Identification of Risk Factors in Clinical Texts through Association Rules

Svetla Boytcheva[1] **Ivelina Nikolova**[1] **Galia Angelova**[1] **Zhivko Angelov**[2]
[1] Institute of Information and Communication Technologies,
Bulgarian Academy of Sciences
[2] Adiss Lab Ltd., Sofia, Bulgaria
svetla.boytcheva@gmail.com {iva,galia}@lml.bas.bg,
angelov@adiss-bg.com

Abstract

We describe a method which extracts Association Rules from texts in order to recognise verbalisations of risk factors. Usually some basic vocabulary about risk factors is known but medical conditions are expressed in clinical narratives with much higher variety. We propose an approach for data-driven learning of specialised medical vocabulary which, once collected, enables early alerting of potentially affected patients. The method is illustrated by experimens with clinical records of patients with Chronic Obstructive Pulmonary Disease (COPD) and comorbidity of CORD, Diabetes Melitus and Schizophrenia. Our input data come from the Bulgarian Diabetic Register, which is built using a pseudonymised collection of outpatient records for about 500,000 diabetic patients. The generated Association Rules for CORD are analysed in the context of demographic, gender, and age information. Valuable anounts of meaningful words, signalling risk factors, are discovered with high precision and confidence.

1 Introduction

Chronic diseases like Chronic Obstructive Pulmonary Disease (COPD) and Diabetes Mellitus are long-lasting disorders with effects that come with time. They are the result of a combination of genetic, physiological, environmental and behaviours factors, and kill over 40 million people each year, equivalent to 70% of all deaths globally[1]. Prevention is focused on reducing the risk factors associated with these diseases. Therefore, establishing the risk rates and early recognition of potential danger will help to decrease the role of the common modifiable risk factors. In the age of big data and given the growing amount of patient-related texts, we believe that Data Mining and Text Mining are key technologies which might help by providing discovery of hidden interdependencies among words (lexical expressions of indicators and assessment of risks) in patient records.

In this paper we demonstrate how automatic analysis of clinical narratives in Bulgarian language allows to identify verbal expressions of risks for patients. Our input data come from the Bulgarian Diabetic Register, which is built using a pseudonymised collection of outpatient records for about 500,000 diabetic patients treated in the period 2010-2016 (Tcharaktchiev et al., 2015). Together with the structured information, the outpatient records contain free texts discussing the patient case history, status, risk factors, treatment etc. Our tools process both structured data and free text of outpatient records in order to extract Association Rules for COPD risk factors. Since Diabetes Melitus and Schizophrenia are also closely related, we study their comorbidity and the risk factors for COPD in patients with Diabetes Melitus and Schizophrenia. By applying unsupervised Data Mining techniques we try to overcome the lack of linguistic and ontological resources that can support successful NLP analysis of clinical narratives in Bulgarian. Thus we demonstrate how new lexical resources can be generated, to be used for better analysis of clinical texts.

The paper is structured as follows. Section 2 overviews related work with focus on the technological solutions. Section 3 presents the method we use, section 4 – the experiments and results. Section 5 contains the conclusion and discusses future work.

[1] World Health Organisation (WHO) factsheets: http://www.who.int/mediacentre/factsheets/fs355/en/

Proceedings of the Biomedical NLP Workshop associated with RANLP 2017, pages 64–72,
Varna, Bulgaria, 8 September 2017.
https://doi.org/10.26615/978-954-452-044-1_009

2 Related Work

Many advanced approaches apply Natural Language Processing (NLP) as a first step in mining entities from free texts and use the latter as input to subsequent biomedical research or decision making tasks. Incorporating NLP has advantages: it systematically links several terms to a concept using databases that standardise health terminologies; avoids manual work for searching term variations; increases the number of patients in the considered cohorts and thus increases the sensitivity of the recognition (Liao et al., 2015). A recent review lists 71 clinical NLP systems, which process free text and generate structured output, in order to address a wide variety of important clinical and research tasks (Kreimeyer et al., 2017). Significant progress has been made in algorithm development and resource construction since 2000 (Luo et al., 2017). Open challenges remain e.g. extraction of temporal information, normalisation of concepts to standard terminologies, interpretation etc. Despite the limitations the conclusion is that today NLP engines are powerful components ready for integration in medical text processing and – due to expected improvements in the near future, e.g. more accurate mappings of terms to medical concepts – the importance of NLP as a valuable supporting technology will grow (Liao et al., 2015). Here we briefly discuss major text analysis technologies that are applied in biomedical domain.

Data mining (DM) is actively used in the field since the middle of 1990's. It employs explorative algorithms to identify meaningful data patterns with acceptable computational efficiency and uncover new biomedical and healthcare knowledge for clinical and administrative decision making. Furthermore it can generate testable evidence-based medical hypotheses from large experimental data, clinical databases, and/or biomedical literature. Today DM is applied for a variety of tasks operating on biomedical entities extracted from free texts. For instance (Luo et al., 2017) states that NLP is a useful tool for extracting information related to adverse drug events (ADE) and pharmaceutical products from electronic health record (EHR) narratives. Since 2012, DM enables successful automation of the ADE discovery so the "NLP-based ADE detection" (as the authors call it) can be soon integrated in practical systems. Moreover, the DM capacity for treatment of heterogeneous data sources is increasingly adopted. (Stubbs et al., 2015) present an overview of the 2014 i2b2/UTHealth NLP shared task focused on identifying medical risk factors related to Coronary Artery Disease (CAD) in the narratives of longitudinal medical records of diabetic patients. Twenty teams participated in this track, and submitted 49 system runs for evaluation. The most successful system used a combination of external lexicons, hand-written rules and Support Vector Machines (a machine learning method). Other machine learning techniques n use were Conditional Random Fields and ensembles of classifiers (CRF, Naïve Bayes, and Maximum Entropy). With six of the top 10 teams achieving F1 scores over 0.90, and all 10 achieving F1 scores over 0.87, the authors conclude that identification of risk factors and their progression over time is within the reach of present automated systems. These examples show that today DM is a key technology for the successful NLP-based medical applications.

Text mining (TM) aims at the delivering of meaningful information from texts, e.g. structuring text units into entities and relationships among them, via NLP applications for shallow analysis. A widely used system of this type is the open-source NLP tool for information extraction from EHR cTAKES (clinical Text Analysis and Knowledge Extraction System)[2]. Another open source system is HITEx (Health Information Text Extraction) which extracts some variables of interest from narrative text (Goryachev et al., 2006). We mention here two more examples how text mining delivers useful information about risk factors and adverse drug events. In (Jonnagaddala et al., 2015) the authors present a system that discovers in free text EHRs information about age, gender, total cholesterol (or low-density lipoproteins cholesterol LDL-C), high-density lipoproteins cholesterol (HDL-C), blood pressure, diabetes history and smoking history for a cohort of 164 diabetic patients. After that the Framingham risk score is calculated to predict the coronary artery disease (CAD) for these patients. The performance of the text extraction system is reliable, however missing data remain a challenging issue. Over 40% of patients in the final cohort are at high risk of CAD and over 50% of the population fitted in the moderate category. The main limitation was the lack of a systematic evaluation of the developed text mining system. In (Harpaz

[2]Official site http://ctakes.apache.org/

et al., 2014) the authors state that TM is sufficiently mature to be applied for the extraction of useful information concerning ADEs from multiple textual sources. Currently such information is collected by manual expert analysis of clinical trial notes and spontaneous reports, and the review of biomedical literature; but progress depends on a comprehensive approach that examines a diverse set of potentially complementing data sources including EHRs. Posting in social media are another source of information about ADEs: 2% of patients and 6% of caregivers share their experiences online, and 18% of all internet users, 31% of all patients with chronic conditions, and 38% of caregivers look at online drug reviews[3]. Despite the challenges, a large body of research has demonstrated that the existing TM tools are capable to extract useful safety-related information from the aforementioned textual sources.

NER and **rule-based approaches** evolved during the last decades from research prototypes to reliable NLP technologies. Mature (and constantly evolving) systems appeared for processing English clinical texts, e.g. KnowledgeMap Concept Identifier which processes clinical notes and returns CUIs (Concept Unique Identifiers) for the recognized UMLS terms (Denny et al., 2003) as well as NegEx, a tool for identification and interpretation of negation in English texts (Chapman et al., 2001), (Gindl, 2006). Identification of temporal events is a hot topic in biomedical NLP. In (Chang et al., 2015) it is proposed to recognise first all temporal expressions and then, after building a temporal model of the context, to assign the corresponding time attributes for all recognised concepts with respect to the creation time of the records. Disease mentions are identified after that, along with their corresponding risk factors and medications. (Chang et al., 2015) shows the progress in processing named entities which represent temporal information. Recently, with the DM development, classical rule-based systems like NegEx can be outperformed by statistical methods (Uzuner et al., 2009); on the other hand the rule-based methods prove to be good in the production of annotated resources and when writing rules that emulate the knowledge of a domain expert (e.g. in ADE discovery).

[3]Pew Research Center, The Social Life of Health Information, 2011: http://www.pewinternet.org/2011/05/12/the-social-life-of-health-information-2011/

3 Methods

Our approach (Fig. 1) has five main phases: *(i)* Structured information processing of the ORs in the repository; *(ii)* Risk Factors Association Rules generation from the training set; *(iii)* Preprocessing of the test sets; *(iv)* Risk Factors Association Rules matching on the test sets; *(v)* Structured information processing of the patients in risk.

3.1 Structured Data Analysis Methods

The Diabetes Register contains pseudoanonymous Outpatient Records (OR) in XML format. Most data necessary for the health management are structured in fields with XML tags which present the Patient ID, the code of doctors' medical specialty, region of practice, Date/Time and ID of the OR. Several free-text fields contain important explanations about the patient: "Anamnesis", "Status", "Clinical examinations" and "Therapy". There are also several XML tags for the main diagnose and additional diagnoses with their codes according to the International Classification of Diseases, 10th Revision (ICD-10)[4]. Each OR contains a main diagnosis with ICD-10 code and ICD-10 codes of up to 4 additional disorders, i.e. in total from 1 to 5 ICD-10 codes.

The study of disorder comorbidities plays an important role in detection and prevention of patients at risk. Chronic diseases constitute a major cause of mortality according to the World Health Organization (WHO) reports and their study is of higher importance for healthcare. For discovering frequent patterns of chronic diseases we use retrospective analysis of population data, by filtering events with common properties and similar significance. One of the major approaches to pattern search is frequent pattern mining (FPM) viewing the events (objects) as unordered sets. This preliminary work was done over outpatient records (ORs) of patients with primary diagnose Diabetes Melitus Type 2 (ICD-10 code E11) *(withdrawn Self-reference)*. We extracted relatively high number of frequent patterns containing different mental disorders – ICD-10 codes F00-F99. This result motivated us to process collection for patients with Schizophrenia (ICD-10 code F20). The study collection *SD* of patients who suffer from both Schizophrenia and Diabetes Melitus Type 2 was

[4] International Classification of Diseases and Related Health Problems 10th Revision. http://apps.who.int/classifications/icd10/browse/2015/en

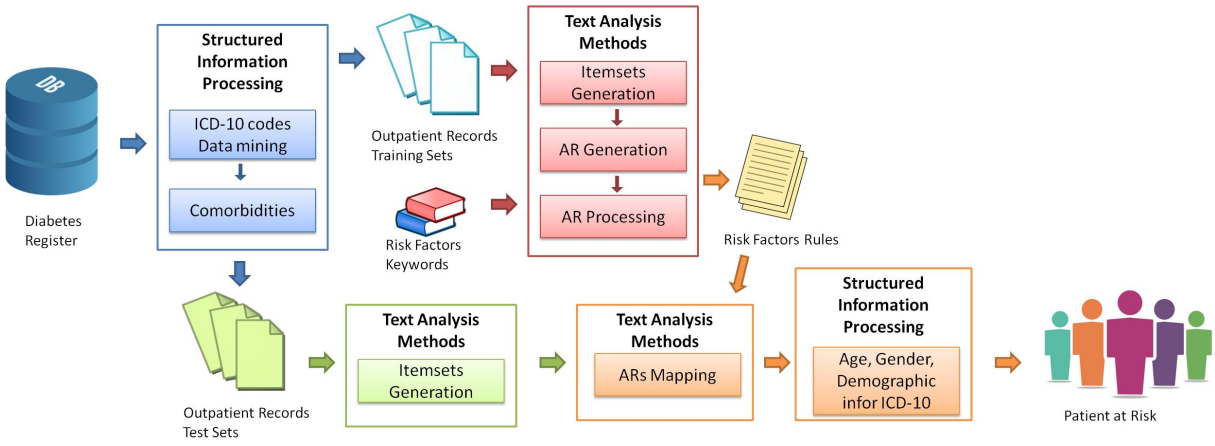

Figure 1: Identification of patients at risk

automatically extracted from the Diabetes Register and contains all ORs for these patient in the period 2012-2014 - approx. 200,000 ORs for 4,080 patients.

Let H be a chronic disease and there exist a frequent itemset F of chronic diseases such that $\{H, E11, F20\} \subseteq F$. The study collection *SD* is split into two subsets *SH* and *ST*. The collection *SH* contains ORs of all patients in SD that also have diagnosis *H*. We will call the set *SH* a training set. The set *ST* is formed as $ST = SD - SH$. We will call the set *ST* a test set.

3.2 Risk Factors Association Rules Generation

Text Analysis has three main phases: *Itemsets Generation* which converts the text documents into itemsets, *Association Rules Generation* based on frequent pattern mining (FPM) techniques and elicitation of ARs, and *Risk Factors Association Rules Filtering* that filters rules by using keywords (Fig. 2).

The system processes input texts in unicode format and is language independent in principle (stemming and stopword filtering can be replaced with modules for another language).

3.2.1 Itemsets Generation

Let *SH* be the training set. We extract for each OR its parts in XML tags for Anamnesis (Patient History) and Status and form separate collections of ORs Anamnesis texts only - *SHa*, and ORs Status texts only - *SHh* correspondingly. We process separately the collections *SHa* and *SHh*.

Let *S* be one collection. Each text in *S* is turned to a sequence of word stems in their original order, using blank spaces and punctuation delimiters as tokenization separators. Stop words and numbers may be essential for some patterns so they are preserved and generalised - replaced by the constants STOP and NUM correspondingly. After this step the punctuation is eliminated. Then we use hashing and substitute each word with an unique number. In addition some compression and sorting is applied. This is necessary to speed up the frequent patterns mining process.

The vocabulary used in all documents of *S* will be called *items* $W = \{w_1, w_2, ..., w_n\}$. For the collection S we extract the set of all different documents $P = \{p_1, p_2, ..., p_N\}$, where $p_i \subseteq W$. This set corresponds to transactions; the associated unique transaction identifiers (*tids*) shall be called ***pids*** (patient identifiers). Each patient interaction with a doctor is viewed as a single document in *P*.

3.2.2 Association Rules Generation

The ORs are written in telegraphic style with phrases rather than full sentences. Usually the ORs list attribute-value (*A*-*V*) pairs - anatomical organ/system and its status/condition. Attribute names contain phrases and abbreviations in Cyrillic and Latin. Values can be long descriptions in case of status complications. The order of *A*-*V* pairs can vary and parts of the value descriptions can surround the attributes. It is also possible that some attributes share the same value. Sample configurations are shown below.

$$A_1V_1, ..., A_nV_n|V_1A_1, ..., V_nA_n$$

$$V_1...V_kAV_{k+1}...V_n$$

$$A_1, A_2, ..., A_nV|VA_1, A_2, ..., A_n.$$

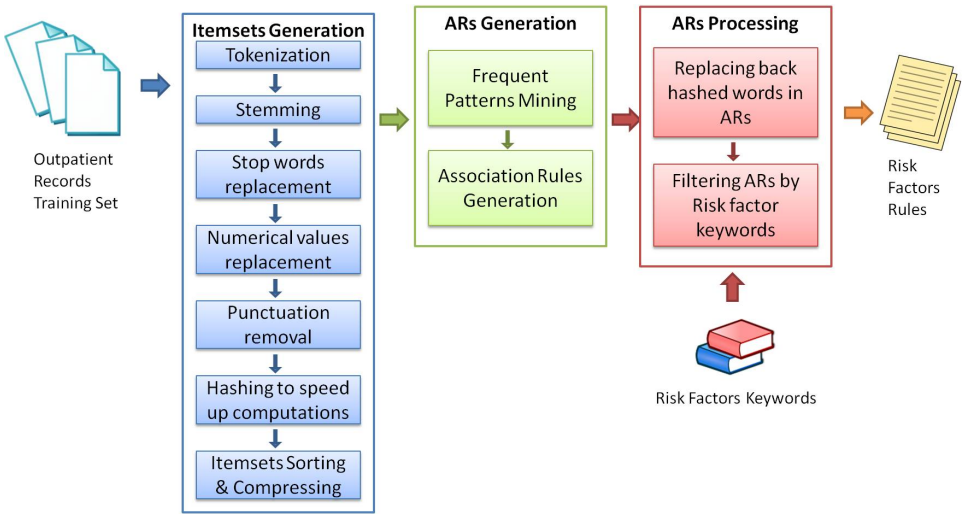

Figure 2: Risk Factors Association Rules Generation

Thus, when searching for frequent patterns, we consider a window of more than 10-12 words around each attribute. The rich terminology and flexible syntax structure hinder the application of traditional methods for extraction of collocations with gaps. Usual collocation extraction approaches would rather find the OR clishe phrases as collocations with highest frequency, moreover many *A-V* pairs would be erroneously considered as *n*-grams. Some FPs are given below.

E.g.: **Positive examples**:
общо състояние (*general condition*)
щитовидна жлеза (*thyroid gland*)
Negative examples:
удължен експириум (*prolonged expiratory time*)
има кашлица (*has a cough*)

Therefore we treat documents as bag of words rather than sequences, they are transformed to itemsets with single word occurrences only.

Given a set of pids S, support of an itemset I is the number of pids in S that contain I. We denote it as $supp(I)$. We define a threshold called *minsup* (minimum support). Frequent itemset (FI) I is one with at least minimum support count, i.e. $supp(I) \geq minsup$. The task of FPM of S is to find all possible frequent itemsets in S.

Most FPM algorithms generate all possible frequent patterns (FPs). The search space grows exponentially with the size of W. Summarised information for data relations can be extracted as maximal frequent itemsets (MFI). The condensed information not only accelerates the process, reducing redundancy, but also decreases significantly the number of frequent patterns for post-analysis.

An implication in the form $I \Rightarrow J$ is called *association rule*, where $I \subset W, J \subset W, I \cap J = \emptyset$. I is called antecedent and J is called consequent. Support of a rule is the number of pids in S that contain $I \cup J$, i.e.

$$sup(I \Rightarrow J) = sup(I \cup J) = P(I \cup J).$$

If *C%* of patient documents in *S* that contain *I*, contain also *J*, then the association rule $I \Rightarrow J$ holds with *confidence C* in S, i.e. this is the condition probability

$$conf(I \Rightarrow J) = P(J|I) = \frac{sup(I \cup J)}{sup(I)}.$$

The task of ARs mining in collection S is to generate all ARs with confidence above the user defined confidence (*minconf*) and support above user defined support (*minsup*). Rules that satisfy both a *minsup* and *minconf* are called strong. However, even for reasonable values of *minsup* and *minconf*, big datasets yield huge amounts of strong ARs. So we use an additional filter called *lift* that is defined as the ratio of the confidence of the rule and the confidence of its consequent.

$$lift(I \Rightarrow J) = \frac{P(I \cup J)}{P(I)P(J)}.$$

The lift represents the strenght of the relation between the consequent and its antecedent. Lift value < 1 indicates independence between them. Lift value > 1 means that the antecedent and consequent appear together more often than expected, i.e. are correlated. Such rules are potentially usefull for predicting the consequent in new sets.

For ARs generation we use algorithms for mining all association rules with the lift measure in a transaction database (Agrawal and Srikant, 1994)

with implementation at SPMF[5]. For experiments is used algorithm for All Association Rule with FPGrowth with lift (Han et al., 2004). Let the two sets of generated ARs for *SHa* and *SHh* correspondingly be *ARa* and *ARh*.

3.2.3 Risk Factors Association Rules Filtering

In order to identify ARs for risk factors we use small lexicon with some keywords - $K = \{k_1, ..., k_m\}$. We convert back the hashed items from the ARs into words and obtain set *ARW*. For the two sets of ARs - *ARa* and *ARh* we have *ARaW* and *ARhW*. Thus the results ARs contain words. We filter those ARs that contain some of the keywords from the lexicon by projection.

$$ARaW_k = \{I \Rightarrow J | I \Rightarrow J \in ARaW \wedge \exists k \in K, k \in I \vee k \in J\}$$

$$ARhW_k = \{I \Rightarrow J | I \Rightarrow J \in ARhW \wedge \exists k \in K, k \in I \vee k \in J\}$$

3.3 Preprocessing of the test sets

Let *ST* be the test set of ORs. All Anamnesis (Patient History) sections formed the text collection *STa*, and all ORs Status texts - the collection *STh*. We process *STa* and *STh* separately. Similarly to the processing of the training set SH, we apply for *STa* and *STh* the first text analysis step - Itemsets Generation - but exclude the last procedures for hashing, compression and sorting.

3.4 Risk Factors Association Rules matching on the test sets

We match the corresponding type ARs to the test collections, i.e. ARs generated from the Anamnesis texts are mapped onto test collections that contain pids for Anamnesis, and the ARs generated from the Status parts of the ORs are mapped onto test collections that contain pids for Status. The result sets contain pids of patients at potential risk of chronic disease *H*.

$$RHa_k = \{p | p \in STa, I \Rightarrow J \in ARaW_k, I \subseteq p \wedge J \subseteq p\}$$

$$RHh_k = \{p | p \in STh, I \Rightarrow J \in ARhW_k, I \subseteq p \wedge J \subseteq p\}$$

[5]http://www.philippe-fournier-viger.com/spmf/index.php?link=algorithms.php

3.5 Structured information processing for patients at risk

Presence of some symptoms is a necessary but not sufficient condition for risk of chronic disease *H*. Some additional factors need further investigation, like related diagnosis with similar symptoms. We also need to study the other current diagnosis of the patient, to take into account age, gender, demographic information, etc. That's why we collect for each patient all pids from RHa_k and RHh_k and the associated structured information with the corresponding ORs from the test *ST*.

4 Experiments and Results

The chronic disease *H* that we investigate here is COPD (ICD-10 code J44), i.e. *H=J44*. The average prevalence of COPD in Bulgaria is 3.197% for 2014 among all Bulgarian citizens (Fig. 3). The average prevalence of both Schizophrenia (ICD-10 code F20) and Diabetes Melitus Type 2 (ICD-10 code E11) in Bulgaria is 0.688% for 2014 among all Bulgarian citizens (Fig. 4). However for 2014 the average prevalence of COPD among patients that suffer by both Schizophrenia and Diabetes Melitus Type 2 is relatively higher 5.576% than the average for the country (Fig. 5).

Some of the typical characteristics of COPD are: starting at middle age; symptoms develop slowly; prolonged smoking is a main reason; patients experience dyspnoea during physical efforts and significant irreversible airflow limitation. Thus in primary interest are ORs written by specialists: in Otolaryngology *(S14)*, Pulmology *(S19)* and Endocrinology *(S05)*. But we try to identify patients at risk, and probably some of them had no visits and consultations yet to such specialists. So we consider also collection of ORs for visits to general practitioners (GP) *(S00)*.

We have 4 text collections with ORs (Table 2): GP (*S00*), Endocrinology(*S05*), Otolaryngology (*S14*), and Pulmology (*S19*). We split these collections into training and test sets, depending on whether they are ORs for patients with *H=J44* or not. In addition we split them into two "*Anamnesis*" and "*Status*" sections of the ORs. Both sections are available for each patients so the training sets *SHa* and *SHh* contain the same number of pids. This is valid also for the test set *STa* and *STh* for each collection.

We can observe that for *SHh* (Table 3) the number of generated FPI and ARs is significantly

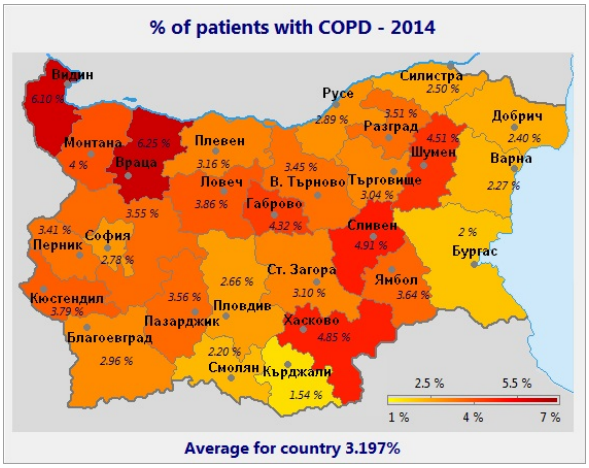

Figure 3: Prevalence of COPD (J44) in Bulgaria, 2014

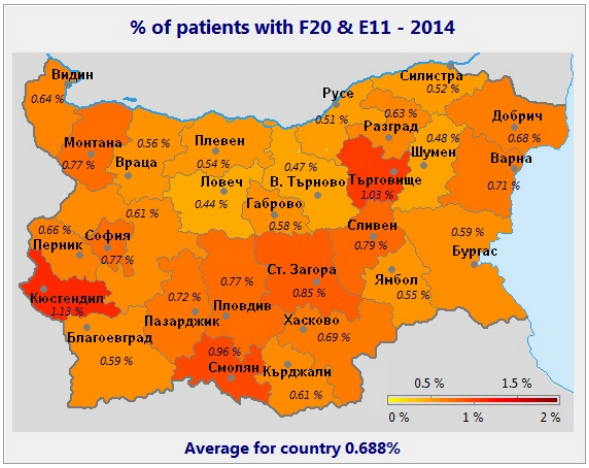

Figure 4: Prevalence of Schizophrenia (F20) and Diabetes Melitus Type 2 (E11) in Bulgaria, 2014

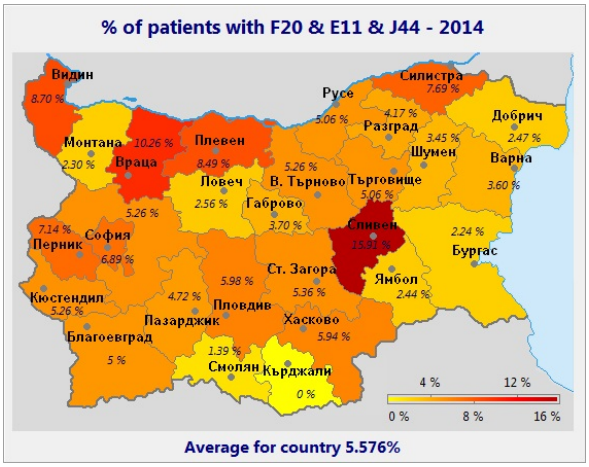

Figure 5: Prevalence of COPD (J44) among patients with both Schizophrenia (F20) and Diabetes Melitus Type 2 (E11) in Bulgaria, 2014

Year	2012	2013	2014	Total
patients	2,929	3,093	3,217	4,080
S00	45,402	46,238	51,894	143,534
S05	2,854	2,900	3,071	8,825
S14	368	351	396	1,115
S19	252	267	344	863

Table 1: Collection *SD* for patients with both Schizophrenia and Diabetes Melitus Type 2

Year	2012	2013	2014	Total
patients	144	166	179	293
S00	3,783	3,796	4,208	11,787
S05	253	273	262	788
S14	45	47	64	156
S19	158	172	202	532

Table 2: Training sets *SHa* and SHh

Set	ARa	FPI	minsup	$ARWa_K$
S00a	647	1,713	0.01	10
S05a	1,695,130	23,677	0.03	0
S14a	82,802	2,499	0.03	34
S19a	278,379	5,431	0.03	249,221

Table 3: Generated Association Rules for Anamnesis with minconf = 1.0 and minlift = 1.05

Set	ARh	FPI	minsup	$ARWh_K$
S00h	1,888,641	286,357	0.08	2
S05h	1,779,462	101,320	0.07	1,264
S14h	1,818	649	0.04	0
S19h	113,718	26,341	0.04	98,185

Table 4: Generated Association Rules for Status with minconf = 1.0 and minlift = 1.1

higher than for *SHa* (Table 4) even for higher $minsup$ values, because the text in Status section is more coherent and contrain less variety of syntax structures. However the projection of these ARs to the keywords set *K* shrinks all the ARs sets in some cases to the ground. And it is not surprice that the majority of the filtered ARs comes from S19a and S19h - ORs from Pulmology.

The keywords for symphtoms of *J44* are:

K = {тежест, задух, кашлица, хрипове, храчки, умора, уморяемост физическа, сърцебиене, трудно, експекторация, експириум} (*Weight, Breathlessness, Cough, Wheezing, Sputum, Fatigue, Tiredness, Physical, Palpitations, Difficult, Expectoration, Expiratory*).

Some generated ARs for COPD risk factors are:

умора експекторац => кашлиц SUP: 17 LIFT: 1.60 (Fatigue Expectoration => Cough)
храчки лесна умора => кашлиц SUP: 17 LIFT: 1.60 (Sputum Easy Fatigue => Cough)
храчки експекторац => задух SUP: 20 LIFT: 2.30 (Sputum Expectoration => Breathlessness)

Patients with potential risk of COPD are identified after matching the filtered rules of AWa_K and AWh_K to *STa* and *STh* correspondingly. The total number of ARs matches over the test sets of ORs is shown on (Table 5) and (Table 6) respectively.

$ARWa_k$	$ARW00a_k$	$ARW14a_k$	$ARW19a_k$
S00Ta	144	1,069	1,154
S05Ta	4	52	96
S14Ta	0	420	0
S19Ta	0	86	464
ORs	20	601	1,018

Table 5: COPD risk factors found in Anamnesis

$ARWa_k$	$ARW00h_k$	$ARW05h_k$	$ARW19h_k$
S00Th	0	3,086,665	829,995
S05Th	0	347,490	125,479
S14Th	0	0	0
S19Th	0	7,806	425,769
ORs	0	33,545	73,485

Table 6: COPD risk factors found in Status

In the following OR excerpt, items from the AR antecedent are highlighted in light blue color and the predicted consequent items are highlighted in pink color.

Association Rule: SUP: 7 LIFT: 9.176
оплаква дишан => затрудн

STOP оплаква STOP честа дразнещ суха кашлиц STOP белезникав храчки задух затрудн дишан заморяван отпадналост STOP главоболи
(*STOP complain STOP frequent irritating dry cough STOP whitish sputum dyspnoea difficult breath* tiredness fainting STOP headache)

Patients that needs to be alerted for COPD risk factors are selected after analyses of some structured information in the ORs: age, gender, demographic region, etc.

COPD develops slowly and usually patient with age above 40s are at a higher risk . Risks are gender specific as well due to the prevalence of male (6.17%) vs. female (5.25%) patients. Demographic information helps to identify patient who live in regions with pollution, close to thermal power stations, etc. On Fig. 5 we can see that such regions in Bulgaria are around the town of Sliven (15.91%), Vidin (8.70%) and Vratsa (10.26%) in comparison with the average prevalence of COPD in the collection 5.576%. Another risk factor that needs further analysis is the patient smoking status because smoking is one of the major causes for COPD development. Some diagnoses related to the CORD symptoms are the following (with the corresponding ICD-10 codes in the parenthesis): *Asthma(J45), Status asthmaticus (J46), Congestive heart failure (I50.0), Bronchiectasis (J47), Tuberculosis (A15-A19), Bronchitis (J40-J42), Acute bronchiolitis (J20-J22), Emphysema (J43)*. So when planning alerts for patients at risk, one should check whether he/she has some of the diagnosis listed above and exclude those patient from the set *RH* for patients with risk alert.

5 Conclusion and Further Work

Here we show how to construct in a reliable manner a "could" of words signalling risks. This is important for a language like Bulgarian where no electornic linguistics resources of medical terminology are available. The existing very large archive of pseudonymised ORs, a nation-wide collection for 2010-2016, enables unique opportunities to acquire automatically lexical resources organised around names of diseases, medical conditions and/or specific groups of patients. The careful pre-selection of training corpora facilitates the explication of association rules; in this experiment we are aware about the comorbidity of CORD and Schizophrenia therefore we extract ORs for a cohort of patients which contains more CORD cases.

Despite the over-generation of ARs, the top rules are a reliable source of information which is easy to filter.

Another important achievement is the sketch of a clear procedure for discovery of patients at risk and issuing alerts to the healthcare authorities who need to take care about their implementation.

Future work involves processing of more complex linguistic constructions (negation) and considering typical risk factors (smoking).

Acknowledgments

This research is supported by the grant SpecialIZed Data MIning MethoDs Based on Semantic Attributes (IZIDA), funded by the Bulgarian National Science Fund in 2017–2019, and the project DFNP-100/04.05.2016 "Automatic analysis of clinical text in Bulgarian for discovery of correlations in the Diabetic Registry" funded by the Bulgarian Academy of Sciences in 2016-2017. The team acknowledges the support of Medical University – Sofia, the Bulgarian Ministry of Health and the Bulgarian National Health Insurance Fund.

References

Rakesh Agrawal and Ramakrishnan Srikant. 1994. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, VLDB '94, pages 487–499. http://dl.acm.org/citation.cfm?id=645920.672836.

Nai-Wen Chang, Hong-Jie Dai, Jitendra Jonnagaddala, Chih-Wei Chen, Richard Tzong-Han Tsai, and Wen-Lian Hsu. 2015. A context-aware approach for progression tracking of medical concepts in electronic medical records. *Journal of biomedical informatics* 58:S150–S157.

Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics* 34(5):301–310.

Joshua C Denny, Plomarz R Irani, Firas H Wehbe, Jeffrey D Smithers, and Anderson Spickard III. 2003. The knowledgemap project: development of a concept-based medical school curriculum database. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, volume 2003, page 195.

Stefan Gindl. 2006. Negation detection in automated medical applications. *Vienna: Vienna University of Technology* .

Sergey Goryachev, Margarita Sordo, and Qing T Zeng. 2006. A suite of natural language processing tools developed for the i2b2 project. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, volume 2006, page 931.

Jiawei Han, Jian Pei, Yiwen Yin, and Runying Mao. 2004. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data mining and knowledge discovery* 8(1):53–87.

Rave Harpaz, Alison Callahan, Suzanne Tamang, Yen Low, David Odgers, Sam Finlayson, Kenneth Jung, Paea LePendu, and Nigam H Shah. 2014. Text mining for adverse drug events: the promise, challenges, and state of the art. *Drug safety* 37(10):777–790.

Jitendra Jonnagaddala, Siaw-Teng Liaw, Pradeep Ray, Manish Kumar, Nai-Wen Chang, and Hong-Jie Dai. 2015. Coronary artery disease risk assessment from unstructured electronic health records using text mining. *Journal of biomedical informatics* 58:S203–S210.

Kory Kreimeyer, Matthew Foster, Abhishek Pandey, Nina Arya, Gwendolyn Halford, Sandra F Jones, Richard Forshee, Mark Walderhaug, and Taxiarchis Botsis. 2017. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *Journal of Biomedical Informatics* .

Katherine P Liao, Tianxi Cai, Guergana K Savova, Shawn N Murphy, Elizabeth W Karlson, Ashwin N Ananthakrishnan, Vivian S Gainer, Stanley Y Shaw, Zongqi Xia, Peter Szolovits, et al. 2015. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *bmj* 350:h1885.

Yuan Luo, William K Thompson, Timothy M Herr, Zexian Zeng, Mark A Berendsen, Siddhartha R Jonnalagadda, Matthew B Carson, and Justin Starren. 2017. Natural language processing for ehr-based pharmacovigilance: A structured review. *Drug Safety* pages 1–15.

Amber Stubbs, Christopher Kotfila, Hua Xu, and Özlem Uzuner. 2015. Identifying risk factors for heart disease over time: Overview of 2014 i2b2/uthealth shared task track 2. *Journal of biomedical informatics* 58:S67–S77.

Dimitar Tcharaktchiev, Sabina Zacharieva, Galia Angelova, S. Boytcheva, Z. Angelov, P. Marinova, G. Nentchovska, L. Maneva, A. Velitchkov, G. Petrova, K. Koprivarova, I. Stoeva, M. Boyanov, R. Savova, R. Radev, L. Stoykova-Tchorbanova, E. Tasheva, P. Dentcheva, E. Foteva, K. Slavtcheva, B. Stoyanov, A. Stoev, S. Alexieva, E. Kotova, I. Kovatcheva, and T. Tomov. 2015. Building a bulgarian national registry of patients with diabetes mellitus. *Bulgarian Journal of Social Medicine* 2:19–21.

Özlem Uzuner, Xiaoran Zhang, and Tawanda Sibanda. 2009. Machine learning and rule-based approaches to assertion classification. *Journal of the American Medical Informatics Association* 16(1):109–115.