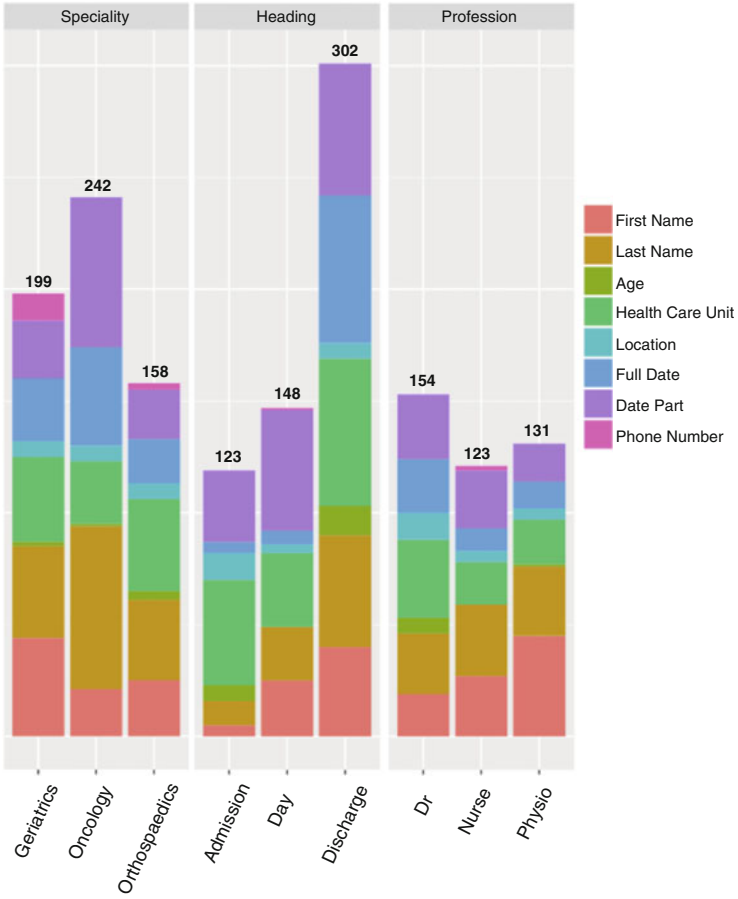

Fig. 9.1 Prevalance and distribution of PHI in various types of clinical notes in HEALTH BANK. Taken from Figure 1 in Henriksson et al. (2017b) (© 2017 The authors—reprinted with permission from the authors and AMIA. Published in Henriksson et al. 2017b)

to the differences in text style. Domain adaption techniques are therefore necessary when training a de-identification system in one domain and processing clinical text from another domain.

The Stockholm EPR PHI Corpus was used for training and the newly annotated *Stockholm EPR PHI Domains Corpus* was used for evaluation. The Stockholm EPR PHI Domains Corpus encompasses the clinical domains *geriatrics, oncology* and *orthopaedics* (including surgery) and contains 1579 annotated instances of PHI: in total there are 63,417 tokens in the corpora. Regarding the Stockholm EPR PHI corpus see Table 9.2.

9.4.2 Pseudonymisation of Electronic Patient Records

The de-identification process has two parts: finding the PHI and then removing or replacing them with an identifier in the form of the class, for example, *patient name, telephone number* or *location*; however, this makes the clinical text less natural to read. Another issue is that PHI may contain important epidemiological information, such as locations where disorders may occur; therefore, is be valuable to maintain some general information about the location.

It is better to replace the class names or the original identified PHI with *surrogates* or *pseudonyms*, which look natural, this process is called *resynthesis* or "re-identification with fake data".

The first attempts to do so were in the de-identification work by Sweeney (1996), where two modules were created, one to de-identify and the other one to replace PHI with surrogates, or what Sweeney called *pseudo-values* in her work.

Each detection algorithm is associated with a replacement algorithm. The strategy is that a date is replaced with a similar date nearby, a personal name is replaced by a fictitious name. These names are produced by orthographical rules creating fictitious names that sounds reasonable but do not belong to a known person. A unique name is always replaced with the same fictitious unique name, a so-called consistent replacement. Sweeney (1996) does not mention how she deals with locations, phone numbers etc.

Douglass et al. (2004) describe how after identification PHI was replaced with surrogates according to the following algorithm: dates were shifted by the same random number of weeks or years, but the days of the weeks were preserved; personal names were replaced with names from the publicly available list of names from the Boston area in the US by randomly substituting first and last names. Locations were replaced by randomly selected small towns; hospitals and hospitals wards were given fictitious names. Moreover, Douglass et al. (2004) explain the process and show the interface of manually validating and correcting the results of the automatic identification and replacement of PHI with surrogates, with the aim of making the corpus available to other researchers.

For Swedish there is a pseudonymisation study by Alfalahi et al. (2012), where personal names are replaced with other personal names in a consistent way, female first names are replaced with common female first names and corresponding is carried out for male first names. Misspelled first names or gender neutral first names are replaced with other gender neutral names, such as *Kim, Pat, Robin* or *Andrea*. Addresses and phone numbers are also replaced, dates are shifted and ages changed slightly. However, locations and healthcare units are replaced by the default location and healthcare unit *Stockholm* and *Solvillan* respectively.

Another study was carried out on the same Swedish data by Antfolk and Branting (2016), where replacement of locations was the focus. The authors replaced locations, such as *places, cities*, and *countries*, with locations that were situated close by to keep possible epidemiological relations, but as the authors also stated there were problems with misspelled or abbreviated locations. Also when a

complete address was tagged with *street, street number* and *city*, they did not change it, since this type of pattern was not in the scope of the study. Prepositions before countries could be a problem in some cases. In Swedish the preposition *i* (English: in) and *på* (English: on) could make the replacement of a country name peculiar in some cases. You live on islands, *på Island* (on Iceland) but you live in countries *i Norge* (in Norway).

In a similar study by Björkegren (2011) location was also automatically replaced by surrogates according to the classes: *countries, cities, streets* and *companies/organisations*, since there is only one annotated class called *location* in the de-identified data, each sub class had to be identified using name lists. An evaluation was carried out where three respondents had to decide which of 17 patient records were pseudonymised or an original record. Half of the records were identified as pseudonymised. This concludes the pseudonymisation program was not good enough and the data was still too complicated for natural pseudonymisation. Natural pseudonymisation needs to have consistent geographical information, such as for example that a street is situated in the correct part of the city. This concludes also that more research work has to be carried out, encompassing better annotation work.

In another approach for pseudonymisation for English by Deleger et al. (2014), the authors used the American English clinical corpus from Physionet, i2b2 and the Cincinnati Children's Hospital Medical Center (CCHMC) corpus for *cross-training*[3] and evaluation of their de-identification algorithm. Part of the research work consisted of creating a replacement algorithm for the identified PHI. Dates were replaced with dates in the same format. Telephone area codes were replaced with other existing area codes. The rest of the phone number was replaced with the same number of random numbers. E-mail addresses were replaced with a random set of characters corresponding to the original number of letters. No replaced PHI had any resemblance with any other PHI occurring in the whole dataset. Personal names were replaced with real names originating from the US Census Bureau, with a frequency above 144 (or 0.002% of the data). Gender of the first name was of course considered. Locations such as streets and places were replaced with new random locations from the corpus, and street numbers were replaced with a random number of same length. No combination of street and number reoccurred as in the original corpus.

In a study by Carrell et al. (2013) the authors describe how to hide non-identified PHI, so-called *residual identifiers*, with the method *Hiding In Plain Sight (HIPS)*, which actually involves to removal of annotation on all de-identified PHI, after replacing them with surrogates. This method hides from plain sight the PHI that could not be de-identified, or residual identifiers. Since neither the de-identification system or the human annotators identify the PHI, the secondary users of the de-identified text believe that identified PHI already has replaced by the pseudonymisation system.

[3]Training on one corpus and evaluating on the other corpus.

State of the art de-identification systems reach recall rates of 95–97%, by using the technique described in Carrell et al. (2013) 90% of the residual identifiers can be effectively concealed the HIPS without improving the de-identification system. The main point is that the risk can be calculated on re-identifying residual identifiers depending on what class of PHI they belong to. For example, ages, dates and organisation names are more difficult to re-identify than personal names.

Meystre (2015) has written a nice overview of the whole process of de-identification and resynthesis.

9.4.3 Re-Identification and Privacy

There is no 100% guarantee that individuals in a de-identified and resynthesised patient record cannot be identified. However, some experiments have been carried out to test this. Meystre et al. (2014) did an experiment that after de-identification and resynthesis (replacing the PHI with realistic surrogates) of 86 discharge summaries let the treating physician try to identify their patient. The physician had written the discharge summary between 1 and 3 months before the study. Of the five physicians in the study none could identify their patient.

Sweeney (2002) writes about something that is today called *privacy preserving data linkage*, whereby adding more and more data sources to a de-identified database may cause re-identification of individuals. An example of this presented in the study was the governor of the state of Massachusetts, who like many other US state employees had filed his medical record for health insurance. The data was considered to be anonymous and was given out for research, but using another database with voter registration for Cambridge, Massachusetts, the governor could be identified using ZIP code, birth date and gender in both databases.

Sweeney (2002) presents her algorithm called *k-anonymity*. In short, it can calculate the risk of data can being de-identified by counting the number of attributes in a database to conclude if it is safe to link the data or not to the database. As she writes in the article: *A release provides k-anonymity protection if the information for each person contained in the release cannot be distinguished from at least k−1 individuals whose information also appears in the release*.

Gkoulalas-Divanis et al. (2014) compare 45 different algorithms for calculating privacy preserving linkage. The 45 algorithms are divided into two privacy models. One model is used for demographics and the other model is used for diagnosis codes. Furthermore, Gkoulalas-Divanis et al. (2014) discuss two models for data access. One is the protected data repository model where researchers can interact with the aggregated data with some restrictions. The other model is data publishing where the user has complete access to the data.

El Emam et al. (2015) there is a nice description of the three sensitivity levels of data: non-public, quasi public, public data, and the risk for identification; however, they do not mention clinical text explicitly.

Andersen et al. (2014) have a description on how to use distributed sources of patient records for statistical analysis and still preserve privacy using a large set of possible secure multi-party computation algorithms and computing graphs.

Black Box Approach

One approach was considered in the master's thesis of Almgren and Pavlov (2016), also published in Almgren et al. (2016). We can call it the *black box approach*, since the electronic patient records were not directly accessed by the authors. Almgren and Pavlov had knowledge of the format of the sensitive manually annotated data in the electronic patient records to be evaluated, in this case clinical entities in Swedish (disorder & finding, pharmaceutical drug and body structure).

The authors trained two models, word-vectors and a recurrent neural network model, on out of domain (non-clinical) training data using Swedish scientific medical text (Läkartidningen). The program code and the trained models were sent to an authorised person with access to the sensitive data in the black box. The person executed the program code and delivered the numerical results in the form of precision, recall and F-score to the authors, hence the researchers could use the clinical data without ethical permission.

Another interesting approach similar to the black box approach is to remove the free text of the patient records and only keep the features of the tokens, for example, the length of a token, whether the token contains numerical characters, or upper or lowercase characters, POS tags, features of the tokens preceding and following the token to be analysed, and if the token is PHI and what type of PHI. Using this information as features in the Random Forest algorithm gave almost as good results in identifying PHI as using CRF with the same features and the real text (Dalianis and Boström 2012). Hence, the approach by Dalianis and Boström (2012), can be used to release a large amount of sensitive data without a risk of revealing the identity of any patient. Such data can be used to train and evaluate the machine learning system as well as to construct commercial systems.

9.5 Summary of Ethics and Privacy of Patient Records for Clinical Text Mining Research

This chapter has described the process of getting access to electronic patient records for clinical text mining, which includes writing an ethical permission application, removing sensitive information (PHI) such as personal names, phone numbers, addresses etc. in the patient records, getting access to the data and keeping it in a safe storage.

The best automatic de-identification systems obtain an F-scores up to 0.97, but also require manual review before the data can be released for research. Generally, the highest density of sensitive PHI is in the assessment and social fields, and in the discharge summaries of the patient records. On average 2% of the information found in clinical text is sensitive.

Various methods and results for pseudonymisation, meaning the replacement of real PHI with surrogates in electronic records were described.

Another approach, the black box approach, was described where the sensitive text is stored in a protected black box that can be used by external users without direct access to the data.

Finally, was privacy preserving data linkage explained. There is a risk when connecting de-identified data with external databases since it may reveal sensitive hidden information, to avoid this risk privacy preserving data linkage can be used.