Adverse drug event classification of health records using dictionary-based pre-processing and machine learning

Stefanie Friedrich and Hercules Dalianis
Department of Computer and Systems Sciences (DSV)
Stockholm University
P.O. Box 7003
164 07 Kista
Sweden
stfr6041@student.su.se, hercules@dsv.su.se

Abstract

A method to find adverse drug reactions in electronic health records written in Swedish is presented. A total of 14,751 health records were manually classified into four groups. The records are normalised by pre-processing using both dictionaries and manually created word lists. Three different supervised machine learning algorithm were used to find the best results; decision tree, random forest and LibSVM. The best performance on a test dataset was with LibSVM obtaining a precision of 0.69 and a recall of 0.66, and a F-score of 0.67. Our method found 865 of 981 true positives (88.2%) in a 3-class dataset which is an improvement of 49.5% over previous approaches.

1 Introduction

Adverse drug events (ADEs) are the seventh largest cause of death in Sweden (Wester et al., 2008). ADEs also cause 3.7% of hospital admissions worldwide (Howard et al., 2007). Drugs have been developed by pharmaceutical companies and tested on a small group of healthy young men (~3,000) (FDA, 2014); however, patients taking the drugs are mostly elderly and multiple sick people, therefore, one needs to perform post-marketing drug safety surveillance in order to detect the ADEs' effect on real patients.

The care of the patient is continuously documented by the physician in health records also called electronic patient records. The health records contain both structured and unstructured information. The structured information is for example the age of the patient, time stamps, drugs, ICD-10 diagnosis code, and microbiological and blood values. Unstructured information is mainly free text. Health records are usually long and written by different authors with different writing styles (Allvin et al., 2011; Wijesekera, 2013). To identify entities in a text, to extract meaning and terms, and to consider their context, advanced methods must carried out. Several strategies and methods have been developed, and some approaches are described in the next section.

2 Related research

There are several studies on automatically identifying ADEs from the text of electronic health records using either rule-based or machine learning-based methods. In this section different approaches and their results are summarised.

2.1 Rule based methods

Several rule-based studies to detect ADEs have been carried out. Eriksson et al. (2013), carried out a rule and dictionary approach to detect ADEs in 6,011 Danish psychiatric patients' hospital records. The system identified 35,477 unique ADEs. They obtained a precision of 0.89 and a recall of 0.75.

Wang et al. (2009) developed a rule-based system to detect the *drug - ADEs relationship* for seven specific drugs. They used 25,074 discharge summaries in English to evaluate the system. The authors obtained a recall and precision of 0.75 and 0.31, respectively, for known ADEs.

Hazlehurst et al. (2009) detected vaccine ADEs among Kaiser Permanente Northwest (KPNW), which encompasses more than 450,000 persons. They compared automated methods MediClass with code-based detection methods; the MediClass method obtained better results than the code-based method - 0.74 versus 0.31 PPV (positive predictive value, which is the same as precision).

Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis (Louhi), pages 121–130,
Lisbon, Portugal, 17 September 2015. ©2015 Association for Computational Linguistics.

2.2 Machine learning based methods

There are several studies on automatically identifying ADEs from text of electronic patient records. One Spanish study by Santiso et al. (2014) used 6,100 concepts and 4,700 adverse drug reactions (ADRs) relations for training, and evaluated on 2,100 concepts and 1,600 ADR relations, and obtained 0.93 precision and 0.85 recall using the Random Forest algorithm.

In a Japanese study by Aramaki et al. (2010), using 3,012 Japanese discharge summaries, the authors annotated 1,045 drugs and 3,601 possible adverse drug effects. They found that around 7.7% of the discharge summaries contained ADE. Of these, 59% could be extracted automatically. They used both support vector machine (SVM) and pattern matching methods(PTM) and obtained slightly better results using PTM; precision was 0.41 and recall 0.92 when using PTM, and SVM gave precision of 0.58 and recall of 0.62.

In a study by Roller and Stevenson (2014), UMLS was used to identify concepts and relations in millions of biomedical articles (for instance, drug contraindications, ADE drug relations), and used them to train a Naïve Bayes classifier obtaining 0.25 precision and 1.00 recall.

Gurulingappa et al. (2012) developed a manually annotated corpora in English consisting of 3,000 medical case reports (i.e. published scientific reports of specific patients, their drugs and their side effects). Three annotators annotated the corpora using the concepts *drugs, drug dosage, adverse effect* and *relationship* among the concepts. The three annotators have an M.Sc. degree in biomedicine, where two of them were experienced and one novice. One annotator was used as standard and the other as reference. Each annotator annotated 2,000 documents, and 1,000 documents were annotated by all three annotators. The F-score was measured. *Drugs* obtained an F-score for partial match from 0.90 down to 0.38. *Adverse effect-Drugs* obtained an F-score of 0.79 down to 0.37. The ADE-corpus is publicly available[1]. The authors performed machine learning experiments both with Naïve-Bayes and Maximum Entropy (MaxEnt) classifiers from the MALLET toolkit, and obtained, as best for MaxEnt, 0.75 precision and 0.64 recall.

2.3 Aim and purpose

Previous approaches to detect adverse events have used either rule-based approaches or machine learning approaches, but none have applied a mixed method.

We aimed to design a method that identifies ADRs in health records. The identification of ADRs is realised with a mixture of keyword and phrase filtering and supervised machine learning algorithms that classify health records. By filtering of ADR related phrases, we achieved less computational effort to obtain the prediction and higher prediction performance. We also aspired to design a flexible method that is able to distinguish between different kinds of ADRs - for example, possible ADRs and ADRs related to a certain drug. Finally, we strived for both classification of medical records concerning ADRs and revealing the drug-symptom relations that are decisive for this classification.

3 Materials and methods

3.1 SEPR Corpus and SEPR Drug Corpus

Stockholm Electronic Patient Record (SEPR) Corpus is a patient record collection encompassing over one million patient records from the years 2006-2014 from Karolinska University Hospital in Stockholm, (Dalianis et al., 2012). Of this SEPR Corpus[2], records were sampled to be used for the machine learning experiment.

The SEPR Corpus is stored in a relational database. The unique serial number of each patient was extracted to identify the corresponding health record written by the physicians. Each entry of the record was ordered in temporal order including the drugs taken by each patient. Although, there are no personal names in the data base there can be personal names mentioned in the free text and therefore the data still may contain confidential information. Thus, the data cannot be made publicly available. The problem is known and there are initiatives to establish an infrastructure of publicly available medical records for research (Dalianis et al., 2015).

[1] https://sites.google.com/site/adecorpus/

[2] This research has been approved by the Regional Ethical Review Board in Stockholm, (Etikprövningsnämnden i Stockholm), permission number 2012/834-31/5.

The SEPR Drug Corpus is the sampled part of the SEPR Corpus. The sampling was performed in two steps. First, five drugs were selected for the focus, and second, four groups were defined to classify the records.

3.2 Methods

When developing our method as presented in Figure 1, we faced two problems, that called for solutions. Firstly, health records are unstructured, not standardised texts, written by many authors with different writing styles. Secondly, health records tend to be long. The average number of words in a health record in the SEPR Drug Corpus is 11,228 words (133 words is the minimum and 74,457 words is the maximum). This volume of data cannot be handled by a Machine Learning application installed on a regular computer since machine learning applications in general need high primary memory capacity. For example, without any text manipulation and filtering, a document vector generated from a sub-dataset with Cefuroxim patients (see SEPR Drug Corpus) would contain 391,761 features.

The first mentioned problem is solved by normalisation of writing format and words. The latter problem is solved by several pre-processing steps, focusing on five drugs and their known ADRs, dividing the dataset into sub-datasets, and keywords and phrase filtering. The designed method is evaluated with the known performance measurements precision (*P*), recall (*R*), (Rijsbergen, 1979) and F-score (Powers, 2015).

3.2.1 Sampling

A sample should be composed in such a way that a machine learning algorithm can work effectively and efficiently. Moreover, a sample should mirror the whole corpus, so that the gained insights can be applied to the whole corpus, and even more importantly, can be generalised. With these defined requirements, a sampling of medical records from the SEPR Corpus was performed in a multi-stage sampling approach. First, five drugs were selected. The choice was assigned and confirmed by our research physician. With this selection, the number of ADRs was narrowed down; however, the method has to work on any drug. Furthermore, independent sub-datasets can be formed to ensure that the results can be generalised. The selected drugs belong to different pharmacological/therapeutically subgroups according to the Anatomical Therapeutic Chemical (ATC) classification system (WHO, 2015) to guarantee that the designed method is applicable to different drugs. Furthermore, each of them is given to at least 1,000 patients to guarantee that the designed method is valid; and two of the drugs are given to a larger number of patients (more than 1,000 patients) to guarantee that the designed method can distinguish between different class labels. To read more details about the requirements and the sampling please see (Friedrich, 2015).

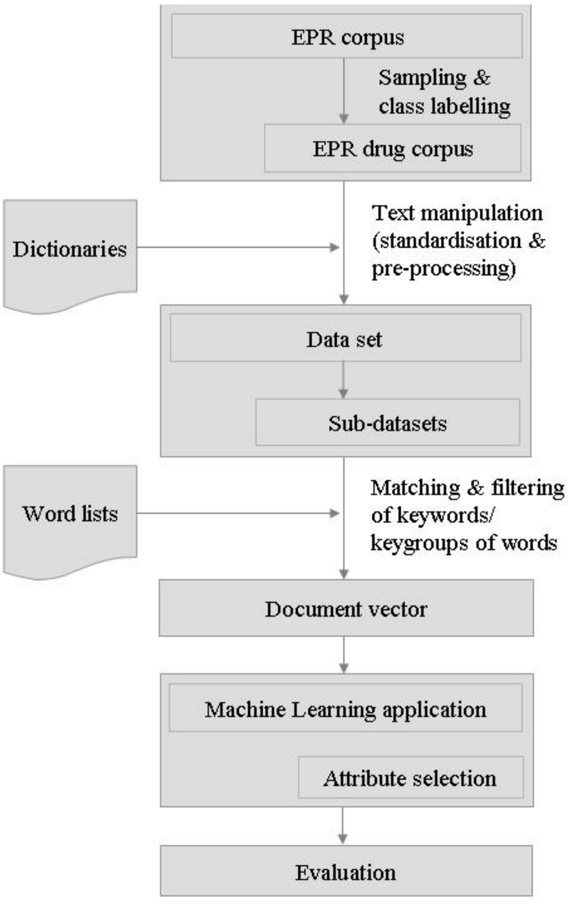
EPR corpus
Sampling & class labelling
EPR drug corpus
Dictionaries
Text manipulation (standardisation & pre-processing)
Data set
Sub-datasets
Word lists
Matching & filtering of keywords/ keygroups of words
Document vector
Machine Learning application
Attribute selection
Evaluation

Figure 1: Method process steps

The following drugs fulfil the above requirements (FASS, 2014a), the ATC codes can be found here [3]:

1. **Cefuroxim** (ATC code J01DC02), antibiotic (Swedish: antibiotika); drug agent Cefuroxime
2. **Imovane** (ATC code N05CF01), psycholeptic (Swedish: neuroleptika, lugnade medel och sömnmedel), also called a tranquilliser or sleeping aid; drug agent Zopiclonum INN

[3] http://www.fass.se/LIF/result?query=&userType=0

3. **Oxynorm** (ATC code N02AA05), analgetica (Swedish: smärtstillande medel), also called a pain killer; drug agent Oxycodone
4. **Prednisolon** (ATC code H02AB06), corticosteroids (Swedish: Kortikosteroider för systemiskt bruk), also called cortisone; drug agent Prednisolon
5. **Primperan** (ATC code A03FA01), functional gastrointestinal disorders (Swedish: funktionella mag-tarmsymptom); to treat nausea and vomiting, drug agent: Metoclopramide

Dosage and dosage form were not considered, nor the companies producing or dealing with any of the above mentioned drugs.

As a second sampling step, four classes were defined, and assigned to health records. Health records that could not be assigned to one of the four classes were excluded; these were mainly health records with contradicting notes about ADRs. The four classes are:

1. **definitely ADRs and related to the chosen drugs**: known and drug-related ADRs according to FASS (2014b)
2. **definitely ADRs but not related to the chosen drug**: ADR because an ADR term (in Swedish e.g. biverkning, läkemedelutlöst) is mentioned in the health record, or diagnosis related to ADR (e.g. IDC G44.4 Drug-induced headache, not elsewhere classified); category A1 according to Stausberg and Hasford (2011)
3. **possible ADRs**: suspicious reaction or suspected adverse drug event, however, neither diagnosis is related to ADR or symptoms are linked to ADR
4. **no ADR**: 'clean' patients with no ADR mentioned, no ADR related diagnosis, or ADR related symptom

3.2.2 Manual class labelling

The 14,751 health records from the SEPR Drug Corpus were manually classified by one annotator who also is a computer scientist. Large parts of the records do not contain any note about ADRs. To aid the annotator the records were pre-annotated; with pre-annotation, only ADR-related passages have to be read. Designed word lists were used, which supported the pre-annotation and the manually performed class labelling. The manual annotation took around seven weeks to perform.

With the performed two sampling steps (five drugs, four classes) the SEPR Drug Corpus was constructed containing 14,751 health records assigned to four class labels, presented in Table 1.

Stausberg and Hasford (2011) categorised 505 ICD-10 codes in seven groups "with respect to its validity as an indicator for an ADE and its definition in the ICD-10". To avoid biased class labelling, patients with a diagnosis related to Stausberg's category A.1 (drug-related causation was noted in the ICD-10) were assigned to class 2 regardless of whether ADRs were mentioned in the health record.

3.3 Text manipulation

To improve both prediction performance and computational speed, normalisation and pre-processing of the unstructured texts were carried out. The number of synonyms, abbreviations, and misspellings that exist for terms decrease the performance of a classifier, thus, normalisation was performed. Pre-processing was carried out to reduce the document vector space. More precisely, the following measurements concerning text manipulation were carried out:

1. Normalisation of text and letter format
   - Case converter, to change capital letters into small letters; that improves keyword search since capital letters do not have to be considered anymore
   - Punctuation marks (e.g. , ; / </n>) are mostly erased or replaced with dots, so that the sentence structure is not ruined. Punctuation marks are not part of a text analysis, but it can be a hint for an unknown ADR or for class label 3. Therefore, the question mark ("?") is changed into the word *maybe* (Swedish: *kanske*)
2. Normalisation of words
   - Expansion to change abbreviations and acronyms to their full word (e.g. *pat* into *patient*)
   - Misspellings and synonyms
   For example, the following abbreviations and variations are found for the term *biverkning* (English: *side effect*): *biv, biverkn, biverkningarna.* For the term *biverkning* 44 misspellings were recognised.
3. Pre-processing
   - Number filter to reduce the vector space and increase computational speed
   - N-Char filter [N=2] to remove words that contain less than N letters
   - Stop Word Filter to filter out stop words (Leskovec et al., 2014)

3.4 Dictionaries and word lists

To support normalisation, three dictionaries were created:

	Total	drug related ADR	general ADR	possible ADR	no ADR
		class 1	class 2	class 3	class 4
Cefuroxim	1,243	11	840	134	258
Imovane	2,329	4	1,790	124	411
Oxynorm	2,886	16	2,120	143	607
Prednisolon	3,411	42	2,134	312	923
Primperan	4,882	34	3,237	296	1,315
Total	14,751	107	10,121	1,009	3,514
Share		0.7%	68.6%	6.8%	23.8%

Table 1: SEPR Drug Corpus with health records per class and drug

- misspellings (contains 460 words)
- abbreviations (contains 217 words)
- synonyms (contains 19 words)

These dictionaries were created by applying a Swedish grammar checker (Semushin, 2015) to a Bag of Word list. Marked words were either abbreviations or misspelling. The list was completed with corrected or expanded words. However, the dictionary for misspellings is not complete; the detection of misspellings was finalised after 460 words since the process was too time consuming.

To support the process of matching and filtering words or phrases, different word lists were created and can be found here [4]:

- **ADR terms** (12 words): they are found with an online thesaurus (SinovumMedia, 2015), online translator (bab.la, 2015), and lexica (Wikipedia, 2015; FASS, 2014b; Healthcare, 2015; Elsevier, 2015).
- **ADR phrases** (195 phrases): for this case a new iterative technique was developed: a term frequency counter for ADR terms (12 words) is applied. Words that co-occur with ADR terms are filtered and combined into one phrase. These three steps (applying a frequency counter, filter terms, and combining them) is executed three times in total. The results are in a list of ADR phrases, which is checked manually.
- **drugs** (5 words): the five chosen drugs
- **drug-related ADR symptoms**: ADR that are known, related to one of the five chosen drugs, and listed in FASS (2014b)
- **general ADR symptoms** (93 words): the list is generated with terms retrieved with the web-content-mining technique from FASS (2014b). The retrieval and extraction process was not part of this research study.

The word list 'ADR terms' contains, for example: *reaktion, bieffekt, biverkning, iatrogen* (in English: *reaction, side effects, adverse events, iatrogenic).* The word list 'ADR phrases' contains, for example: *mycket vanlig biverkning, förmodligen biverkning, känd biverkning, misstanke om biverkning* (in English: *very common side effect, possible side effect, known side effect, suspected adverse reaction)*

Both dictionaries and word lists were reduced to their stem with the Swedish version of Snowball stemmer (Porter, 2001) to improve hit accuracy and to bypass spelling errors.

3.5 Matching and filtering

To reduce the word vector space even more, words and phrases were matched according to the created word lists and filtered from the health records. This match and filter process involves two dimensions: matching words that belong to three different intensities, and filtering the matched terms into three different levels.

An ADR is a drug-symptom relation and is usually indicated in a sentence with a combination of terms for drug, reaction (like *biverkning*), and symptom, such as *Cefuroxim gav biverkning* (English: *Cefuroxim gave ADR*), or *Patient har fått huvudvärk av Cefuroxim* (English: *Patient got headache from Cefuroxim*). A negated or suspected ADR may be indicated with a combination of a negated or suspected reaction, drug and (negated) symptom, for example *inga biverkningar* (English: *no ADR*), *ingen huvudvärk av Cefuroxim* (English: *no headache from Cefuroxim*). Thus, four types of terms were distinguished to match ADRs: reactions, drugs, symptoms, and help words. These terms were matched in the records in three different intensities to meet the variants of an ADR/no ADR/possible ADR that can be expressed and to investigate which matching intensity gains best prediction perfor-

[4] http://dsv.su.se/health/dictionaries

mance. The matching intensities (dimension) are:

1. Intensity 1 - **ADR terms**: terms that indicate an ADR (e.g. *biverkning, bieffekt, biverkan, läkemedelsutlöst*) were tagged as [Reaction]; additionally the names of the five chosen drugs were tagged as [Drug]. The tagged terms are mostly nouns.
2. Intensity 2 - **ADR terms plus help words**: helping words and terms that may indicate an ADR (e.g. *förmodligen, misstänkt*) or no-ADR (e.g. *ingen, utan*) are tagged as [HelpWord] additionally to intensity level 1. The additionally tagged terms are adjectives, adverbs, and verbs.
3. Intensity 3 - **ADR terms, help words, plus symptoms**: in addition to the two levels above, two types of symptoms were tagged: general symptoms that often are ADRs as [Symptoms], and additionally, symptoms that are ADR related to the drug as [Symptoms]. The additionally tagged terms are mostly nouns.

With matching and tagging a text is modified to: "Patient har inga[HelpWord] biverkningar [Reaction] av strålbehandlingen ännu, mår inte [HelpWord] illa [Symptom] och har inga [HelpWord] huvudvärk [Symptom]. Tar Cefuroxim [Drug]." (English: "The patient has no [HelpWord] adverse drug reactions [Reaction] of radiotherapy yet, do not [HelpWord] feel bad [Symptoms] and have no [HelpWord] headaches [Symptoms]. Taking Cefuroxime [Drug].")

Since words create their meaning in a context, a second dimension was chosen to filter the matched words in three different levels of filtering. For example, a term such as *adverse reaction* influences the class label depending on whether the term is part of *patient has no adverse reactions* or of *patient has an adverse reaction to Cefuroxim*. That is why the context of the matched term is considered with the filter level. Eriksson et al. (2013) applied designed dictionaries to filter ADR relevant compounds in clinical texts. However, the number of features should also be reduced without losing important information so that a machine learning algorithm runs in an appropriate time. Therefore, the following filter levels are defined:

1. Filter level A - **only matched words**: only the tagged words are filtered. This means that if the term *biverkning* (English: *ADR*) is tagged, it is filtered, no matter if the term is part of the sentence *ingen biverkning* (English: *no ADR*) or *pratar om biverkning* (English: *talking about ADR*); however, helping words like *ingen* (English: *no*) are also filtered. The disadvantage of only matching and filtering keywords is that the order of words is not considered nor are negations.
2. Filter level B - **matched words and their neighbours**: tagged words and their word neighbours in the sentence are filtered (N-grams on word level). The idea behind this is to find and include words in the classification model that are close by the keywords and may influence their meaning, but were not tagged on purpose. Still, neither the order of words nor negations are considered.
3. Filter level C - **phrases and tag as one attribute**: ADR phrases are filtered, in addition to tagged words depending on the intensity level. The document vector contains of both words and word groups, where a word represents one feature and an ADR phrase represents one feature. Here, negations and word order concerning ADR phrases are considered.

The reasons for this choice are, firstly, the attributes in a document vector usually consist of one word (as in level 1); however, the filtering should not be biased (therefore, level 2 was created). Secondly, a solution for considering negations and word order was needed (therefore, level 3 was defined). With the filtered words and phrases, a document vector is formed, and machine learning algorithms are applied.

3.6 Machine Learning

As mentioned in the Methods section, the dataset containing of five drugs and four classes (14,751 records) must be reduced to run a Machine Learning tool in an appropriate amount of time. Moreover, dividing the dataset into sub-datasets helps to distinguish different combinations of classes, and thus, a comparison for determining which combination yields higher prediction results. It also meets the aim of designing a flexible method that is able to distinguish between different kinds of ADRs. Therefore, three sub-datasets were created that contain:

1. Health records of patients that take either Imovane or Oxynorm and are class-labelled with 2 (general ADRs), 3 (maybe ADRs), or 4 (no ADRs) to distinguish patients that may have ADRs from patients that definitely have ADRs or no ADRs at all.
2. Health records of patients that take Prednisolon and are class-labelled as 1 (ADR related to Prednisolon) or 2 (general ADRs) to distinguish patients with Prednisolon related ADR from patients with ADRs related to other drugs.
3. Health records of patients that take Cefuroxim and are class-labelled with 1 (Cefuroxim related ADRs), 2 (general ADRs), 3 (maybe ADR), or 4 (no ADR).

Other combinations are conceivable and are planned for the future.

Decision tree (DT) (Quinlan, 1993), Random Forest (RF) (Breiman, 2001) and Library Support Vector Machine (LibSVM) (Chih-Chang and Lin, 2001) were chosen as supervised machine learning algorithms. The produced model is applied on a test dataset containing 30% of the instances gained with stratified sampling. The predicted classes of the test data instances are evaluated with precision *(P)*, recall *(R)*, and F-score as performance measurer.

Furthermore, as the machine learning tool KNIME (Berthold et al., 2007) was chosen with WEKA add-ons for decision tree, random forest and libSVM (KNIME, 2015).

4 Results

With this research project a method has been developed that is able to do both predict, if an ADR/possible ADR in a health record occur, and identify an ADR as a drug-symptom relation. The latter was realised with attribute selection. With the presented method there were two results: firstly, prediction performance measurements as presented in Table 2, and secondly, attributes that are selected and may reveal drug-symptom relationships.

4.1 Best results for classification

The best results for a classification on a test dataset that contains of 30% of the health records chosen with stratified sampling were achieved with libSVM (default parameters): precision of 0.69, recall of 0.66 and F-score of 0.67. The prediction was performed on a 3-class sub-dataset containing patients that take Imovane or Oxynorm, and health records that were class-labelled with 'general ADR' (class 2), 'possible ADR' (class 3), or 'no ADR' (class 4). 10-cross-fold validation was also carried but led to lower prediction performance.

4.2 Iterative filter and tag technique

A new iterative technique was introduced to filter ADR phrases, to combine them into one feature, and to build a document vector. This technique improves prediction performance for the best achieved result (3-class problem, intensity 1) from an F-score of 0.46 to an F-score of 0.65 (both with DT) and from an F-score = 0.44 to an F-score of 0.67 (with libSVM). This is an F-score improvement of 41% and 52%, respectively, for the 3-class problem, intensity 1.

4.3 Results for drug-symptom relation extraction/feature selection

With feature selection, the most important features concerning the class labels are selected. A selected attribute is important if it supports distinguishing one class from the other. Here, it means an attribute helps to distinguish patients that, for example, have an ADR from patients that have none. However, a selected term does not mean that it is an ADR.

For the sub-dataset that is reduced to patients taking the drug Prednisolon (2-class problem, intensity level 3, filter level C, F-score of 0.62), a classifier must distinguish patients that show Prednisolon-related ADR from patients with general ADR. With attribute selection, 36 attributes were considered important. Four of them were categorised as ADR phrases, three can be related to Prednisolon, five terms are mentioned in FASS (FASS, 2014b) as either known ADR of Prednisolon or symptoms that are treated with Prednisolon. Eight of the 37 selected attributes are tagged as symptoms, but they are not mentioned as ADRs of Prednisolon nor as a symptom treated with Prednisolon (FASS, 2014b). One of these symptoms is *body weight*, which is a known side effect of cortisone, however, it is not mentioned as side effect of Prednisolon. Sixteen attributes cannot be evaluated clearly.

For the sub-dataset that contains patients taking the drug Cefuroxim (4-class problem, intensity level 1, filter level C, F-score of 0.48), 13 attributes are considered important with attribute selection, and all belong to ADR phrases.

4.4 Term frequencies

The term frequencies of tagged ADR terms were compared in the health records. Synonyms for an ADR are not equally distributed across the classes. For example, in class 2 the word *biverkning* (English: *adverse reaction*) occurs 320 times per 100 health records, whereas in class 3, it occurs only 165 times per 100 health records. In class 3 the ADR term *biverkan* (English: *adverse effect*) is preferred with 139 occurrences compared to class 2 with 66 occurrences. For all three classes, the terms *biverkning* (English: *adverse reaction*), *reaktion* (English: *reaction*), *biverkan* (English: *adverse effect*), and *bieffekt* (English: *side effect*) are the preferred terms to describe an ADR. The term *överkänslig* (English: *hypersensitive*) is mentioned 63 times per 100 health records in classes 2 and 3, whereas in class 4 (no ADR) it is only mentioned 36 times per 100 health records.

			libSVM			DT			RF		
classification problem	filter level	intensity	P	R	F-score	P	R	F-score	P	R	F-score
4-class classification											
Cefuroxim	A	1	0.43	0.40	0.41	0.55	0.35	0.35	0.39	0.29	0.28
class labels 1,2,3,4	A	3							0.18	0.25	0.20
	C	1	0.17	1.00	0.20	0.50	0.55	0.51	0.45	0.33	0.34
	C	2	0.70	0.45	0.48				0.43	0.28	0.27
	C	3	0.40	0.46	0.35						
3-class classification											
Imovane & Oxynorm	A	1	0.50	0.43	0.44	0.61	0.45	0.46	0.25	0.33	0.28
class labels 2,3,4	A	3				0.59	0.48	0.51	0.25	0.33	0.28
	C	1	**0.69**	**0.66**	**0.67**	0.70	0.62	0.65	0.25	0.33	0.28
	C	3				0.66	0.57	0.60	0.25	0.33	0.28
2-class classification											
Prednisolon	A	1	0.60	0.57	0.58	0.49	0.50	0.49	0.49	0.50	0.49
class labels 1,2	A	2	0.74	0.54	0.56	0.99	0.54	0.57	0.49	0.50	0.49
	A	3	0.49	0.50	0.49	0.82	0.58	0.62	0.49	0.50	0.49
	C	1	0.69	0.57	0.61	0.99	0.54	0.57	0.49	0.50	0.49
	C	2	0.74	0.54	0.56	0.49	0.50	0.49	0.49	0.50	0.49
	C	3	0.49	0.50	0.49	0.99	0.54	0.57	0.49	0.50	0.49

Table 2: Performance measurer for a 30% test dataset

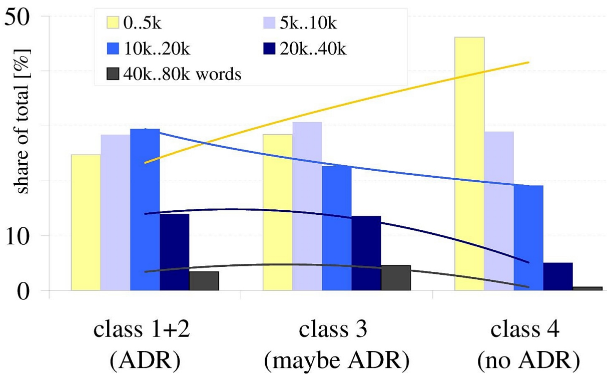

Figure 2: Distribution of record length per class.

4.5 Text length of health records

The length of a health record differs over the classes, as presented in Figure 2. The proportion of short texts shifts if patients are classified as 4 (no ADR), whereas health records classified as 1 or 2 (ADR occurred) tend to be longer.

5 Conclusions

Human beings are complex systems, so there can be great diversity in their reaction to a medical treatment. Clinical tests cannot carry out on all variants, even if all variants are known. Therefore, we have presented a method to perform ADR detection as post-marketing drug safety surveillance.

Supervised machine learning algorithms were applied on the SEPR Drug Corpus with 14,751 class labelled health records. A good prediction performance was yielded (F-score of 0.67, precision of 0.69 and a recall of 0.66). Hazlehurst et al. (2009) identified vaccine adverse effects with a supervised machine learning approach and found 181 of 319 true positives (57%). Aramaki et al. (2010) conducted machine learning and predicted 59% of the adverse drug events correctly. The present method finds 865 of 981 true positives (88.2%) in a 3-class dataset, which is an improvement of 49.5%.

The high ratio of misspelled terms for selected attributes calls for smart spell checks specialised for medical texts. The supervised machine learning algorithms prefer rare terms to distinguish classes. Misspellings disturb and mislead this process immensely.

The reason behind the different text length over the class labels (see figure 2) 1 to 4 may be that getting ADRs is just a question of time. Also, it may be that weakened, more sensitive, or ailing patients are prone to getting sick more often, and thus getting more medical treatment. Cascading effects arise, and therefore, it is more likely that an ADR occurs.

The fact that the word *överkänslig* (English: *hypersensitive*) occurs more often in health records of class 2 and 3 (139 occurrences vs 66 occurrences in class 4 per 100 health records) underscores our observation that patients with certain sensitivities are more likely to develop an ADR. If there is a correlation between ADRs and a patient's sensitivity, then this is even more of a reason to invest in post-marketing drug safety surveillance, since sensitive persons are limited in their participatiion in medical tests. The fact that the number of people that develop allergic reactions and other hypersensitivities has increased in recent years, highlights the urgency of post-marketing drug safety surveillance to better understand drug-symptom relations under special circumstances.

In the future, we plan to apply a spell check for Swedish, NER and parser techniques, to make the pre-processing faster and the prediction performance more accurate. Unfortunately, we used only one annotator; in the future, we will use at least two annotators to calculate the inter-annotation agreement (IAA). We also plan to test different methods of ADR expression extraction to perform machine learning and to obtain improved results.

Acknowledgements

The authors would like to thank Dr. Maria Kvist for her contributions as medical expert.

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

References

Helen Allvin, Elin Carlsson, Hercules Dalianis, Riitta Danielsson-Ojala, Vidas Daudaravicius, Martin Hassel, Dimitrios Kokkinakis, Heljä Lundgrén-Laine, Gunnar H Nilsson, Øystein Nytrø, Salanterä Sanna, Suominen Hanna, and Velupillai Sumithra. 2011. Characteristics of Finnish and Swedish intensive care nursing narratives: a comparative analysis to support the development of clinical language technologies. *Journal of Biomedical Semantics*, 2 Suppl 3:1–11.

Eiji Aramaki, Yasuhide Miura, Masatsugu Tonoike, Tomoko Ohkuma, Hiroshi Masuichi, Kayo Waki, and Kazuhiko Ohe. 2010. Extraction of adverse drug effects from clinical records. *Stud Health Technol Inform*, 160(Pt 1):739–743.

bab.la. 2015. Online dictionary for 27 languages. http://bab.la/. accessed 31/5/2015.

Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Christoph Sieb, Kilian Thiel, and Bernd Wiswedel. 2007. KNIME: The konstanz information miner. In Christine Preisach, Hans Burkhardt, Lars Schmidt-Thieme, and Reinhold Decker, editors, *GfKl*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 319–326. Springer.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Chih-Chang and Chih-Jen Lin. 2001. *LIBSVM: a library for support vector machines.* http://www.csie.ntu.edu.tw/˜cjlin/libsvm.

Hercules Dalianis, Martin Hassel, Aron Henriksson, and Maria Skeppstedt. 2012. Stockholm EPR Corpus: A Clinical Database Used to Improve Health Care. In *Swedish Language Technology Conference, SLTC-2012*, pages 17–18.

Hercules Dalianis, Aron Henriksson, Maria Kvist, Sumithra Velupillai, and Rebecka Weegar. 2015. HEALTH BANK - A workbench for data science applications in healthcare. In John Krogstie, Gustaf Juel-Skielse, and Vandana Kabilan, editors, *CAiSE Industry Track*, volume 1381 of *CEUR Workshop Proceedings*, pages 1–18. CEUR-WS.org.

Elsevier. 2015. Das Roche Lexikon Medizin. https://www.tk.de/rochelexikon/. accessed 31/5/2015.

Robert Eriksson, Peter Bjødstrup Jensen, Sune Frankild, Lars Juhl Jensen, and Søren Brunak. 2013. Dictionary construction and identification of possible adverse drug events in Danish clinical narrative text. *Journal of the American Medical Informatics Association*, 20(5):947–953.

FASS. 2014a. Fass atc register, http://www.fass.se/lif/atcregister?usertype=0. http://www.fass.se/LIF/atcregister?userType=0. accessed 31/5/2015.

FASS. 2014b. Läkemedelsindustriföreningens Service AB (LIF). Retrieved from 31/5/2015http://www.fass.se. accessed 31/5/2015.

U.S. Food Drug Administration FDA. 2014. *Inside Clinical Trials: Testing Medical Products in People.* http://www.fda.gov/Drugs/ResourcesForYou/Consumers/ucm143531.htm, accessed 31/5/2015.

Stefanie Friedrich. 2015. Detecting clinical entities using machine learning - How to find and predict (patients with) adverse drug reactions in medical records. Master thesis, Stockholm University.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of biomedical informatics*, 45(5):885–892.

Brian Hazlehurst, Allison Naleway, and John Mullooly. 2009. Detecting possible vaccine adverse events in clinical notes of the electronic medical record. *Vaccine*, 27(14):2077–2083.

Alliance Healthcare. 2015. Gesundheit.de. http://www.gesundheit.de/. accessed 31/5/2015.

Rachel L Howard, Anthony J Avery, Serena Slavenburg, Simon Royal, G Pipe, Peter Lucassen, and Munir Pirmohamed. 2007. Which drugs cause preventable admissions to hospital? a systematic review. *British journal of clinical pharmacology*, 63(2):136–147.

KNIME. 2015. Knime weka data mining integration. https://www.knime.org/update/. accessed 31/5/2015.

Jurij Leskovec, Anand Rajaraman, and Jeffrey D. Ullman. 2014. *Mining of massive datasets*. Cambridge University Press, pub-CAMBRIDGE:adr, second edition.

Martin F Porter. 2001. Snowball: A language for stemming algorithms. http://snowball.tartarus.org.

David M. W. Powers. 2015. What the F-measure doesn't measure: Features, flaws, fallacies and fixes. *CoRR*, abs/1503.06410. http://arxiv.org/abs/1503.06410, accessed 31/5/2015.

J. Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

Cornelis Joost Van Rijsbergen. 1979. *Information Retrieval*. Butterworth-Heinemann.

Roland Roller and Mark Stevenson. 2014. Self-supervised relation extraction using umls. In *Information Access Evaluation. Multilinguality, Multimodality, and Interaction*, pages 116–127. Springer.

Sara Santiso, Alicia Pérez, Koldo Gojenola, IXA Taldea, Arantza Casillas, and Maite Oronoz. 2014. Adverse Drug Event prediction combining shallow analysis and machine learning. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)@ EACL*, pages 85–89.

Sergey Semushin. 2015. Notepad++ Spell-checking Plug-In, Swedish. https://github.com/Predelnik/DSpellCheck.git. accessed 31/5/2015.

SinovumMedia. 2015. Synonymer.se. http://www.synonymer.se/. accessed 31/5/2015.

Jürgen Stausberg and Joerg Hasford. 2011. Drug-related admissions and hospital-acquired adverse drug events in Germany: a longitudinal analysis from 2003 to 2007 of ICD-10-coded routine data. *BMC health services research*, 11(1):134.

Xiaoyan Wang, George Hripcsak, Marianthi Markatou, and Carol Friedman. 2009. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *Journal of the American Medical Informatics Association*, 16(3):328–337.

Karin Wester, Anna K Jönsson, Olav Spigset, Henrik Druid, and Staffan Hägg. 2008. Incidence of fatal adverse drug reactions: a population based study. *British journal of clinical pharmacology*, 65(4):573–579.

WHO. 2015. Who collaborating centre for drug statistics methodology. http://www.whocc.no/atc/structure_and_principles/. accessed 31/5/2015.

Nandalal Wijesekera. 2013. *Documenting medical records - A handbook for doctors*. Health Information Systems Knowledge Hub School of Population Health,The University of Queensland, 1 edition.

Wikipedia. 2015. Adverse effect - Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Adverse_effect. accessed 31/5/2015.