Integrating Data Analysis Tools for Better Treatment of Diabetic Patients

Svetla Boytcheva[1], Galia Angelova[1], Zhivko Angelov[2], Dimitar Tcharaktchiev[3]
[1] Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Sofia, Bulgaria
[2] Adiss Lab Ltd., Sofia, Bulgaria
[3] Medical University Sofia, University Specialized Hospital for Active Treatment of Endocrinology, Sofia, Bulgaria

svetla.boytcheva@gmail.com, galia@lml.bas.bg, angelov@adiss-bg.com, dimitardt@gmail.com

Abstract. This paper presents the construction and usage of an anonymous Diabetes Register for patients in Bulgaria. The Register is generated automatically from outpatient records submitted to the Bulgarian National Health Insurance Fund in 2010-2014 and continuously updated using outpatient records for 2015-2016. The construction relies on advanced automatic analysis of free text information as well as on Business Analytics technologies for storing, maintaining, searching, querying and analyzing data. Original frequent pattern mining algorithms enable to find patterns and sequences taking into account temporal information. The paper discussed the software environment as well as experiments in frequent pattern mining that enable knowledge discovery in the very large repository underlying the Register (currently 262 million pseudonymized outpatient records submitted to the Bulgarian National Health Insurance Fund in 2010-2016 for more than 5 mln citizens yearly). The claim is that the synergy of modern analytics tools transforms a static archive of clinical patient records to a sophisticated software environment for knowledge discovery and prediction.

Keywords: synergy of data management, data mining and text mining tools; clinical data; frequent pattern mining; data analytics; natural language processing; knowledge discovery

1 Introduction

Medicine is known as a Data Intensive Domain: due to the recent penetration of the Information and Communication Technologies (ICT) in all areas of our society, a rapidly increasing amount of medical data is produced by the healthcare sector, on the one hand, and by biomedical research on the other hand. In the healthcare sector, ICT applications support health diagnostics, development and maintenance of medical Electronic Health Records, telemedicine and telecare, patient administration, almost all aspects of healthcare management and healthcare delivery as well as medical education and training. In biomedical research, progress in deeper understanding of medical phenomena is sought by construction of big data models: e.g. virtual physiological human, models of brain, in computational genetics and so on. Public access to health information is changing the relationship between the patients and the health institutions that are responsible for care delivery. The monitoring and control function of patient organizations is facilitated by the modern ICT tools as well. Today we are still in an early phase of a long-term technological and social shift that will be implied by advancing further the ICT fundamentals and tools.

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10-13, 2017

In this paper we present the integration of various ICT tools for automatic generation of a Diabetes Register for Bulgarian patients. The huge amount of clinical data, underpinning a repository of Outpatient Records (ORs), enabled to construct interfaces that support both *monitoring* functionalities (oriented to the health management authorities) and *research*-oriented functionalities for knowledge discovery. The monitoring functionalities are based on business analytics while research tools use data mining and pattern search. The software environment includes also components for automatic analysis of free texts in Bulgarian. These components facilitate the Register generation and its update because they deliver values of clinical tests and lab data which are described as unstructured text only.

This paper is structured as follows. Section 2 overview related work in several areas that are relevant to the subject: Diabetes registers, Natural Language Processing (NLP) for clinical narratives, Business Intelligence (BI) and analytics, Frequent Pattern Mining (FPM). Section 3 presents the experimental study context and summarizes the developments during the last 3-4 years (because the Register was built iteratively). Section 4 presents relevant achievements in automatic analysis of clinical narratives in Bulgarian language. Section 5 discusses recent algorithms for frequent pattern mining and presents experiments related to knowledge discovery in the Register repository. Section 6 contains the conclusion and plans for future work.

2 Related Work

2.1 Diabetic Registers

There are several nation-wide Diabetes Registers in the world, e.g. in Denmark [1], Sweden [2], Norway [3]. Registers explicate the number of patients who are diagnosed with Diabetes and provide good monitoring and control. Constructing registers is expensive and burdening the patients as well as the medical experts with additional administrative work. Furthermore, in some countries chronic disease management is not recognized as a part of general medical practice. As for the construction, most medical experts agree that Registers are a must since Diabetes is a chronic disease with significant social consequences. Electronic patient registration systems are proposed like the one in Ireland [4] (but it is not implemented yet). It is interesting to mention that in Sweden, during the Diabetic Register development phase 2001-2005, the registration rate of patients gradually increased and reached 75% which in 2010 still remains stable and is one of the highest in the country [5]. Thus infrastructure construction is a critical issue but data collection and update are further problems that can be solved only by persistency and diligence.

2.2 NLP of Clinical Narratives

Usually automatic analysis of clinical narratives is implemented partially: only fragments of the text are considered. The phrases, selected as "interesting", are typically picked up due to the presence of a word or an entity which are considered "significant". This approach for shallow analysis is called "Information Extraction" (IE). IE from clinical texts matures only recently but its accuracy gradually improves and often exceeds 90% [6]. The review [6] stated in 2008 that "current applications are rarely applied outside of the laboratories they have been developed in, mostly because of scalability and generalizability issues". Today, however, this is valid for languages other than English because, with the active contribution of numerous research groups in the USA, NLP for English clinical narratives has much better performance at present. Comprehensive language resources exist for English, such as UMLS [7] as well as tools like KnowledgeMap Concept Identifier [8] which processes clinical notes and returns CUIs (Concept Unique Identifiers) for the recognized UMLS terms. Another important tool is the public NegEx system which identifies and interprets negations in English clinical texts [9, 10]. We also mention the open-source cTAKES[1] (clinical Text Analysis and Knowledge Extraction System) and the Health Information Text Extraction (HITEx) system [11]. A recent study [12] enumerates the advantages to incorporate NLP for English in medical systems: it systematically links several terms to a concept using databases that standardize 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. Despite the NLP limitations, the conclusion is that NLP engines are powerful components ready for integration in medical data mining and – due to improvements expected in the future, e.g. more accurate mappings of terms to medical concepts – the importance of NLP as a valuable supporting technology will grow.

Here we consider NLP for Bulgarian clinical text. No comprehensive resources exist for Bulgarian medical language; the International Classification of Diseases ICD-10 is the only terminological resource which is available in electronic format. Our experience shows that within 2-3 years one can achieve good performance in separate extraction tasks. We apply software prototypes developed some years ago that are gradually improved. The most useful tools are a drug extractor (it finds in the free text the drug name, dosage, frequency and route of admission [13]) as well as an extractor of numeric values of lab data and clinical tests [14].

2.3 Big Data, Business Intelligence Tools

Big Data usually designates a massive volume of structured and unstructured data, too large or too dynamic to be processed by traditional software tools and techniques. The popular "3Vs" features of Big Data were first introduced by Gartner (previously META group): "high Volume, high Velocity, and/or high Variety" [15]. Wikipedia is an example for big data consisting of unstructured texts, images and hyperlinks. Big data analytics is the process of collecting, organizing and analyzing big data to discover useful information. Business Intelligence tools analyze big data of enterprises in order to provide historic, present and predicted views to the business processes. Predictive analytics for establishment of trends is the preferred functionality in contrast to databases that deal with data items and extract subsets of data values. Visualization is an important feature of BI tools because they show generalizations and tendencies in one screen [16]. Another necessary feature is the speed of processing since big data often appear in real time.

In our project we use a BITool which stores data in *n*-dimensional cubes and explores multi-dimensional data i.e. hyperplanes [17]. The user can split the dataset into groups of objects with similar features. If temporal dimension is included the user can track changes of object characteristics over time by animation. BITool enables the discovery of similar situations over time when a search pattern is specified for a particular period.

2.4 Frequent Pattern Mining

There are two principal tasks in pattern search: frequent pattern mining (FPM) where the events (objects) are considered as unordered sets, and frequent sequence mining (FSM). Approaches for solving the FPM task vary from the naïve BruteForce and Apriori algorithms, where the search space is organized as a prefix tree, to Eclat algorithm that uses tidsets directly for support

[1] http://ctakes.apache.org/

computation by processing prefix equivalence classes [18]. Most FPM and FSM methods do not consider contextual information about extracted patterns. They usually build a (huge) prefix tree. Most FPM algorithms generate all possible frequent patterns (FPs). Summarized information for data relations can be extracted as maximal frequent itemsets (MFI) in order to reduce redundancy and decrease significantly the number of FPs for post-analysis. All classic algorithms for FPM can be modified for MFI search.

We have proposed a novel algorithm for mining sets of events in order to identify strong co-occurrence of patterns [19]. It is a cascade data mining approach for FPM enriched with context information which aims at the discovery of complex relations between medical events with respective timestamps. Experiments with this approach are presented in Section 5 to illustrate the functionality of the Diabetes Register as a research tool.

3 Experimental Study

3.1 Principal Objective

A pseudonymized Register of diabetic patients was generated in 2015 from the Outpatient Records, collected by the Bulgarian National Health Insurance Fund (NHIF), in compliance with all legal requirements for safety and data protection [20]. The usual patient registration process was kept without burdening the medical experts with additional paper work. NHIF is the only obligatory Insurance Fund in Bulgaria so we note that working with ORs ensures 100% registration of all patients who contacted the healthcare system at all (however there are Bulgarian citizens who are not insured and some others who have ORs but are not properly diagnosed with Diabetes). The data repository, underpinning the Register, currently contains more than 262 mln pseudonymised ORs submitted to the NHIF in 2010-2016 for more than 7.3 mln Bulgarian citizens (more than 5 mln yearly), including 483,836 diabetic patients. In Bulgaria ORs are produced by General Practitioners (GPs) and Specialists from Ambulatory Care whenever they contact patients. Despite the primary accounting purpose ORs summarize sufficiently the case and motivate the requested reimbursement. They are semi-structured files with predefined XML-format. Many indicators in the Register copy the structured data submitted to NHIF in ORs: *(i)* date and time of the visit; *(ii)* pseudonymized personal data, age, gender; *(iii)* pseudonymised visit-related information; *(iv)* diagnoses in ICD-10; *(v)* NHIF drug codes for medications that are reimbursed; *(vi)* a code if the patient needs special monitoring; *(vii)* a code concerning the need for hospitalization; *(viii)* several codes for planned consultations, lab tests and medical imaging.

ORs contain also important values presented in free text fields: glycated haemoglobin (HbA1c), body mass index (BMI), weight, blood glucose and blood pressure etc. These values are essential for a Diabetic Register so they are extracted automatically from four XML fields: *(i) Anamnesis:* summarizes case history, previous treatments, often family history, risk factors; *(ii) Status*: summary of patient state, height, weight, BMI, blood pressure etc.; *(iii) Clinical tests*: values of clinical examinations and lab data listed in arbitrary order; *(iv) Prescribed treatment:* codes of drugs reimbursed by NHIF, free text descriptions of other drugs. Integration of large scale text analysis is a real novelty in this field.

3.2 Analytics Using BITool

Today the system BITool supports the Diabetes Register at the University Specialized Hospital for Active Treatment of Endocrinology ″Acad. Ivan Penchev″, Medical University – Sofia (this Hospital was authorized by the Bulgarian Ministry of Health to host the Register of diabetic patients in Bulgaria). BITool's functionalities enable the monitoring of significant indicators like glycated hemoglobin (HbA1c) and blood glucose values. In this way the Register achieves its objective: to provide an adequate monitoring strategy for diabetic patients and to improve the healthcare and quality of life for the patients and their families. Two examples illustrate the services. Figure 1 shows the number of diabetic patients in the dimensions age-gender (at certain moment). Here BITool operates on the structured information from the NHIF archive: patient pseudonym, age and gender. Further statistics of this kind might concern explorations of diabetic patients per region code, types of diabetes and diabetes complications, per GPs, per types of medication, according to frequency of visits etc.

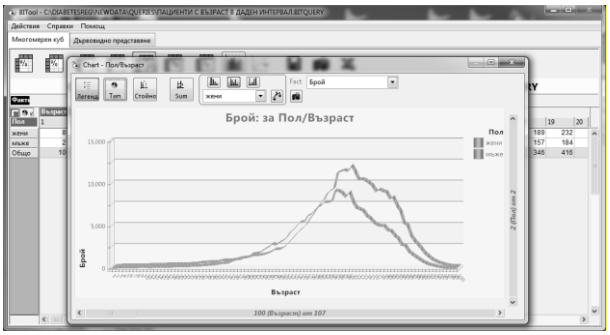

Figure 1 Number of diabetic patients grouped by age

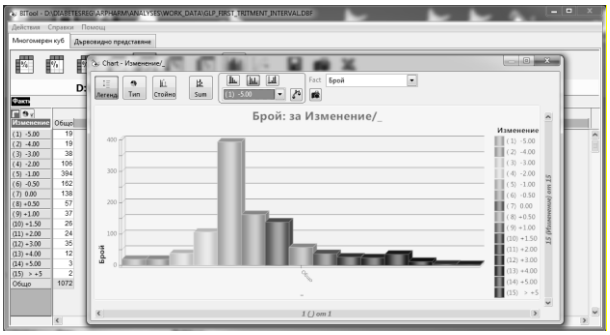

Figure 2 Reduction of HbA1c levels after application of incretin[2] based drugs

[2] https://www.drugs.com/drug-class/incretin-mimetics.html

Figure 2 explores the tendency in the development of treatment. It displays the number of patients who had changes in the HbA1c levels within the interval [-5,5] units for certain period of time. For most patients the HbA1c level decreased by 1 unit. The HbA1c levels are extracted from the free text of ORs for the corresponding patients with timestamp.

Finally we show the Register interface during the process of exploring the collection of ORs (Figure 3). The names and personal identifiers of patients and GPs are replaced by pseudonyms; only the name of the city/village remains in the address field.

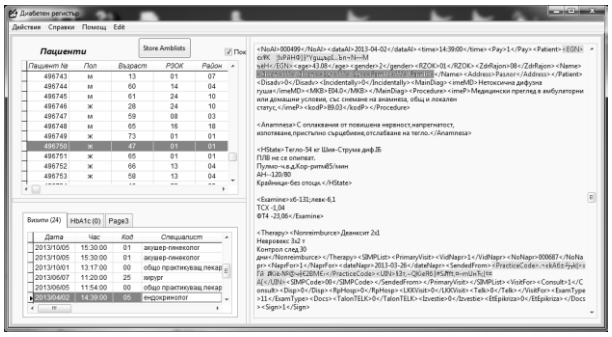

Figure 3 Exploring outpatient records in the Register

4 NLP for Bulgarian Clinical Narratives

Design and implementation of software for automatic extraction of patient-related entities from a Big Data collection is a quite challenging task. One needs to scale up existing research prototypes to process millions of patient records, coping with noisy and missing data, and still providing reliable results. Some numeric entities refer to key risk factors for development of Diabetes Mellitus (levels of glycated hemoglobin HbA1c and blood glucose) and cardio-vascular diseases (high blood pressure). Unfortunately in the Bulgarian clinical practice these values are usually documented in free text paragraphs, presented in a huge variety of formats, so their automatic identification is difficult. We note that according to some studies, today more than 80% of the patient-related clinical information is stored as free text in the Electronic Health Record systems.

In [21] we proposed a hybrid method for automatic generation of grammar rules for IE from clinical data. The experiments were made and evaluated over approximately 9.5 million of ORs. Here we cite only the evaluation of blood pressure extraction from the ORs of about 1,800,000 patients with arterial hypertension for 3 year period: all available values are about 38.3 million and the extraction was performed with precision 92% and recall 98%. The variety of recording formats and explanations written by thousands of medical professionals require constant evaluation of grammar coverage and extraction accuracy in general. Some of the main advantages of the proposed method, beyond its reliable performance and good precision in text mining, are the modularity, extensibility, and scalability.

5 Research in Frequent Pattern Mining

5.1 Contextual Information

Most FSM and FPM approaches do not use contextual information about extracted patterns. These algorithms extract general templates but do not answer the major question whether they are influenced in some way by the context and whether they are valid in various aspects. Existing methods which search for patterns using contextual information are based on attributes that are organized into hierarchical structures and on attributes' generalizations and specializations.

Context information is organized as attributes of itemsets and tidsets. Attributes may have different organization - structured or unstructured. This enables to explore the context-dependent templates. Rabatel et al. [22] propose an approach in marketing domain taking into account not only the transactions that have been made but also various attributes associated with customers like age, gender etc. Attributes have a hierarchical structure $(H(Age), H(Gender))$ and explore patterns at different levels of attributes abstraction – lattice H (Figure 4). Traditional methods consider only the top level $[*,*]$ - for any age and regardless of gender, i.e. without attributes. Rabatel et al. designed the algorithm Gespan and made experiments with about 100,000 product descriptions from *amazon.com*.

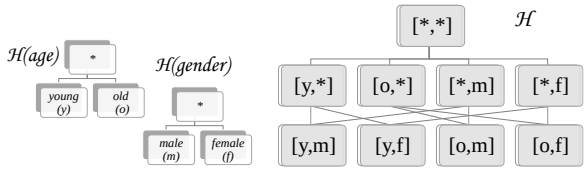

Figure 4 Structuring attributes in marketing domain

Ziembiński [23] proposes a new approach for extracting small contextual models from smaller collections of data that later are summarized in generalized models using information from contextual models with common information. This approach applies a metrics for measuring distance of context models. All values for similarity assessment are normalized in the range between 0.0 and 1.0. Attribute values are considered identical if the similarity function returns 1.0. In the opposite case the result is 0.0. This approach allows extracting patterns for data that would otherwise have to be dropped out of the templates because of its dispersion and low frequency.

5.2 Experimental Setup

We apply a retrospective analysis for patients from the Diabetes Register with Diabetes Type 2. The period of interest is two years preceeding the onset of the Diabetes Type 2, i.e. the so called prediabetes condition. In order to illustrate the potential of contextualized FPM we present results in searching comorbidities for patients in prediabet condition. Text mining modules are used to convert raw text descriptions to structured event data.

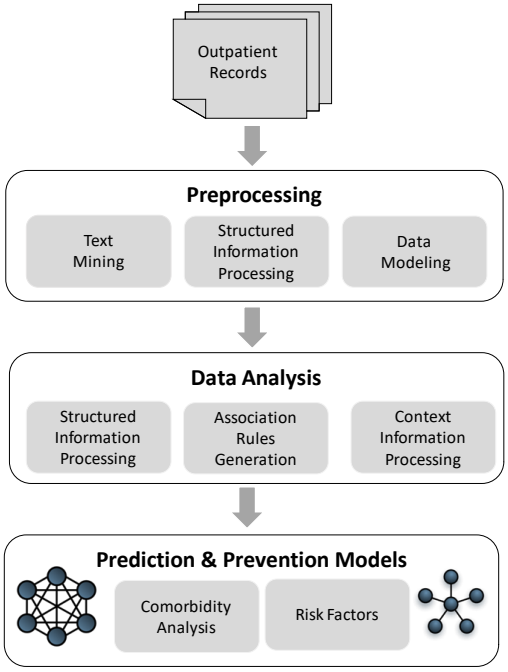

Figure 5 System Architecture

The search space is very large: the database is big, the number of diseases is also large. We propose a tabular method using a vertical database, depth-first traversal as well as set intersection and diffsets [19]. Further processing of the maximal frequent itemsets (MFI) is applied to remove diagnostic-related groups. In addition some context information is added to each MFI to investigate comorbidities. Furthermore association rules with lift are generated. The context information is represented as attribute-value tuples for each patient; the post-processing identifies the importance of different attributes for each MFI.

The architecture of the experimental workbench is shown on Figure 5. Our research [19] aims to develop further the ideas of the two contextual approaches for data mining [22, 23].

For the collection S of ORs we extract the set of all different patient identifiers $P = \{p_1, p_2, \dots, p_N\}$. This set corresponds to transaction identifiers (*tids*) and we call them *pids* (patient identifiers). We consider each patient visit to a doctor as a single event. For each patient $p_i \in P$ an event sequence of tuples $\langle event, timestamp \rangle$ is generated: $E(p_i) = \big(\langle e_1, t_1\rangle, \langle e_2, t_2\rangle, \dots, \langle e_{k_i}, t_{k_i}\rangle\big),\ i = \overline{1, N}$. Let $\mathcal{E}$ be the set of all possible events and $\mathcal{T}$ be the set of all possible timestamps. Let $I = \{id_1, id_2, \dots, id_p\}$ be the set of all diseases ICD-10[3] codes, which we call *items*. Each subset $X \subseteq I$ is called an *itemset*. We define a projection function $\pi: (\mathcal{E} \times \mathcal{T})^N \to 2^I$: $\pi\big(E(p_i)\big) = I(p_i) = \big(id_{1i}, id_{2i}, \dots, id_{m_i}\big)$, such that for each patient $p_i \in P$ the projected time sequence contains only the first occurrence (onset) of each disorder recorded in $E(p_i)$. Let $D \subseteq P \times 2^I$ be the set of all itemsets in our collection after projection π in the format $\langle pid, itemset \rangle$. We shall call D a *database*. We are looking for itemsets $X \subseteq I$ with frequency $(\sup(X))$ above given $minsup$. Let $\mathcal{F}$ denote the set of all frequent itemsets, i.e. $\mathcal{F} = \{X |\ X \subseteq I\ and\ \sup(X) \geq minsup\}$. A frequent itemset $X \in \mathcal{F}$ is called *maximal* if it has no frequent supersets. Let $\mathcal{M}$ denote the set of all maximal frequent itemsets, i.e. $\mathcal{M} = \{X |\ X \in \mathcal{F}\ and \nexists\, Y \in \mathcal{F}, such\ that\ X \subset Y\}$. Let 2^X denote the power set (set of all subsets) of itemset X. Then each subset of $X \in \mathcal{F}$ is also a frequent itemset, i.e. $\forall\, Y \in 2^X\ implies\ that\ Y \in \mathcal{F}$. For each item $id \in I$ we define the set called *pidset*: $p(\mathrm{id}) = \{p_i |\ \langle p_i, I(p_i)\rangle \in D\ and\ id \in I(p_i)\}$.

To study the nature of comorbidities we need to investigate the context in which they occur. Therefore we add some semantic attributes to each event – demographics of patients, age and gender, treatment, status, lab data and etc.

We define a set of attributes of interest $A = \{a_1, a_2, \dots, a_k\}$. Context Q for some patient $p_i \in P$ is defined as the set of attribute-value pairs from patient profile information:

$$Q(p_i) = \{\langle a_1, q_1\rangle, \langle a_2, q_2\rangle, \dots, \langle a_k, q_k\rangle\}.$$

In order to decrease the number of possible values of attributes we apply some aggregation of data. For instance age value is categorized according to the World Health Organization (WHO) standard age groups. Data for body mass index (BMI) are also categorized according to the WHO[4] standard classification - *underweight*, *normal weight*, *overweight*, *obesity*.

For some data concerning demographic information, like region ID we have large number of distinct values. For such data we add also some additional properties concerning background information for the region – e.g. whether it is *south*, *north*, *west*, *east*, *central*, *northwest* etc., and *mountain*, *river*, *sea*, *thermal spring*, *urban region* etc. For status and clinical test data we take the worst value for the period, according to the risk factors deinition.

In primary interest for Diabetes Type 2 are BMI, glycated haemoglobin, blood pressure (RR – Riva Roci), blood glucose, HDL-cholesterol.

From $Q(p_i)$ we generate a feature vector $v(p_i) = (v_{1i}, v_{2i}, \dots, v_{mi})$, where each attribute $a_j \in A$ with N_j possible values is represented by N_j consecutive positions in the vector. For the set of maximal frequent itemset $\mathcal{M}$ with cardinality $|\mathcal{M}| = \mathrm{K}$ we have K classes of comorbidities. We apply classification of multiple classes in order to generate rules for each comorbidity class. We use large scale multi class classification because we deal with a big database and a large group of comorbidity classes. We use Support Vector Machines (SVM) and optimization based on block minimization method described by Yu et al. [24].

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

[4] WHO, BMI Classification http://apps.who.int/bmi/index.jsp?introPage=intro_3.html

Table 1. Data analysis results for patients in prediabetes condition

Set	2013		2014		2013-2014	
items	ICD-10 3 signs	ICD-10 4 signs	ICD-10 3 signs	ICD-10 4 signs	ICD-10 3 signs	ICD-10 4 signs
Patients	27,082	27,082	27,902	27,902	29,205	29,205
Outpatient Records	267,194	267,194	296,129	296,129	556,323	556,323
ICD-10 codes	1,142	4,701	1,145	4,834	1,257	5,503
minsup	0.01	0.01	0.01	0.01	0.01	0.01
Total MFI	203	486	219	512	521	1,406
Longest MFI	5	8	5	9	6	9
Frequent Itemsets	608	7,452	689	8,935	1,909	32,093
Association Rules	686	58,299	810	78,052	2,722	381,012

5.3 Experiments and Results

We report results for patients with Diabetes Type 2 onset in 2015. The ORs of these patients for the period 2013-2014 were excerpted from the Diabetes Register when, as we assume, these patients were in a pre-diabetes condition. The idea of this experiment is to check whether we can successfully discover risk factors for these patients looking only at their ORs in 2013 and 2014. Then, maping our hypotheses to the real data for 2015, we test whether our approach is reasonable. (We note that due to the relatively short period of observation and lack of data about mortality, at the moment we cannot follow diabetes development in longer periods.)

In the Register each OR, corresponding to a single visit, cointains up to 4 diagnoses encoded in ICD-10. Some diagnoses are presented by 4-sign encodings, i.e. in a more specific way, while others use the more general 3-sign encoding. Due to the hierarchical organization of ICD-10 we shall analyse individually two collections: the original one, that is more specific (with 4-sign codes - see Example 1) and we shall generalise also all diagnoses to more general classes (with 3-sign codes - see Example 2). The examples present collections of diagnoses for a patient with ID 2196365.

Example 1:
I(2196365)={I10, M10.9, M10, K76.9, K76, L94.1, L94, M06.9, G57.9, Z00.8, H53, M51.1, M33.9}

Example 2:
I(2196365)={I10, M10, K76, L94, M06, M51, M33, H53, Z00, G57}

For some patients, the available ORs contain no information about certain attributes of the context information (Table 2). It is well known that missing data in medical documentation is inevitable. Thus some attribute values are replaced by the value NA, which is considered as the most general value.

For example the context information for the patient with ID 2196365 is:

$$Q\ (2196365) = \{\langle age, 58\rangle,\ \langle gender, 1\rangle,\ \langle region, 03\rangle,\ \langle bmi, 29.32\rangle,\ \langle hba1c, NA\rangle,\ \langle blood\_glucose, 6.39\rangle, \langle hdl\_cholesterol, 1.15\rangle\}$$

Table 2. Data for attributes in the collections

A	attribute	2013	2014	2013-2014
a_1	*age*	27,082	27,902	29,205
a_2	*gender*	27,082	27,902	29,205
a_3	*region*	27,082	27,902	29,205
a_4	*bmi*	21,659	22,413	27,928
a_5	*HbA1c*	153	238	370
a_6	*HDL cholesterol*	4,917	4,815	6,952
a_7	*blood glucose*	11,925	12,185	17,016

One of the generated Maximal Frequent Itemsets (comorbidity class), whose support contains the pid= 2196365, is:

```
MFI#12: Z00 I10 M51 #SUP: 453
```

The following charts show the distribution of patients in the support of "MFI#12" according to their age (Figure 6), gender (Figure 7), BMI (Figure 8), and the HDL Cholesterol (Figure 9) correspondingly.

We can observe that most patients in this support set have higher risk of Diabetes Type 2, due to the presence of multiple risk factors as obesity, medium or high levels of cholesterol and hypertension (diagnose with ICD-10 code I10).

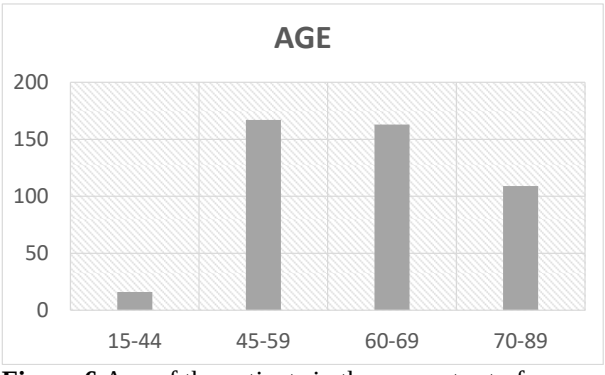

Figure 6 Age of the patients in the support set of "MFI#12"

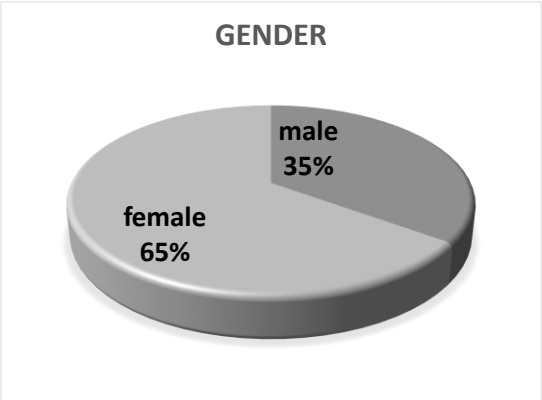

Figure 7 Gender of the patients in the support set of "MFI#12"

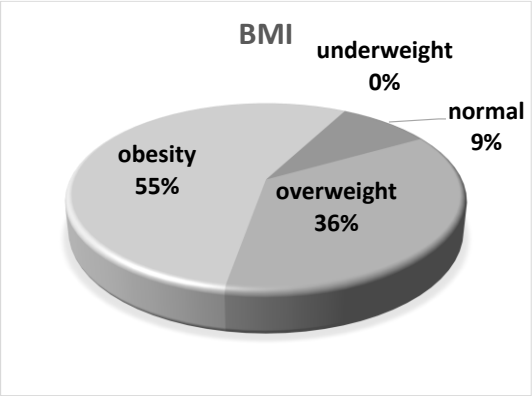

Figure 8 BMI of the patients in the support set of "MFI#12"

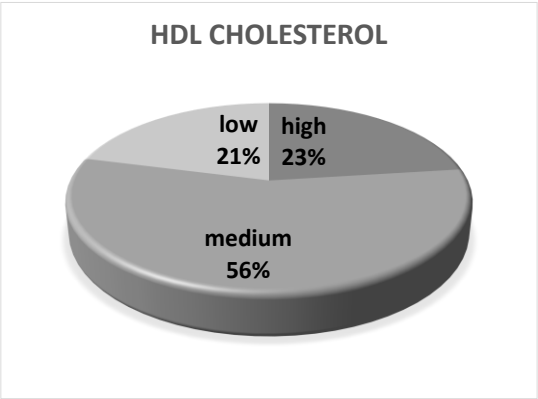

Figure 9 Levels of HDL Cholesterol of the patients in the support set of "MFI#12"

Data about HbA1c are available only for 3 out of 453 patients, that is why we consider this attribute as a more general value ANY. But we note that the lack of HbA1c measurements is not surprising because tests for HbA1c are made when the Diabetes is diagnosed (and this has happened in 2015 for the selected patient cohort).

Data for blood glucose are available only for 30% of these patient and for 50% of them the values were high.

Deeper analyses reveal medical arguments why higher risk exist especially for the patients in the support set of MFI#12: Z00 I10 M51 #SUP: 453. The diagnose with ICD-10 code M51 (Thoracic, thoracolumbar, and lumbosacral intervertebral disc disorders) means that the patients have lower motor activity and sedentary lifestyle, which causes obesity, overweight, higher values of cholesterol and blood pressure and therefore increases the risk of developing Diabetes. Actually this has happened in 2015. We note that in general the ICD-10 diagnose M51 is not considered risky for Diabetes. But our algorithm reveals this unknown and latent interrelationship.

6 Conclusion and Future Work

In this paper we present a software environment for collection and processing of Big Data in medicine - a Data Intensive Domain. The Diabetes Register has been developed stepwise and its research functionality is still under construction. We believe that the integration of various technologies is the proper way to approach the challenges of large-scale information processing because the integration ensures flexible multi-functionality and enables reuse of results.

The nation-wide Diabetes Register of Bulgaria is now visible in Internet[5] together with some public statistical information. We plan to develop the Register further as a predictive and preventing tool using the synergy of advanced technologies which enable to discover risk groups of patients that have predisposition to various socially-significant diseases. We have shown here that the present software environment is mature enough to identify patients with complexes of risk factors for development of Diabetes, e.g. risks like: family history (relatives with Diabetes); obesity; arterial hypertonia (RR≥140/90); low physical activity; giving birth to a baby with weight more than 4 kg or gestational Diabetes; established impaired fasting glycaemia or impaired glucose tolerance; other states of insulin resistance (e.g. acanthosis nigricans, a specific hyperpigmentation of the skin that might be due to endocrine disorders); HDL-cholesterol≤0.90 mmol/l or triglycerides≥2.2 mmol/l (≥2.82 mmol/l according to ADA); diagnosed polycystic ovarian syndrome, a cardio-vascular disease, or mental disorders etc. These risk factors are explicated in the patient-related documents either by values of clinical tests or by keywords and typical phrases that describe the factor. The patients with predisposition suffer from disorders and syndromes, diagnosed by various medical specialists in various time periods, but without any chance to establish connections between the medical doctors – e.g. a connection between a Psychiatrist and a Cardiologist that have consulted the patient. Elaborating further the analytics facility of the Register will provide functionality to monitor patient status over time, in the context of all available information, and to issue alerts for coincidence of risk factors that open the door to Diabetes and other chronic diseases. In this way we believe that in the foreseeable future it will become possible to identify the Bulgarian citizens who have predisposition to develop Diabetes Mellitus.

[5] http://usbale.com/Register_Diabetes.htm

Acknowledgements

The research presented here is partially supported by the grant 02/4 *SpecialIZed Data MIning MethoDs Based on Semantic Attributes* (IZIDA), funded by the National Science Fund in 2017–2019. The support of Medical University – Sofia, the Bulgarian Ministry of Health and the National Health Insurance Fund is acknowledged.

References

[1] Carstensen, B. et al.: The Danish National Diabetes Register: Trends in incidence, prevalence and mortality. Diabetologia. 51(12), 2187–2196 (2008). doi: 10.1007/s00125-008-1156-z

[2] Hallgren Elfgren, I. M., Grodzinsky, E., Törnvall, E.: The Swedish National Diabetes Register in clinical practice and evaluation in primary health care. Prim. Health Care Res. Dev. 17(6), 549-558 (2016). doi: 10.1017/S1463423616000098

[3] Cooper, J. G., Thue, G., Claudi, T., Løvaas, K., Carlsen, S., Sandberg, S.: The Norwegian Diabetes Register for Adults – an overview of the first years. Norsk Epidemiologi. 23(1), 29-34 (2013)

[4] O'Mullane, M., McHugh, S., Bradley, C. P.: Informing the development of a national diabetes register in Ireland: a literature review of the impact of patient registration on diabetes care. Inform. Primary Care. 18(3), 157-68 (2010)

[5] Hallgren Elfgren, I.M., Törnvall, E., Grodzinsky, E.: The process of implementation of the diabetes register in Primary Health Care. Int. Journal of Qual. Health Care. 24(4), 419-424 (Aug 2012)

[6] Meystre, S., Savova, G., Kipper-Schuler, K., Hurdle, J. F.: Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research. IMIA Yearbook of Medical Informatics, pp. 138-154. (2008)

[7] UMLS, the Unified Medical Language System. https://www.nlm.nih.gov/research/umls/

[8] Denny, J. C., Irani, P. R., Wehbe, F. H., Smithers, J. D., Spickard, A.: The KnowledgeMap Project: Development of a Concept-Based Medical School Curriculum Database. In: AMIA Annu Symp Proc., pp. 195–199. (2003)

[9] Chapman, W., Bridewell, W., Hanbury, P., Cooper, G. F., Buchanan, B.: A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries**.** Univ. of Pittsburgh (2002)

[10] Gindl, S.: Negation Detection in Automated Medical Applications. TUW (2006)

[11] HITEx Manual: https://www.i2b2.org/software/projects/hitex/hitex_manual.html

[12] Liao, K. P., Cai, T., Savova, G. K., Murphy, S. N., Karlson, E. W., Ananthakrishnan, A. N., Gainer, V. S. et al.: Development of phenotype algorithms using electronic medical records and incorporating natural language processing. British Med. J., 350 (1): h1885 (2015)

[13] Boytcheva, S.: Shallow Medication Extraction from Hospital Patient Records. Studies in Health Technology and Informatics. vol. 166, pp. 119-128. IOS Press (2011)

[14] Tcharaktchiev, D., Angelova, G., Boytcheva, S., Angelov, Z., Zacharieva, S.: Completion of Structured Patient Descriptions by Semantic Mining. Studies in Health Technology and Informatics, vol. 166, pp. 260–269. IOS Press (2011). doi: 10.3233/978-1-60750-740-6-260

[15] Laney, D.: 3D Data Management: Controlling Data Volume, Velocity, and Variety. META Group Research Note, 6, 10 (2001) https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

[16] Top 238 Business Analytics Tools. Predictive Analytics Magazine (Feb 2012). http://www.predictiveanalyticstoday.com/top-business-intelligence-tools/

[17] Angelova, G., Nikolova, I., Angelov, Zh.: Embedding language technologies in a data analytics tool. Advances in Bulgarian Sciences, pp. 29-42. National Centre for Information and Documentation (2016). ISSN: 1314-3565

[18] Nasreen, S., Azam, M. A., Shehzad, K., Naeem, U., Ghazanfar, M. A.: Frequent Pattern Mining Algorithms for Finding Associated Frequent Patterns for Data Streams: A Survey. Procedia Computer Science, 37, 109-116 (2014)

[19] Boytcheva, S., Angelova, G., Angelov, Z., Tcharaktchiev, D.: Mining Comorbidity Patterns Using Retrospective Analysis of Big Collection of Outpatient Records. Health Inf Sci Syst. Journal, Springer (2017). ISSN: 2047-2501 (*to appear*)

[20] Tcharaktchiev, D., Zacharieva, S., Angelova, G., Boytcheva, S., et al. Building a Bulgarian National Registry of Patients with Diabetes Mellitus. Journal of Social Medicine. 2, 19-21 (2015) (*in Bulgarian*)

[21] Boytcheva, S., Angelova, G., Angelov, Z., Tcharaktchiev, D.: Text Mining and Big Data Analytics for Retrospective Analysis of Clinical Texts from Outpatient Care. Cybernetics and Information Technologies, 15(4), 58-77 (2015). doi: 10.1515/cait-2015-0055

[22] Rabatel, J., Bringay, S., Poncelet, P.: Mining sequential patterns: a context-aware approach. Advances in Knowledge Discovery and Management, pp. 23-41. Springer (2013)

[23] Ziembiński, R. Z.: Accuracy of generalized context patterns in the context based sequential patterns mining. Control and Cybernetics. 40, 585-603 (2011)

[24] Yu, H. F., Hsieh, C. J., Chang, K. W., Lin, C. J.: Large linear classification when data cannot fit in memory. ACM Transactions on Knowledge Discovery from Data (TKDD), 5(4), 23 (2012)