Anonymization of General Practioner Medical Records

Amund Tveit[a,c], Ole Edsberg[a,c], Thomas Brox Røst[a,c], Arild Faxvaag[a,b], Øystein Nytrø[a,c], Torbjørn Nordgård[d], Martin Thorsen Ranang[c] and Anders Grimsmo[a,b]

[a]*Norwegian Center for Patient Record Research, Faculty of Medicine, Norwegian NTNU, NO-7489 Trondheim, Norway*
[b]*Faculty of Medicine, NTNU, NO-7489 Trondheim, Norway*
[c]*Department of Computer and Information Science, NTNU, NO-7491 Trondheim*
[d]*Department of Linguistics, NTNU, NO-7491 Trondheim, Norway*

Abstract

The Electronic Patient Record (EPR) is both a legal document and a tool for use by physicians and other health personnel during provision of health care. Its primary purpose is to provide and store information about the patient in clinical settings, but it's also a source of medical knowledge (e.g. epidemiology and quality of care). Due to the sensitive nature of the data they must be handled in a secure manner with a high awareness of privacy concerns. This problem can be partially avoided by applying an anonymization procedure to the data. For large volumes of data (e.g. thousands of patient records) such a procedure must be partially automated. We aim to develop techniques and methods for semi-automated anonymization of medical record information. We first present the requirements and goals of anonymization. Relevant goals for designing an anonymization method are: complying with national laws, and making the anonymization as automated as possible. We discuss anonymization challenges, including linguistic issues (e.g. spelling and ambiguity) and determining which parts of the data that is sensitive. Finally we propose methods including utilization of database structure, dictionaries, heuristics and natural language processing for anonymizing patient records in general, but with focus on general practioner records gathered from a Profdoc Vision database.

Keywords[1]: Anonymization, De-identification, Electronic Patient Records, Computational linguistics

Introduction

Electronic medical records, in addition to serving as legal documents and tools for clinical workers, are of great interest for medical research. They contain evidence of what actually happens in medical care, information that if properly extracted and presented would be valuable in deciding how to improve medical care. Due to the private and potentially damaging information present in medical records, there are government regulations controlling who can access them and for what ends they can be used. We will not discuss these regulations here, except for noting that strategies for protecting the privacy of the patients must be planned and documented before applying for permission to use medical records in research. One such strategy is *anonymization*; the act of modifying the data in order to make it harder to connect the records to the actual people they describe.

When designing an anonymization method, the following are principal goals:

1. The method should provide the level and type of privacy protection that is required from it.
2. There should be an easy and convincing way of demonstrating that the method satisfies the requirements.
3. The method should be feasible to carry out. For large amounts of data, a high degree of automation is necessary.
4. The method should not cause any more distortion or loss of the information in the record than necessary.

(Berman 2002)[1] give a general review of privacy issues and privacy protection methods for medical data mining, our paper differs from his since we're primarily focusing on personally-identifying information in Norwegian language medical records.

Terminology

Personally-identifying information is information that can be used to connect a medical record to an identified person. ***Sensitive information*** is information that could be damaging to a person if it became know by other people. A piece of data is said to ***de-identified*** when it is ***stripped*** of personally identifying information. If personally-identifying information is consistently replaced by aliases, the data is said to have been ***pseudonymized***. ***Computational linguistics*** is a subfield of linguistics that is concerned with the computational modeling of natural language. ***Natural language processing*** is a disci-

[1] Keywords should, preferably, be drawn from the MeSH thesaurus

pline that is closely related to computational linguistics and focuses on the processing and manipulation of natural language. ***Unrestricted text*** is text that has few if any enforced rules governing its content and structure.

Related Work

Existing approaches to the problem of anonymizing electronic medical records differ in the types of records they are constructed for, in the type and scope of the anonymization task they aim to perform, in the methods they apply and in the way they evaluate their results.

The ***Scrub*** system (Sweeney 1996)[2] defines the anonymization task, which they call "scrubbing", as finding and replacing personally-identifying information. Their (English language) data set consisted of 275 patient records and 3198 letters to referring physicians. According to the article, the data contains large amounts of unrestricted text with spelling and grammatical errors, nicknames and cryptic abbreviations. Their system is based on a study of how manual anonymization is performed and consists of a set of detection algorithms that each attempts to detect a particular type of entity. The algorithms are apparently based on common-sense information encoded as templates and word lists. They report that their system found 99-100% of personally-identifying references.

(Ruch 2000)[3] presents another system for locating and removing personally-identifying information. The data set consisted of 600 post-operative reports, 200 laboratory and test results and 200 discharge summaries. One would guess that the free text in these types of documents is relatively diverse and unrestricted. The authors claim that their system performs the same type of task as the Scrub system and that the distinguishing feature of their approach is the use of natural language processing (NLP) tools. The NLP tools used are a medical semantic lexicon, a word-sense tagger and a morpho-syntactic tagger (part-of-speech). It appears that the approach relies on detecting so-called identity markers like "Doctor" and "Ms." and using the taggers to resolve any ambiguities. 20% percent of the data was used to iteratively test and modify rules in the system; the remaining 80% was used for evaluation. The authors claim that their method found 98-99% of all personally-identifying information.

(Taira 2002)[4] presents an approach that uses a maximum entropy classifier operating with respect to semantic constraints generated from a manually tagged training set to determine which words in urology reports are patient names.

(Thomas 2002)[5] presents a tool for finding and replacing proper names in pathology reports. It uses a list of medical terms and a list of proper names, and it exploits the fact that proper names often come in pairs, often start with a capital letter and often are preceded by honorifics such as "Dr." The authors claim that their method finds 98.7% of the proper names in their data set

The ***Concept-Match*** algorithm (Berman 2003)[6] attempts to solve the tasks of removing both personally-identifying information and other information that is incriminating or otherwise private. Their data sets are a set of 567 921 pathology phrases and a set of invented free text sentences. Their method consists of performing the following actions in order: parsing the input into sentences and then into words, detecting and preserving common stop words in their positions in the sentences, detecting medical terms from a metathesaurus (UMLS 2003)[7] and replacing them with synonyms and codes when possible and, finally, replacing all the remaining terms with the marker "***". They claim that their method works perfectly according to the specification of what sorts of words should be removed, but they note that breaches of privacy is still possible if the compromising words are terms that also have medical meanings. Another problem with the approach is that the information content in the text will likely be damaged. Important information may be written with words that are not medical terms or that are not in the metathesaurus, and the removal of words makes sentences hard to parse and even to interpret by inspection. These problems will likely be more severe when the method is applied to texts that contain more varied and general information that pathology reports. This can be seen from the examples given of the results of applying the method to invented free text sentences, where in all cases the whole meaning of the sentence seems to be lost. Another problem mentioned is that the replacement of medical terms with synonyms and codes can be a source of errors when faced with ambiguity in terminology.

Note: *When attempting to compare the approaches described above, it becomes clear that the reported percentage of personally-identifying references found is not a good measure, since both the properties of the data sets and the types of entities that are counted as personally-identifying may vary. Such a measure also doesn't take into account the destructive effects of removing false positives from the data. A good measure would be a blind test on a common data set, but the sensitive nature of authentic test data would make this difficult to carry out.*

(Gupta 2004)[8] gives an interesting description of the possible interplay between the evaluation and construction of an anonymization system. The system itself is not thoroughly described in the article, but it appears to use metathesauri, including (UMLS 2003)[7], and manually designed rules. Pathology reports were used as the data set. The evaluation was performed by pathologists in three cycles of 300 to 1000 reports each. In the first and second cycles, defects were detected and fixed. After the third cycle, the system was said to perform extremely well. It seems to us that such an iterative approach is almost a necessity in designing an anonymization system, for it will be very hard to correctly guess all the possible personally-identifying entities in the design phase. When performing evaluate/fix iterations, it is very important that the evaluators are only given data that has not been used in previous iterations. Otherwise the evaluation scores would likely be inflated by adapting the rules to the particular data set in use.

(Øhrn and Machado 1999)[9] describes a mathematically well-founded method for anonymizing databases by altering the data such that "for each patient in the database, there has to be at least one other patient from which the first patient cannot be discerned, at least with respect to the database fields that are most likely to be used for linking." (Vinterbo 2003)[10] considers the task of finding the least destructive modification that will make it impossible for an adversary that only has access to the modified data to connect sensitive information to identifying information in the database. He then formally defines the task as an optimization problem and gives a proof that the problem as defined is NP-hard. Consider a database column with fields of unrestricted text that possibly contains personally-identifying references. Each field will most likely be unique. Since the methods of (Øhrn 1999)[9] and (Vinterbo 2003)[10] operate at the field level, it will be impossible or very difficult to anonymize the database without destroying the useful content in the same fields.

Our Anonymization Approach

How does our planned approach, as described in this article, distinguish itself from these other systems? First, our texts use a different language, which will mean that linguistic aspects will be different. For example, identity markers (e.g. Dr. and Mrs.) that were used in the approaches (Ruch 2000)[3] and (Thomas 2002)[5] are not as prevalent in Norwegian, and Norwegian hyphenation patterns are also different**.** Second, our data set consists of general practitioner patient records. Because of the *general* nature of general practice, lots of valuable information is entered as unstructured text. Furthermore, we cannot count on there having been strictly enforced guidelines for where and how personally-identifying information should be encoded, meaning that the entire database must be suspected. The Concept-Match (Berman 2003)[6] algorithm would certainly have a devastating effect on the information content of the input when applied to our data rather than pathology reports. Third, in addition to simple NLP and external dictionaries, we plan to use dictionaries built from our own corpus and modified through reference with other dictionaries and by manual inspection. This ***semi-automatic corpus-driven approach*** is possible because we plan to use our dataset for extensive research and therefore are willing to put in the effort required.

The nature of the anonymization task and the nature of electronic medical records give rise to many problems that must be overcome in designing an anonymization method. An overview of our proposed 6-step anonymization approach is shown in the figure below. In order to anonymize the patient record data we need to replace or remove words or numbers that can be used to identify persons. Examples of such words are geographical locations and person names.

Step A – Dictionaries

Our goal is to anonymize general practioner data - both structural and free-text - found in the patient record system Vision provided by the vendor Profdoc.

Profdoc Vision uses a relational Oracle database for storing both the doctor's notes as well as all related, mainly structural, information (e.g. patient personalia, prescriptions and medications, data for governmental social security institutions, etc.). Words and numbers found in structural data found in Profdoc Vision can extracted into *local dictionaries* with corresponding type (e.g. of patient names, social security numbers, postal codes, health institution names, health personnel names and locations). In addition one should create dictionaries of local acronyms, initials, abbreviations and terms; this could be done by interviewing the local doctor(s).

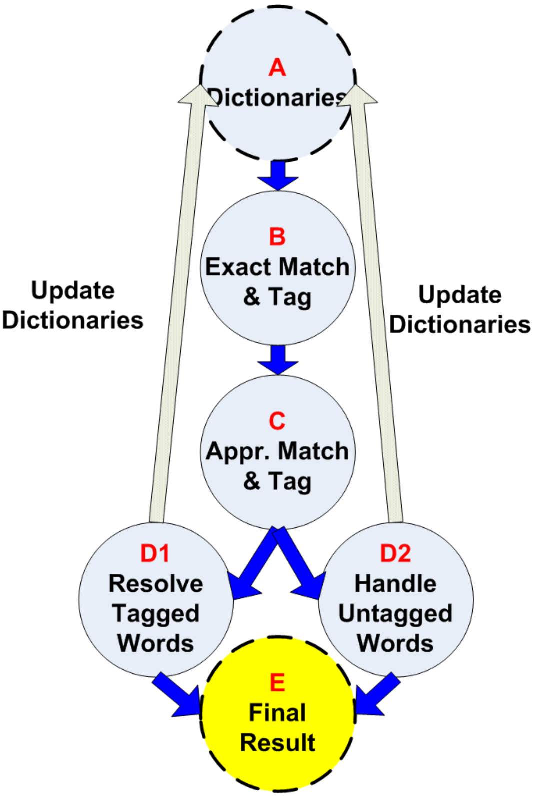

Figure 1 – Our anonymization process

In order to reduce the large amount of free-text from doctor's notes, social security notes and discharge summaries we choose to count occurrences of unique words found in the database, this count is called a unigram. With this unigram we also get an overview of the whole patient record's number of unique words and statistics for how often they occur. By removing common non-sensitive and relatively non-informing words, the amount of free-text data can be significantly reduced.

Finally we gather external dictionaries from various sources:

- Medical Names (e.g. ICPC)
- Medication Names (e.g. the Norwegian PAS)
- Norwegian person names (e.g. Statistics Norway and National Register)
- Geographical names and locations (e.g. National Map and Postal Services)
- Names of Health Institutions (e.g. Directorate for Health and Social Affairs)
- Statistics for activities and occupations (e.g. Statistics Norway)
- Norwegian Language e.g. from Norcomplex or Kunnskapsforlaget
- Word co-occurences and frequencies in general, e.g. by gathering statistical information about words

found in traditional Norwegian Dictionaries by counting their occurrences on all or a subset of Norwegian web sites.

Step B – Exact Match and Tag

In this and the following steps we focus on processing the unigram created from the free-text notes. Based on all the local and external dictionaries we create a combined dictionary where one can look up words and get their corresponding type(s), in order to achieve high computational performance for dictionary lookups we use a suffix tree (also called a trie).

We then proceed with looking up each word in the unigram in the combined dictionary and if there is match add a type tag to the word in the unigram, e.g. *<tag type=medication>* ***paracetamol*** *</tag> and <tag type=surname>* ***Olsen*** *</tag>*. There can also be words with multiple tags, e.g. *<tag type=noun,firstname>* ***bjørn*** *</tag>*

In order to tag non-textual symbols such, e.g. dates, phone numbers and social security numbers, we apply regular expression matching.

The Norwegian language, as opposed to the English, contains a significant amount of composite words. One example is the Norwegian word "**bjørnejakt**" (bjørn+jakt) as opposed to the English expression ("**bear hunt**"). In theory the Norwegian language allows infinite concatenation of nouns in order to create a new Norwegian noun, but in practice it is only a few (rarely more than five). For Norwegian it means that dictionaries are unlikely to be complete, but they might be relative complete with respect to each noun of composite words. Other issues that have to be handled are inflected forms and joint characters (e.g. the 'e' in the combination of "**bjørn**" and "**jakt**"). By using the combined dictionary together with a finite-state automata one can tag composite words, and since each composite word contains several words they would naturally have more than one tag type, e.g. *<tag containstype=noun,firstname,verb>**bjørnejakt**</tag>*

This step results in the unigram being tagged with direct matches of words and symbols found in the combined dictionary or by the regular expressions. But many words are likely to not be tagged during this step, e.g.:

- Words with spelling errors (e.g. ***paracetamol*** *incorrectly spelled as* ***paracetemol***)

Step C – Approximate Match and Tag

Patient records may contain erroneously spelled words, and in many cases they might be only slightly incorrectly spelled. Edit distance (also called Levenstein distance) is a measurement of how many character replacements, additions and removals a word must undergo in a transition to another word. By going through untagged words in the unigram and allowing an edit distance of 1 one can find candidate misspellings and relate them to the combined dictionary. This can be handled by transforming the combined dictionary suffix tree into an approximate matching suffix tree (Navarro 2001)[11], and then go through each untagged word and do a lookup in the approximate match suffix tree. Note that composite words with spelling errors would not be tagged by step B or C, but they are likely to be few. There will likely be other untagged words not found in the dictionaries, but they are likely to be of a small number.

Step D1 – Resolve Tagged Words

For words in the unigram with a single tag it has to be replaced if it is an identifying name (e.g. person name), otherwise not. When words have multiple type tags, they have to be replaced or removed if one or more of the tags are of the identifying type. But there could be cases when there are multiple tags that the word probably shouldn't be replaced or removed (e.g. *<tag type=noun, firstname>* ***bjørn*** *</tag>*, this has to be investigated further and adapted to the current data sets.

Step D2 – Handle Untagged Words

Untagged words in the unigram needs to be investigated manually by the local clinician for tagging, the result of this investigation could be additional entries for the local or external dictionaries.

Step E – Final Result

The final result contains patient record text with identifying entities replaced by pseudonyms. In order to ensure an acceptable level of anonymization the result must be validated according to the requirements that motivated the anonymization in the first place.

Conclusion and Future Work

We have presented our work in progress on the development of an anonymization approach of Norwegian-language general practioner records.

Ongoing work is the implementation and empirical testing of our approach on medical records from a medium-size Norwegian municipality. While doing so, it is likely that our understanding of the problem will increase, leading to revisions of our method. After gaining that understanding we would also like to turn our method into a framework for deciding which words to remove when anonymizing texts. That would involve making the dictionary selection, tagging, tag reduction and rule generation steps more systematic and general so that it would be easy to adapt the framework to new tasks with other data sets, resources and requirements.

We also have in mind some specific ways in which our approach can be extended and improved: Our method currently ignores the textual context in which a word appears in the record. For example, some words can have different meanings depending on the grammatical role they have in a sentence. With a word-sense tagger, this information could be taken into account in the classifying words as sensitive or not sensitive. Since we would then be classifying words at the level of word instances in the text rather than at the level of dictionary entries, tagging would have to be carried out in two stages, first on the dictionary entries in order to find the subset of words that must be checked manually and then on the instances in the text in order to perform the final decision of which words to remove.

Acknowledgments

This work has been supported in part by grants from the Norwegian Research Council.

References

[1] Berman J. Confidentiality Issues for Medical Data Miners. Artificial Intelligence in Medicine, 2002.

[2] Sweeney L. Replacing Personally-Identifying Information in Medical Records, the Scrub System. Proc AMIA, 1996.

[3] Ruch P, Baud RH, Rassinoux A-M, Bouillon P, Robert G. Medical Document Anonymization with a Semantic Lexicon. Proc AMIA Symp, 2000.

[4] RK Taira et. al. Identification of patient name references within medical documents using semantic selectional restrictions. Proc AMIA Symp., 2002.

[5] SM Thomas et. al. A Successful Technique for Removing Names in Pathology Reports Using an Augmented Search and Replace Method. Proc AMIA Symp,. 2002.

[6] Berman J. Concept-Match Medical Data Scrubbing. How Pathology Text Can Be Used In Research. Arch Pathol Lab Med, 2003.

[7] Berman JJ. A tool for sharing annotated research data: the "Category 0" UMLS (Unified Medical Language System) vocabularies. BMC Med Inform Decis Mak. 2003.

[8] Gupta D, Saul M and Gilbertson J. Evaluation of a Deidentification (De-Id) Software Engine to Share Pathology Reports and Clinical Documents for Research. Am J Pathol, 2004.

[9] Vinterbo SA. Privacy: A Machine Learning View. IEEE Transactions on Knowledge and Data Engineering, 2003.

[10] Vinterbo S A. Privacy: A Machine Learning View. IEEE Transactions Knowledge and Data Engineering, 2003.

[11] Gonzalo Navarro. A Guided Tour of Approximate String Matching. ACM Computing Surveys, No. 1, March 2001, pp.31-88

Address for correspondence:

Dr. Amund Tveit, IDI/NTNU, NO-7491 Trondheim, Norway
amund.tveit@idi.ntnu.no – http://www.idi.ntnu.no/~amundt/
phone: +47 416 26 572 – fax: +47 7359 4466