Telematics and Informatics 35 (2018) 727–736

Contents lists available at ScienceDirect

Telematics and Informatics

journal homepage: www.elsevier.com/locate/tele

DEDUCE: A pattern matching method for automatic de-identification of Dutch medical text

Vincent Menger[a,*], Floor Scheepers[b], Lisette Maria van Wijk[a], Marco Spruit[a]

[a] Department of Information and Computing Sciences, Utrecht University, P.O. Box 80089, 3508 TB Utrecht, The Netherlands
[b] Department of Psychiatry, University Medical Center Utrecht, P.O. Box 85500, 3508 GA Utrecht, The Netherlands

ARTICLE INFO

Keywords:
De-identification
Dutch medical text
Pattern matching
Protected Health Information
Patient privacy

ABSTRACT

In order to use medical text for research purposes, it is necessary to de-identify the text for legal and privacy reasons. We report on a pattern matching method to automatically de-identify medical text written in Dutch, which requires a low amount of effort to be hand tailored. First, a selection of Protected Health Information (PHI) categories is determined in cooperation with medical staff. Then, we devise a method for de-identifying all information in one of these PHI categories, that relies on lookup tables, decision rules and fuzzy string matching. Our de-identification method DEDUCE is validated on a test corpus of 200 nursing notes and 200 treatment plans obtained from the University Medical Center Utrecht (UMCU) in the Netherlands, achieving a total micro-averaged precision of 0.814, a recall of 0.916 and a F_1-score of 0.862. For person names, a recall of 0.964 was achieved, while no names of patients were missed.

1. Introduction

Data from Electronic Health Records (EHR) is since long being used for medical research purposes, and with the increasing digitalization in the medical world even more so (Milovic, 2012; Murdoch and Detsky, 2013). Since many hospitals today have adopted an EHR system, the produced data can be used for clinical research with few further limitations (Koh and Tan, 2005; Lee et al., 2013). Both health care institutions and patients can directly benefit from the results of this kind of research that can improve diagnosis, treatment, hospital operations and more (Jensen et al., 2012; Menger et al., 2016). The use of patient data for research however puts a strain on patient privacy, since this requires getting the data out of their health care context (Patel et al., 2014). This entails for example copying the data to different databases, where it can be accessed by data analysts (Simon et al., 2000). Technical staff such as data managers or data analysts typically do not have a treatment relation with the patient, and therefore should not be able to identify individual patients in a research dataset. Medical staff on the other hand is allowed to see patient information under medical confidentiality, but lack the technical skills to perform advanced analysis that is needed for obtaining direct clinical value from data.

Multiple ways exist to solve this problem for structured data. One rigorous approach is to remove all variables that could identify a person, such as patient names, addresses and social security numbers from the dataset (El Emam et al., 2006). A more sophisticated solution is a k-anonymity method, where the information of a patient is indistinguishable from at least $k-1$ individuals (Toledo and Spruit, 2016). In recent years however, the more widespread availability and quality of text mining approaches have shifted attention to analyzing unstructured textual data in addition to structured data. It is becoming ever more apparent that these approaches offer

* Corresponding author.
E-mail addresses: v.j.menger@uu.nl (V. Menger), f.e.scheepers-2@umcutrecht.nl (F. Scheepers), l.m.vanwijk2@students.uu.nl (L.M. van Wijk), m.r.spruit@uu.nl (M. Spruit).

http://dx.doi.org/10.1016/j.tele.2017.08.002
Received 31 December 2016; Received in revised form 8 May 2017; Accepted 1 August 2017
Available online 08 August 2017

substantial benefits for data driven research (e.g. Harpaz et al., 2014; Patel et al., 2014; Maenner et al., 2013), and that textual data from the medical domain holds valuable information that should not be disregarded. Therefore, if we require research datasets to be anonymous to mitigate the potential negative impact on patient privacy, we must also de-identify the medical free text variables.

From a patient perspective, protecting the private details of a disease from the public is essential in retaining the trust bond between a physician and the patient (Krishna et al., 2007). Any violation of this confidentiality can therefore have serious consequences for the relation between a healthcare institution and a patient. A patient may be adverse to their data being used for research when non medical staff has access to private details, and might even consider seeking treatment elsewhere. Moreover, a potential data breach may expose private patient information to the general public. In the USA alone, between 2010 and 2013, 29,000,000 patient records were compromised (Liu et al., 2015). Clearly, such events have serious consequences for both the hospital and the patient.

From a legal perspective, on a European level the Directive 95/46/EG of the European Parliament on the protection of individuals with regard to the processing of personal data and on the free movement of such data was introduced in 1995 (European Data Protection Directive, 1995). All members of the European Union are obliged to take this directive into account, but how they implement it varies for each member. For example, in Sweden regional Ethics Committees give permission for the reuse of electronic patient records if the information that can identify a patient is removed (Velupillai et al., 2009). Similar measures are implemented in France in the French Data Protection Authority (Grouin et al., 2009). In the USA, the stricter Health Insurance Portability and Accountability Act (HIPAA) protects the privacy of healthcare data, requiring 18 different categories of identifying information ranging from person names to biometric identifiers such as fingerprints to be removed from medical data (HIPAA, 1996).

In the Netherlands no specific laws on the reuse of medical data exist, but there are general rules for dealing with personal data, that can be applied to medical data as well. Since EHR data is used for retrospective research, is not specifically collected for research purposes, and human subjects are only indirectly involved, the Medical Research Involving Human Subjects Act (WMO) does not need to be taken into account. Only the Agreement on Medical Treatment Act (WBGO) and the Personal Data Protection Act (WbP), which is the implementation of the European Directive 95/46/EG mentioned above, play a role in this situation. Retrospective research with medical records needs to be proposed to the medical ethics committee (METC), which verifies that the proposed research is in line with privacy legislation. An exception to this is when only anonymized data is used, which is the case if the de-identification process is executed perfectly.

For the two reasons above it is therefore important to de-identify as much of research data as possible, both to retain patient privacy and to be able to comply with legal requirements. For de-identification of personal data, a distinction can be made between directly identifying information and indirectly identifying information. Directly identifying information, such as names, phone numbers and citizen service numbers allow identification of a person with just that information. Indirectly identifying information, such as postal codes and birth dates, is not directly relatable to a person, but if pieces of indirectly identifying information are combined it is easy to identify someone (Borking and Raab, 2001). In medical text data both directly and indirectly identifying information can be present, and both types of information need to be removed to successfully de-identify medical text. Although manual de-identification is possible, it is time consuming and generally prone to error, while automatic de-identification is feasible and easily scalable to large numbers of records (Deleger et al., 2013). For this reason, we choose to develop an automatic de-identification method. Our goal is to remove as much de-identifying information as possible, while ensuring the de-identified text is still human readable, so that research can still be carried out. Even strict de-identification methods still retain good readability of the remaining text (Meystre et al., 2014). We therefore strive to balance towards developing a method with a high recall while also maintaining a good precision.

Since there are differences in legislation that exists on a national level, and because of language-specific problems that occur in the different the types of identifying information, it is clear that a separate de-identification method must be developed for each language (Grouin et al., 2009). Although many research into the de-identification problem has been performed in English, a reliable method for de-identifying Dutch medical text has yet to be developed. For the English language, most notably Neamatullah et al. (2008) obtained a recall of 0.967 using their pattern matching method that was developed on a test corpus of 1836 nursing notes. Uzuner et al. (2008) managed to achieve a 0.97 F_1-score on medical discharge summaries, based on a machine learning approach. A hybrid approach was developed by Ferrández et al. (2013), achieving a 0.922 recall by combining both pattern matching and machine learning techniques. Many more approaches in English exist (e.g. Friedlin et al., 2008; Douglass et al., 2005; Fenz et al., 2014).

Apart from text-processing the English language, Velupillai et al. (2009). attempted to port the Neamatullah et al. (2008) algorithm to Swedish, but in their own words with “poor results”. The average score over all de-identified categories was a 0.65 F_1-measure. Over a decade ago already, Ruch et al. (2000), successfully managed to de-identify discharge letters written in French with a recall of 0.98 using their MEDTAG framework for semantic tagging. For notes that are written in Korean and English, Shin et al. (2015) developed a method based on regular expressions that achieved a 0.963 recall on several categories combined. In Dutch, Scheurwegs et al. (2013) managed a recall of 0.89 on a previously unseen dataset with a machine learning approach, achieving reasonable success using limited training data.

As can be seen, the two most common methods to de-identify medical text are pattern matching based or machine learning based (Meystre et al., 2014). In the former, lookup tables and decision rules are used to determine what parts of the text contain identifying information. In the latter approach, machine learning techniques are used to classify each piece of text. A third option is a hybrid approach, combining the two. In literature, it is not immediately clear whether pattern matching or machine learning based methods perform better, although hybrid approaches generally outperform other approaches. A clear downside of the machine learning approach (and therefore also the hybrid approach), however, is the need for a large annotated training corpus, which requires extensive manual labour and is thus expensive to obtain (Neamatullah et al., 2008). For this reason, such a corpus is currently unavailable in Dutch. As Ferrández et al. (2012) furthermore shows, pattern matching based methods may even achieve better recall on unseen data than machine learning approaches. For these reasons, we opt to develop a pattern matching based de-identification method.

The development of our DE-identification method for DUtch mediCal tExt, that we name DEDUCE, will be structured along the Design Science Research Process (DSRP) methodology (Peffers et al., 2006). The first two phases comprise the identification of the de-identification problem and the objectives of a solid de-identification method, largely described above. In the Method section, the phases 3 and 4 concerning the design, development and demonstration of our de-identification method DEDUCE will be described. We will elaborate on the selection of relevant Protected Health Information (PHI) categories as well as the data that we used to develop and test our method, and in detail describe how all PHI categories are de-identified. In the Results and Discussion section finally, we will describe how the method is evaluated and deployed (phases 5 and 6).

2. Method

2.1. Development corpus

The de-identification method is developed and tested on data from the Psychiatry department of the University Medical Center Utrecht (UMCU) in the Netherlands. Specifically, nurse notes and treatment plans in the period January 2012–December 2015 are made available. Nurse notes are written by nurses about all inpatients during each of the three daily shifts, and include information about the current wellbeing and activities of the patients. The treatment plan is typically written at the start of a treatment, and describes the activities that will take place in the context of an admission or outpatient treatment, such as therapies or medication prescription. While the nurse notes tend focus on the daily routine, and thus mention names of patients and treatment staff frequently, a treatment plan has a stronger focus on the long term treatment, and therefore more often mentions locations and other healthcare institutions where a patient has previously been treated or might be referred after treatment. Using both these data sources ensures a diverse corpus is used for developing and testing the de-identification method, and thus improves the generalizability to other medical text written in Dutch. Some more descriptive statistics about the two data sources can be found in Table 1.

To the records of both the nurse notes and the treatment plans the first names and last names of a patient as described in the EHR system are added. The resulting data set comprises a total of 113,553 nurse notes and 4012 treatment plans. From both data sources 1000 records are sampled at random, the 2000 resulting records will comprise the development corpus. To ensure our de-identification method is not biased, the method will be developed on this development corpus without using any other records, and later be validated on a disjoint test corpus.

2.2. PHI selection

Since no clear guideline on which PHIs to remove exists in the Netherlands, we decided to base our method on the strict HIPAA guidelines that are implemented in the USA. In the HIPAA guidelines, 18 patient characteristics are identified: 1) names; 2) all geographic subdivisions smaller than a state; 3) all elements of dates except years and all ages above 89; 4) telephone numbers; 5) fax numbers; 6) electronic mail addresses; 7) social security numbers; 8) medical record numbers; 9) health-plan beneficiary numbers; 10) account numbers; 11) certificate and licence numbers; 12) vehicle identifiers and serial numbers, including licence plate numbers; 13) medical device identifiers and serial numbers; 14) URLs; 15) Internet Protocol (IP) addresses; 16) biometric identifiers including fingerprints and voiceprints; 17) full-face photographic images and any comparable images; 18) any other unique identifying number, characteristic or code. These guidelines do not only apply to the patient, but also to relatives, household members and hospital staff.

During initial exploration of the de-identification problem, by means of manual inspection of the data and conversation with hospital staff, not all of the 18 categories were found to occur in our medical dataset. To come up with a subset of PHIs to de-identify in our method, four health care professionals that work with the text data in our corpus on a regular basis were asked to score each PHI category as occurring “never”, “sometimes” or “regularly” in either the nurse notes or the treatment plan. In these surveys, seven of the 18 PHIs were indicated to be present “sometimes” or “regularly” by at least one of the four health care professionals. During a qualitative follow up, participants indicated that in addition to these seven PHIs, names of institutions where patients are treated can be mentioned in the data as well, possibly revealing something about the identity of a patient. Our de-identification method therefore

Table 1
descriptive statistics about the two data sources used to develop and test the de-identification method.

	Nurse notes	Treatment plan
Nr. records	113,553	4012
Nr. patients	1452	2025
Avg. nr. words	100	658
Type of patient	Inpatient	Inpatient and outpatient
Written by	Nurse	Physician
Type	Free-text	Free-text

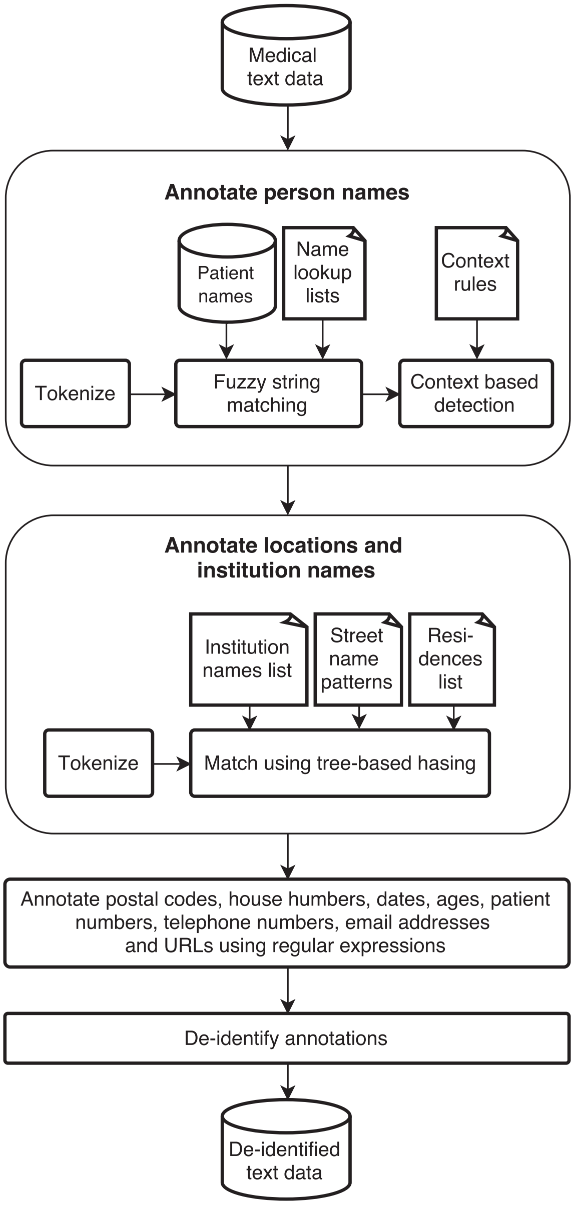

Fig. 1. An overview of the de-identification method DEDUCE.

focuses on the following eight PHI categories - other categories fall outside the scope of this paper.

1. Person names, including initials
2. Geographical locations smaller than a country
3. Names of institutions that are related to patient treatment
4. Dates
5. Ages
6. Patient numbers
7. Telephone numbers
8. E-mail addresses and URLs

In the next subsections, the annotation each of these PHI categories will be elaborated upon. An overview of the de-identification method DEDUCE is depicted in Fig. 1.

2.3. Person names

Person names are the most common PHI in our dataset, including both first names and last names of patients, family members and hospital employees. Only person names are included in this category, names of for example medication are not annotated.

As a preprocessing step, all non-ascii characters are mapped to their ascii counterparts, and all text is tokenized using a custom tokenizer that segregates based on sequences of alpha characters separated by non-alpha characters. The assumption behind tokenizing this way is that person names are always sequences alpha characters with the exception of prepositions that can occur in Dutch names (such as ‘van der‘, which can be abbreviated as ‘v/d’ or ‘v.d.’). Therefore, a list of prepositions is obtained, which will be regarded as one token by the tokenizer.

2.3.1. Non-context-based

The most logical first step in detecting names is using the first names and last names that are available in the EHR system. It is common for a patient to have multiple first names and/or multiple last names. For matching names to tokens, we use fuzzy string matching with a Damerau-Levenshtein (DL) distance of at most 1, which means at most one character insertion, deletion or swap is allowed to transform the string into one of the names. By setting a threshold of 1, essentially all simple misspellings of a name are captured. Although in some rare cases two misspellings occur in a single name, increasing the threshold to 2 greatly increases the number of false positives.

Furthermore, for names with 3 or less characters we require an exact match to occur, since otherwise short Dutch names (such as ‘Jan’) are frequently matched with common Dutch short words (such as ‘aan’ or ‘van’). For matching the names of patients in the EHR system we use three rules:

N1 Tokens that can be matched with a first name are annotated as < PATIENT FIRST NAME > .

N2 Sequences of tokens that can be matched with a sequence of last names (i.e. obtain a fuzzy match for each token) are annotated as < PATIENT LAST NAME > .

N3 A token that is equal to the first character of one of the first names is annotated as < PATIENT INITIAL > .

Most patient names can be annotated using the above three rules, however more challenging are names of co-patients, hospital staff or family members. To be able to annotate these names, we obtain the following lookup lists:

- For first names, we obtain a list of the most common 10,000 Dutch first names, and a list of the most common 10,000 Dutch last names. The number of names that are needed to successfully de-identify text varies in literature from 1.8 million (Thomas et al., 2002) to 10.000 (Velupillai et al., 2009), and although it is unclear where the optimum is, shorter lists appear to perform better. The two lists are obtained from Netwerk Naamkunde of the Meertens Institute (Meertens Instituut, 2016).
- In Dutch, it is not uncommon to have an preposition between the first and last name such as ‘van’, ‘van der’ or ‘in de’. A list of all prepositions in use in the Dutch language is obtained from the internet.
- A list of prefixes, such as ‘dhr’ (sir) or ‘mw’ (madam) is obtained, since they contain valuable clues about a possible presence of a name. During the development of the method, other prefixes such as ‘pt’ (short for patient) and ‘vpk’ (short for nurse) are manually added to this list.
- In order to prevent over-annotation, a whitelist of words that should not be annotated as names is compiled, consisting of the 1000 most common words in Dutch, a list of stop words and a list of medical terms.

For matching tokens with values on lookup lists, exact string matching is used - this includes capitalization. Empirically, fuzzy string matching as described above results in a huge amount of false positives. Using the lookup lists, the following rule for annotating names is added:

N4 Tokens that are on the list of first names or on the list of last names are respectively annotated as < UNKNOWN FIRST NAME > or < UNKNOWN LAST NAME >

2.3.2. Context based

Based on the lists of prefixes and prepositions, clues about possible names enable annotating mentions of unknown persons.

N5 If a token is on the list of prefixes, and the next token starts with a capital, the token is annotated as < UNKNOWN PERSON > . This includes for example occurrences like “Mr. Smith” or “patient Jones”, whose relation to the patient cannot be automatically determined.

N6 If a token is on the list of prepositions, and the next token starts with a capital, the tokens are annotated as < UNKNOWN LAST NAME >

Although a reasonable share of the names present in the text can be annotated using the rules above, a lot of names are still under-annotated, meaning that the annotation needs to be extended to the next or previous token(s) to completely annotate the name. For this, the following rules are added:

N7 If a token is a single capital letter, and the next token is annotated as either a initial or a last name, the token is annotated as an initial. This rule captures initials in front of last names, or multiple consecutive initials.

N8 For tokens that are on the list of prepositions, the context is checked to contain any annotated names. For example, if the previous token is annotated as first name or an initial, and the next token starts with a capital, it is annotated as a last name. Similarly, the token before a preposition can be annotated as a first name or initial, if the next token is annotated as a last name.

N9 If a token is annotated as an initial, first name or last name, and the next token starts with a capital, it is annotated as a last name. Although this rule produces some false positives as well, it captures an important amount of names that were not annotated because they do not occur on the lists of common last names.

N10 Finally, a token that is preceded by an annotated name and the token 'en' (and) is annotated as a name. During development, sentences like "Patient A spent its day with Patients B and C" were observed, where especially the C was commonly not annotated.

Naturally, the rules above only apply to names that were not filtered using the lists of first and last names because they are not so common in the Dutch language. In all cases, the newly annotated tokens are only annotated if they are not on the whitelist. The difference between a patient name and an unknown name is retained, by tagging the new annotation with the appropriate name based on in which context it was annotated.

Based on these ten rules (N1–N10), a very large part of the development corpus is de-identified of names: a manual scanning of a random subset of 500 pieces of text is found to contain 4 false negatives. Exact results will be elaborated upon in the Results section.

2.4. Geographical locations

Locations that are present in the data may not immediately identify a patient, but combined with other data might reveal something about the identity of a patient. For this reason, all addresses in our dataset are annotated. In Dutch, an address contains a street name, a house number with possible suffix, a postal code and a place of residence.

For annotating places of residence a list of places of residence is compiled by combining a list of all Dutch cities and villages, and a list of major European cities. The list contains a total of 2647 places of residence. Street names can be annotated using a regular expression that matches Dutch suffixes such as 'straat' (street), 'laan' (lane) or 'plein' (square).

Since removing all values on the lists from each piece of text is computationally not feasible, for this a trie-based hashing technique is used. Initially, all locations are tokenized in the same way the medical text is tokenized, after which these sequences of tokens are stored in a trie. Then, locations are detected by iterating over all tokens in the text and efficiently matching the longest possible sequence of tokens in the trie using hashing. With this method, a reasonable runtime for the de-identification of locations can be obtained.

Dutch postal codes adhere to a clear format: four digits followed by two characters, e.g. 1234AB. Some minor variations on this format are observed, such as 1234ab or 1234 AB, all these variations of the postal codes are matched using regular expressions.

House numbers and possible additions can also easily be matched using regular expressions, when they are preceded by a street name that is already annotated in the previous step.

Additionally, the development corpus is observed to contain at least one occurrence of a mailbox number, the format for this is the word 'postbus' (mailbox) followed by five digits - again this can be matched using regular expressions.

Occurrences in one of the above categories are annotated with < LOCATION > .

2.5. Names of institutions

It is possible for patients to be treated in multiple care institutions consecutively or even in parallel, especially in psychiatric care this is not uncommon. For this reason, names of care institutions where a patient is treated is regarded as indirectly identifying information that we will annotate in our dataset.

For annotating names of institutions, we combined institution names from the following sources:

- Internal data about the most common institutions where patients are also treated
- Psychiatric care institutions in the region of the UMCU
- Large psychiatric care institutions in the Netherlands
- Institutions that were identified in the text during development of the method

Choosing these lists clearly limits annotating institution names to our specific dataset, and not to Dutch medical text in general. Obtaining a complete list of all health care institutions in the Netherlands however proved to be impossible. Other users of the de-identification method however can easily compile lists that suit their data, thereby enabling the possibility of hand-tailoring the algorithm to the users specific needs.

Although most institutions names that are mentioned in the text are on the list, institutions are not always referred to by their official name in colloquial language. To mitigate this problem, some preparations on the list are performed:

- For institution names that contain articles or prepositions in their name (e.g. 'De Hoogstraat' which translates to 'The Hoogstraat'), the institution name without the preposition (i.e. simply 'Hoogstraat') is also added to the list.
- For institutions with three or more words in their name, the abbreviation of the institution name is also added (i.e. 'University Medical Centre Utrecht' can be abbreviated 'UMCU').
- Some common abbreviations of words are also substituted and added to the list, such as 'zkh' for 'ziekenhuis' (hospital).

Our final list of institution names contains 742 values. Again, to keep things computationally efficient, our trie-based hashing method is used. All institutions that are found are annotated < INSTITUTION > .

2.6. Dates and ages

A date could for example constitute a date of birth or date of admission, which may identify a patient. As in the HIPAA guidelines, only combinations of days and months are regarded as potentially identifying information; years do not need to be de-identified. In Dutch, dates usually follow the day-month-year format. Using the following two patterns, dates can be annotated < DATE > :

- A number between 0 and 31, followed by a number 1–12, possibly followed by a two or four digit number. Between the number groups, white spaces, slashes and dots are allowed.
- A number between 0 and 31, followed by the name of a month or an abbreviation (e.g. ‘Jan’ for ‘January’)

The HIPAA guidelines state that ages over 85 should be removed from medical text, we however opt to remove all ages from the dataset. In combination with other information, the age of a patient may reveal the patient’s identity. Moreover, from the EHR the year of birth can easily be obtained for selections of patients if needed. Using simple patterns such as a number followed by “year” or “year old”, ages can be detected and annotated < AGE > .

2.7. Patient numbers

Patient numbers cannot directly identify information about a patient, but may allow connecting other data sources to the text data in which the patient number is found. Unfortunately, no other structure in a patient number is found than the fact that is a 7-digit number - we therefore match all 7 digit numbers and tag it as < PATIENT NUMBER > . Although this strict rule results in false positives, no other less rigorous matching of patient numbers is possible.

2.8. Telephone numbers, email addresses and URLs

Lastly, telephone numbers, email addresses and URLs are annotated because they can clearly directly identify a patient. For all three, an abundance of regular expressions exist on the web, that after some minor tweaks generally annotate these three PHI types well. Telephone numbers are tagged < TELEPHONE NUMBER > , both email addresses and URLs are tagged as < URL > .

2.9. De-identifying the annotations

After annotating the PHIs, de-identification can take place. Although it is possible to simply remove all annotated PHIs from the text or replace them with a default string, this reduces the legibility of the text and is thus not preferable. We therefore choose the following method to de-identify the annotations:

- First, all adjacent annotations are merged into a single annotation if the annotation type matches. This ensures that no tags like < NAME > < NAME > occur in the text.
- For patient names, all occurrences of < PATIENT FIRST NAME > , < PATIENT LAST NAME > and < PATIENT INITIAL > are replaced by simply < PATIENT > .
- For all other tags, the tag is replaced by the name of the tag with a number that uniquely identifies the occurrences of the same value. For example, all occurrences of the name of a specific nurse within one piece of text are replaced with < PERSON-1 > , while all occurrences of the name of a co-patient in that text are replaced with < PERSON-2 > , etc. For person names, geographical locations and names of institutions, fuzzy string matching is used to make sure misspellings or slightly different spellings are matched to the same tag. For all other PHI categories, exact string matching is used.

3. Results and discussion

The final de-identification method is validated on a test corpus that consists of a random sample of 200 nurse notes and 200 treatment plans. The test corpus is fully disjoint from the development corpus, in other words no text that is used to develop the method can be selected in the test corpus. The test corpus is annotated using our de-identification method, after which the annotations are validated by a human rater. This is done by visually presenting a piece of text with annotations marked in different colors for each PHI category. The rater then for each category specifies the number of false positives (i.e. pieces of text that were erroneously marked PHI) and false negatives (i.e. pieces of text that are not marked as PHI while they contain a PHI). If a PHI is correctly annotated, but in the wrong category, a false positive is also registered.

Furthermore, since all pieces of text are written about a single patient, it regularly occurs that the same name of a patient is written multiple times and is thus annotated multiple times. Counting all of these occurrences as separate true positives would strongly overestimate the performance of the method, and therefore only the number of unique annotations in each category is counted. Consequently, false positives and false negatives are also only uniquely counted in one piece of text.

To verify that one human rater can accurately validate the annotations, a random selection of 150 pieces of text from the test corpus were validated by a second rater. Although some minor differences occurred, no major false negatives were overlooked and all precision, recall and F_1-scores matched to within a small tolerable margin of error.

The results of the validation on the test corpus can be seen in Table 2.

Table 2
For each PHI category, the results for de-identifying the test corpus (n = 400). The totals displayed are based on micro-averaging over the PHI categories. For the micro-averaged precision and recall respectively, the standard errors are 0.017 and 0.012.

	Nr. of PHIs	Precision	Recall	F_1-score
Person names and initials	270	0.742	0.964	0.839
Geographical locations	13	1	0.867	0.929
Names of institutions	102	0.99	0.756	0.857
Dates	99	0.78	0.98	0.868
Ages	50	0.98	0.98	0.98
Patient numbers	5	1	1	1
Telephone numbers	3	1	0.6	0.75
E-mail addresses/URLs	0	–	–	–
Total	542	0.814	0.916	0.862

From Table 2 it can be seen that the de-identification method achieves generally good results, with a micro-averaged precision of 0.814, a recall of 0.916 and a F_1-score of 0.862. The recall of the method shows how likely it is for a PHI to be missed, and thus how likely it is for re-identification to occur, while the precision measures the amount of false positives, thereby estimating the amount of information loss that is a result of applying the method. These results are in line with research in other languages. More importantly, for patient names, which is the most directly identifying PHI, a good recall of 0.964 was achieved, while no patient names were missed. These numbers combined show that automatic de-identification of medical text written in Dutch is possible to a high degree using our method DEDUCE.

A total of 10 occurrences of person names were missed by the method. In one case, this was the name of a school teacher which was misspelled. In other cases, names of hospital staff that were not capitalized or abbreviated in an uncommon way (such as concatenating initials with a last name) were not properly de-identified. One occurrence of a first name, and four occurrences of last names were missed, other occurrences concerned only initials. No person names of patients were missed by the method. For locations, two misspellings of cities were missed. The most tricky PHI to correctly annotate are names of institutions where patients are treated, with a recall of 0.756. In many cases, different spellings, abbreviations or colloquial variants of the names of institutions made annotating difficult, and in some cases the name of an institution was not on one of the lookup lists. For telephone numbers, in one piece of text the last two digits of two phone numbers in a format from a different country were not properly included in the annotation, resulting in two false negatives. Although the recall score of 0.6 seems low, this can solely be attributed to these two phone numbers, where the text after de-identification still does not reveal any identifying information other than the two last digits of the phone numbers.

For person names, both over-annotation as well as some annotation of text that are no names takes place. In the first case too much text is annotated, this mostly concerns person names followed by a word with a capital letter, in the latter case person names that are also Dutch words are annotated. For dates, some false positives occurred when the numerical part medication dosages (such as 2.5 milligram) were annotated as a date. This can relatively easily be detected in an improved version of the method, but is not included in the validation. In other categories, very few false positives occurred. Although some email addresses and URLs were present in the development corpus, none were present in the test corpus and thus no score is added for this category.

Finally, it must be addressed that for the geographical locations, patient numbers and telephone numbers, relatively few occurrences were found in the text. Although the recall shows that most of these occurrences were found by the algorithm, the small sample size affects the reliability of the results in these categories. This could be a topic of further research, using a dataset that is more rich in these categories.

3.1. Generalizability to other Dutch medical text

Our method strives to be applicable to Dutch medical text in general, some fitting to the dataset that we used to develop and test the method is however inevitable.

De-identification of person names should generalize well to data of other institutions, because it relies on national lists of popular names and generic rules. It must be noted that this part of the de-identification process partly relies on the patient names that are registered in the EHR system. In case this data is not available, performance of the method may decrease. For institution names, de-identification heavily relies upon the available list of treatment institution names. Without a list that is specific to a health care institution, performance will likely not be very good - so for this PHI category obtaining a good list of institution names is essential. For the de-identification of geographical locations, dates, ages, telephone numbers, email addresses and URLs, generalizability is expected to be good, since very little information specific to our dataset was used. Patient numbers finally will most likely not be easily detected in other datasets, since they may not follow the same 7-digit format. This part of the method can however easily be hand-tailored.

To conclude, most of the method has been designed to be simple and generic, and to thusly be more generalizable to other Dutch medical text. This is also supported by for example Ferrández et al. (2012), who showed that pattern matching based methods generally obtained a good recall on unseen datasets. The method is generic with respect to finding PHIs in all categories, except names of institutions and patient numbers, which can be hand tailored to a specific data set with relative ease.

4. Conclusion

There is no doubt that medical text data holds great potential for research, and that it can be a great asset in creating direct clinical value by applying data analysis techniques. Using text data however puts a strain on patient privacy: in many cases, identifiable information about patients or patient family is mentioned in text data, which should not be accessible by data analysts. Even more serious is the risk of a potential data breach, in which private details of a treatment could become known to the public.

To mitigate these drawbacks of using medical text data for research, we set out to develop a de-identification method, that we name DEDUCE: DE-identification method for DUtch mediCal tExt.

In cooperation with local medical staff, we decided to focus on 1) Person names, including initials; 2) Geographical locations smaller than a country; 3) Names of institutions that are related to patient treatment; 4) Dates; 5) Ages; 6) Patient numbers; 7) Telephone numbers and 8) E-mail addresses and URLs. To develop and test the de-identification method, treatment plans and nurse notes from the Psychiatry department of the University Medical Center Utrecht (UMCU) in the Netherlands were used to create a development corpus (n = 2000) and a disjoint test corpus (n = 400).

Our de-identification method for person names relies on the names of patients that are known in the EHR system, lookup lists of Dutch first and last names, and subsequently applies fuzzy string matching and context based rules to annotating the names. For geographical locations and names of institutions, a tree-based hashing method was used to efficiently annotate all occurrences on lookup lists. For the other categories, regular expressions were employed to match all PHI occurrences. After annotating all PHIs, de-identification took place by replacing all annotations with appropriate de-identified strings.

Validation of our method DEDUCE on the test corpus shows generally good results, with a total micro-averaged precision of 0.814, a recall of 0.916 and a F_1-score of 0.862. Another notably good result of the method is a recall of 0.961 for names, only missing names of treatment staff and no single patient name. Although our method is to some extent fitted to our specific dataset, we believe it will be applicable to medical text written in Dutch in general, after some simple extra preparation to hand tailor the algorithm. The validation furthermore shows that de-identification of medical text written in Dutch in an automated manner using our method DEDUCE is possible with good results.

Implementation code

A Python implementation of the de-identification method is made available on https://github.com/vmenger/deduce/

Conflicts of interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

Borking, J.J., Raab, C., 2001. Laws, PETs and other technologies for privacy protection. J. Inf. Law Technol. 1, 1–14.

Deleger, L., Molnar, K., Savova, G., Xia, F., Lingren, T., Li, Q., Ellipsis Solti, I., 2013. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. J. Am. Med. Inform. Assoc. 20 (1).

Douglass, M.M., Cliffford, G.D., Reisner, A., Long, W.J., Moody, G.B., Mark, R.G., 2005. De-identification algorithm for free-text nursing notes. In: Comput. Cardiol. IEEE, pp. 331–334.

El Emam, K., Jabbouri, S., Sams, S., Drouet, Y., Power, M., 2006. Evaluating common de-identification heuristics for personal health information. J. Med. Internet Res. 8 (4), e28.

European Data Protection Directive (1995). Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.

Fenz, S., Heurix, J., Neubauer, T., Rella, A., 2014. De-identification of unstructured paper-based health records for privacy-preserving secondary use. J. Med. Eng. Technol. 38 (5), 260–268.

Ferrández, O., South, B.R., Shen, S., Friedlin, F.J., Samore, M.H., Meystre, S.M., 2012. Evaluating current automatic de-identification methods with Veteran's health administration clinical documents. BMC Med. Res. Methodol. 12, 109.

Ferrández, O., South, B.R., Shen, S., Friedlin, F.J., Samore, M.H., Meystre, S.M., 2013. BoB, a best-of-breed automated text de-identification system for VHA clinical documents. J. Am. Med. Inform. Assoc.: JAMIA 20 (1), 77–83.

Friedlin, F.J., McDonald, C.J., South, B.R., Shen, S., Samore, M.H., Dorr, D., Ellipsis Uzuner, O., 2008. A software tool for removing patient identifying information from clinical documents. J. Am. Med. Inform. Assoc. 15 (5), 601–610.

Grouin, C., Rosier, A., Dameron, O., Zweigenbaum, P., 2009. Testing tactics to localize de-identification. Stud. Health Technol. Inform. 150, 735–739.

Harpaz, R., Callahan, A., Tamang, S., Low, Y., Odgers, D., Finlayson, S., Ellipsis Shah, N.H., 2014. Text mining for adverse drug events: the promise, challenges, and state of the art. Drug Saf. 37 (10), 777–790.

HIPAA (1996). Health Insurance Portability and Accountability Act of 1996.

Jensen, P.B., Jensen, L.J., Brunak, S., 2012. Mining electronic health records: towards better research applications and clinical care. Nat. Rev. Genet. 13 (6), 395–405.

Koh, H.C., Tan, G., 2005. Data mining applications in healthcare. J. Healthcare Inform. Manage.: JHIM 19 (2), 64–72.

Krishna, R., Kelleher, K., Stahlberg, E., 2007. Patient confidentiality in the research use of clinical medical databases. Am. J. Public Health 97 (4), 654–658.

Lee, J., McCullough, J.S., Town, R.J., 2013. The impact of health information technology on hospital productivity. Rand J. Econ. 44 (3), 545–568.

Liu, V., Musen, M.A., Chou, T., J, A.-M., 2015. Data breaches of protected health information in the United States. JAMA 313 (14), 1471.

Maenner, M.J., Yeargin-Allsopp, M., Braun, K.V.N., Christensen, D.L., Schieve, L.A., Newschaffer, C., Ellipsis Foldy, S., 2013. Development of a machine learning algorithm for the surveillance of autism spectrum disorder. PLoS ONE 11 (12), e0168224.

Meertens Instituut Netwerk Naamkunde. (n.d.). Retrieved December 29, 2016, from http://www.naamkunde.net/?page_id=289.

Menger, V., Spruit, M., Hagoort, K., Scheepers, F., 2016. Transitioning to a data driven mental health practice: collaborative expert sessions for knowledge and hypothesis finding. Comput. Math. Methods Med. 2016, 1–11.

Meystre, S.M., Ferrández, Ó., Friedlin, F.J., South, B.R., Shen, S., Samore, M.H., 2014. Text de-identification for privacy protection: a study of its impact on clinical text information content. J. Biomed. Inform. 50, 142–150.

Milovic, B., 2012. Prediction and decision making in Health Care using Data Mining. Int. J. Public Health Sci. (IJPHS) 1 (2).

Murdoch, T.B., Detsky, A.S., 2013. The inevitable application of big data to health care. JAMA 309 (13), 1351.

Neamatullah, I., Douglass, M.M., Lehman, L.H., Reisner, A., Villarroel, M., Long, W.J., Ellipsis Clifford, G.D., 2008. Automated de-identification of free-text medical records. BMC Med. Inform. Decis. Mak. 8, 32.

Patel, R., Jayatilleke, N., Jackson, R., Stewart, R., McGuire, P., 2014. Investigation of negative symptoms in schizophrenia with a machine learning text-mining approach. Lancet 383, S16.

Peffers, K., Peffers, K., Tuunanen, T., Gengler, C. E., Rossi, M., Hui, W., Ellipsis Bragge, J., 2006. The Design Science Research Process: A Model for Producing and Presenting Information Systems Research. 1st International Conference on Design Science in Information Systems and Technology (Desrist, 83–106).

Ruch, P., Baud, R. H., Rassinoux, A. M., Bouillon, P., Robert, G., 2000. Medical document anonymization with a semantic lexicon. Proceedings. AMIA Symposium, 729–733.

Scheurwegs, E., Antwerpen, A. H., Luyckx, K., Schueren, F. Van Der, Bulcke, T. Van Den, 2013. De-Identification of Clinical Free Text in Dutch with Limited Training Data: A Case Study.

Shin, S.-Y., Park, Y.R., Shin, Y., Choi, H.J., Park, J., Lyu, Y., Ellipsis Hripcsak, G., 2015. A de-identification method for bilingual clinical texts of various note types. J. Korean Med. Sci. 30 (1), 7.

Simon, G.E., Unützer, J., Young, B.E., Pincus, H.A., 2000. Large medical databases, population-based research, and patient confidentiality. Am. J. Psychiatry 157 (11), 1731–1737.

Thomas, S. M., Mamlin, B., Schadow, G., McDonald, C., 2002. A successful technique for removing names in pathology reports using an augmented search and replace method. Proceedings. AMIA Symposium, 777–781.

Toledo, C. van, Spruit, M., 2016. Adopting privacy regulations in a data warehouse: A case of the anonymity versus utility dilemma. In Fred, A. et al. (Ed.), Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (pp. 67–72). KDIR 2016, November 11–13, 2016, ScitePress, Porto, Portugal.

Uzuner, O., Sibanda, T.C., Luo, Y., Szolovits, P., 2008. A de-identifier for medical discharge summaries. Artif. Intell. Med. 42 (1), 13–35.

Velupillai, S., Dalianis, H., Hassel, M., Nilsson, G.H., 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. Int. J. Med. Informatics 78 (12), e19–e26.