Data Mining and Analytics for Exploring Bulgarian Diabetic Register

Svetla Boytcheva[1(✉)], Galia Angelova[1], Zhivko Angelov[2], and Dimitar Tcharaktchiev[3]

[1] Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, 25A Acad. G. Bonchev Street, 1113 Sofia, Bulgaria
svetla.boytcheva@gmail.com, galia@lml.bas.bg
[2] ADISS Lab Ltd., 4 Hristo Botev Blvd, 1463 Sofia, Bulgaria
angelov@adiss-bg.com
[3] University Specialized Hospital for Active Treatment of Endocrinology – Medical University Sofia, 2 Zdrave Street, 1431 Sofia, Bulgaria
dimitardt@gmail.com

Abstract. This paper discusses the need of building diabetic registers in order to monitor the disease development and assess the prevention and treatment plans. The automatic generation of a nation-wide Diabetes Register in Bulgaria is presented, using outpatient records submitted to the National Health Insurance Fund in 2010–2014 and updated with data from outpatient records for 2015–2016. The construction relies on advanced automatic analysis of free clinical texts and business analytics technologies for storing, maintaining, searching, querying and analyzing data. Original frequent pattern mining algorithms enable to discover maximal frequent itemsets of simultaneous diseases for diabetic patients. We show how comorbidities, identified for patients in the prediabetes period, can help to define alerts about specific risk factors for Diabetes Mellitus type 2, and thus might contribute to prevention. We also claim that the synergy of modern analytics and data mining tools transforms a static archive of clinical patient records to a sophisticated knowledge discovery and prediction environment.

Keywords: Big data analytics · Data mining · Frequent pattern mining
Text mining · Health informatics

1 Introduction

Diabetes is an increasingly common disease and a global public health problem that places a considerable economic burden on society. The World Health Organization (WHO) reports that diabetes prevalence among adults has risen from 4.7% in 1980 to 8.5% in 2014. It is expected that diabetes will be the seventh leading cause of death in 2030 [1]. In the recent Global Report on Diabetes WHO recommends: "Strengthen national capacity to collect, analyze and use representative data on the burden and trends of diabetes and its key risk factors. Develop, maintain and strengthen a diabetes registry if feasible and sustainable" [2]. All countries in Europe have national plans for

L. Kalinichenko et al. (Eds.): DAMDID/RCDL 2017, CCIS 822, pp. 19–33, 2018.
https://doi.org/10.1007/978-3-319-96553-6_2

discovery, treatment and prevention of diabetes [3]; seven countries have diabetic registers in 2014 [4]. However, one hardly finds information about the execution of national diabetes plans, monitoring of various plan measures and evaluation of their success. Positive health outcomes are difficult to assess too, moreover this needs to be done dynamically, at national level in order to improve the treatment plans. From a technological point of view, the general impression is that healthcare authorities lack understanding about the potential of modern Information and Communication Technologies (ICT) as an enabling tool that facilitates data collection, monitoring of indicators, knowledge discovery, early alerting and automatic sending of feedbacks, evaluation of updated indicators and automatic preparation of aggregated recaps.

In this paper we discuss the automatic generation of a national Diabetic Register, using outpatient records submitted to the Bulgarian National Health Insurance Fund (NHIF) for the period 2010–2016 and present research efforts to explore the register data by extracting useful information about patients and disease development over time. Some results concern discovery of correlations among data items and have more scientific value while other outcomes are actually aggregated reports addressing the healthcare management. The authors believe that these developments, which are already integrated in the software infrastructure underlying the Diabetic Register and regularly used by the national healthcare authorities, will influence the forthcoming implementation of Electronic Health Record (EHR) system in Bulgaria.

This paper presents novel results extending [5]. Section 2 briefly overviews the need and construction of diabetic registers in Europe. Section 3 presents the Bulgarian Diabetic Register which was generated automatically using a national collection of more than 262 mln outpatient records. We emphasize on the originality of our approach: starting from a very large repository of full-text clinical records, we had to employ more sophisticated software solutions in order to cope with the input data and to provide dynamic exploration of the constantly growing archive of pseudonimyzed outpatient records. Some examples of aggregated reports, prepared by a business analytics tool, demonstrate the potential of the software behind the register. Section 4 shows another data mining tool for discovery of correlations. It sketches an original method for frequent pattern mining and discusses its application for searching of comorbidities in the register. Section 5 contains the conclusion and plans for future work.

2 Diabetic Registers in Europe

The Euro Diabetes Index 2014 compares the figures of diabetes prevalence to previous ones and concludes that prevention and screening in Europe have improved after 2008 because less people die [4]. Patient awareness is raising, devices for self-monitoring become much more accessible, and the variety of medications is growing. However, still a very high number of diabetic patients are undiagnosed and half of the European countries cannot provide reasonably good data of procedure indicators. It is claimed that “as long as important data is not systematically reported and transformed into methodology, diabetes care will remain inefficient and, at worst, haphazard” [4].

On the other hand, it is well known that availability of high-quality data is hard to achieve. Information about diabetic patients is often not collected nationally but rather in hospitals or at regional level, with limited comparability of collected indicators. Available data often come from isolated national projects with fixed duration or EU-funded initiatives like EUBIROD (European Best Information through Regional Outcomes in Diabetes, 2008–2012) [6]. After the project ends, no strategic plans are built by the respective political or governing institutions and in this way projects that started and proved to be successful remain feasibility studies without practical effects.

Seven European countries have diabetic registers in 2014: Sweden, Denmark, Norway, Netherlands, UK, Switzerland, and Hungary. Without making detailed overview of data collection procedures, we emphasize that data input to the registers listed in [4] is ensured either by self-registration or by burdening medical professionals with additional documentation tasks. However self-registration means that a significant percent of the patients remains unregistered. For instance in Sweden, which according to Euro Diabetes Index 2014 is the country with the best diabetes care delivery in Europe, the register was constructed by self-registration. During its development phase 2001–2005 the self-registration rate of patients gradually increased and reached 75% which in 2010 still remains stable and is one of the highest in the country [7]. No information is available about the procedures for register update and maintenance.

The Euro Diabetes Index 2014 summarizes the situation with the nice phrase "*No data, no cure*". Surprisingly, no attempts for automatic extraction of registers from available EHR repositories are mentioned in 2014. In the next section we present our achievement for building a national Diabetes Register as a component of the healthcare system, where clinical narratives can be reused dynamically for ensuring good diabetes care to patients, on the one hand, and reducing the documentation burden to many healthcare professionals, on the other hand.

3 Bulgarian Diabetes Register and Its Exploration

3.1 Automatic Register Generation

A pseudonymized Register of diabetic patients was generated in 2015 from the Outpatient Records (ORs), collected by the Bulgarian NHIF, in compliance with all legal requirements for safety and data protection [8]. 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 using 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 pseudonymized 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 these ORs summarize sufficiently the case and motivate the requested reimbursement. ORs are semi-structured files with predefined XML-format.

Many indicators in the Diabetic Register copy the structured data submitted to NHIF in ORs: *(i)* date and time of the visit; *(ii)* pseudonymized personal data, age, gender; *(iii)* pseudonymized 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.

The ORs contain also values of clinical tests and lab data, presented in the free text fields. Using extractors for automatic text analysis of Bulgarian texts, which have been developed in our previous projects, we mine these values from four OR 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; *(iv) Prescribed treatment*: codes of drugs reimbursed by NHIF, free text descriptions of other drugs.

We develop text mining tools for clinical texts in Bulgarian language since many years. The focus was placed mostly on clinical narratives discussing diabetic patients due to the social importance of this chronic disease. Initially various indicators concerning the patient status were extracted from hospital discharge letters [9], later the attention was shifted to extraction of numeric values of clinical tests and lab data from NHIF outpatient records [10, 11]. A brief overview of natural language processing (NLP) from clinical narratives is provided in [5].

3.2 Business Analytics as an Exploratory Tool

The Diabetes Register contains more than structured indicators in a database; actually data about subsequent visits of all patients to medical doctors is kept so the patient records in the Register have variable length. In addition, all underlying pseudonimyzed outpatient records for all diabetic patients in Bulgaria can be accessed in an efficient manner for detailed full text inspection. Due to this reason, the usual database functionality is insufficient to provide the necessary capacity for search and exploration of the Register repository. Moreover the archive size excludes direct observations by database tables. Our solution is based on business intelligence. As far as we know, this approach to construction and maintenance of medical Registers is unique.

The system BITool supports the Diabetes Register at the University Specialized Hospital for Active Treatment of Endocrinology "Acad. Ivan Penchev", Medical University – Sofia, Bulgaria (authorized by the Bulgarian Ministry of Health to host the Register of diabetic patients in Bulgaria). BITool shows correlations among various indicators, significant for diabetes and its complications, and the prescribed drugs. Given detailed and semi-structured descriptions of all case histories, BITool identifies the importance of various risk factors combinations for diabetes development over time. The relatively complex business analytics functionalities with appropriate visualization extend the main Register purpose from monitoring to prevention. Some examples illustrate the services.

BITool displays the correlation between the compensation of Diabetes and Hypertension for the diabetic patients included in the Register at certain moment (Fig. 1). Age groups show clear distinction between children and adults. Here BITool operates on the structured information from the NHIF archive as patient pseudonym,

age and types of diabetes using also aggregated lab test data. 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 visit frequency etc.

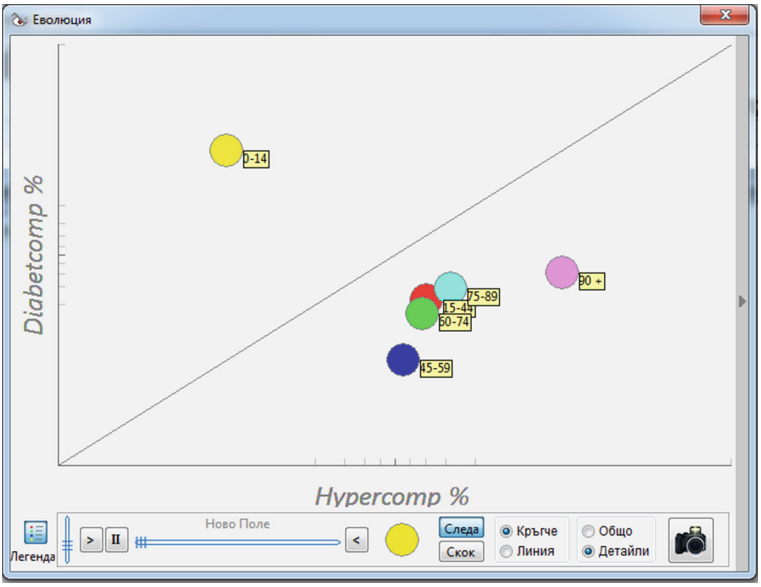

Fig. 1. Compensation of diabetes and hypertension by age groups

BITool easily finds the support (number of patients) for combinations of five risk factors for diabetes development in a cohort of patients without Diabetes (Fig. 2). The patients are outside the Register and data is extracted from the respective ORs using the same software tools that generated the Register. The latter is updated yearly with information provided within an archive of pseudonimyzed ORs for the respective year.

BITool integrates drill down functionality as well; clicking on some item, aggregating a list, moves the user to a level of greater detail. For instance, Fig. 3 shows an aggregated report about drugs prescribed to diabetic patients for 2016. Patient numbers are listed in age groups and genders. Clicking on any number (e.g. "2" in line A10AD, age 0–14, 2 boys in the second column) will open a list with these two patient identifiers and their basic Register indicators, from where access to all the information about them is provided.

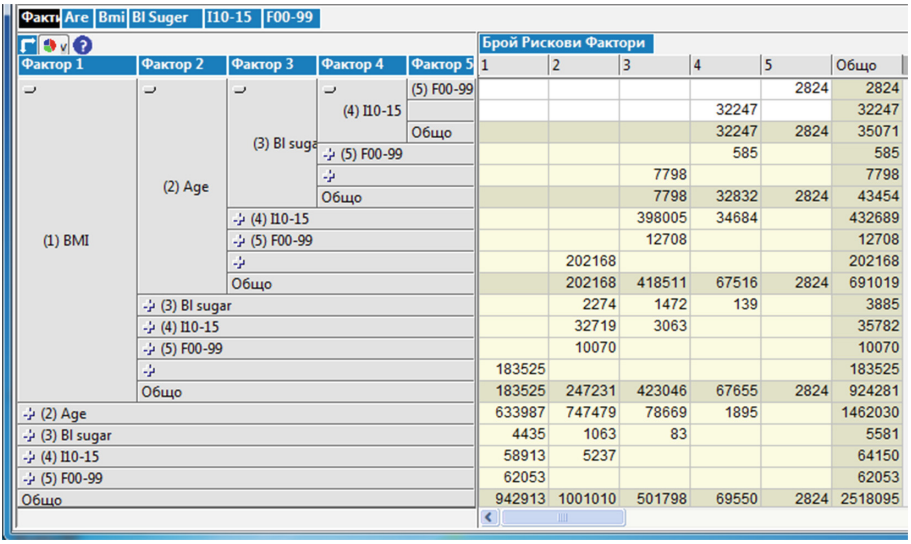

Fig. 2. Monitoring number of citizens with risk factors for diabetes development

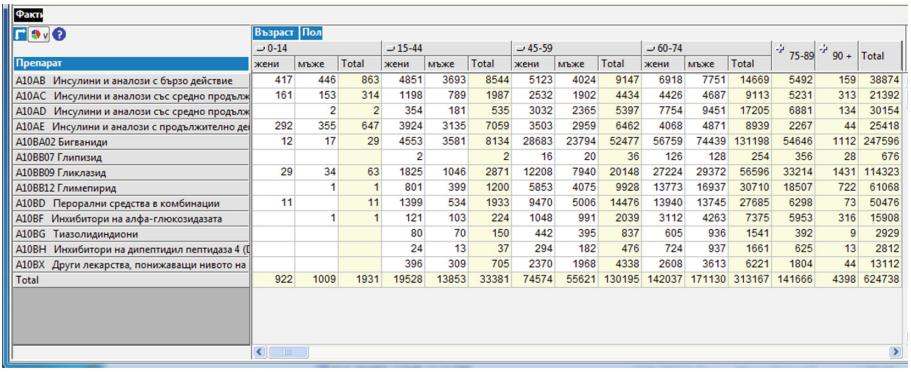

Fig. 3. Groups of drugs (by ATC codes) prescribed to diabetic patients (age, gender) in 2016

4 Frequent Pattern Mining for Knowledge Discovery

4.1 Motivation and Context

The Register is pseudonimyzed, i.e. all ORs for each patient are linked in one case history. Then data mining can be used to discover unknown associations among data items in the Register. The algorithm MixCO for finding Maximal Frequent Itemsets (MFI) in Frequent Pattern Mining (FPM) has been developed [5, 11] and recently we apply it to study associations between diseases (so called comorbidities) for patients with Diabetes Mellitus Type 2 (DM2). Given the importance of early diabetes discovery and prevention, our aim is to identify risk factors using the Register data.

We consider the patients in prediabetes condition taking ORs from the period of two years preceding the onset of DM2. Below we show how retrospective analyses are done using the ORs: some comorbidities are identified for the prediabetes period, they are analyzed and given to medical experts who can define alerts about more complex risk factors for DM2. In general comorbidities are considered as frequent patterns of diagnoses.

Formally, for the collection S of ORs we extract the set of all different patient identifiers $P = \{p_1, p_2, \ldots, p_N\}$. 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: $E(p_i) = (\langle e_1, t_1 \rangle, \langle e_2, t_2 \rangle, \ldots, \langle e_{k_i}, t_{k_i} \rangle), i = \overline{1,N}$. Let $\mathcal{E}$ be the set of all possible events and $\mathcal{T}$ be the set of all possible timestamps. Let $I = \{id_1, id_2, \ldots, id_p\}$ be the set of all diseases ICD-10[1] codes, which we call *items*. Each subset $X \subseteq I$ is called an *itemset*. We define a projection function $\pi: (\mathcal{E} \times \mathcal{T})^N \rightarrow 2^I$: $\pi(E(p_i)) = I(p_i) = (id_{1i}, id_{2i}, \ldots, id_{m_i})$, 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. $\mathcal{F} = \{X | X \subseteq I \text{ and } \sup(X) \geq minsup\}$. 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. $\mathcal{M} = \{X | X \in \mathcal{F} \text{ and } \nexists\, Y \in \mathcal{F}, \text{ such that } X \subset Y\}$. 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*: $p(id) = \{p_i | \langle p_i, I(p_i) \rangle \in D \text{ and } id \in I(p_i)\}$.

The majority of FPM and MFI algorithms consider no contextual information of the processed data [12]. Only few methods for contextual FPM and FSM (frequent sequence mining) use structured background knowledge: hierarchies [13] and ontologies [14], or some metrics to measure distances between the frequent patterns context [15]. Rabatel et al. [13] propose a hierarchical organization of attributes that allows different levels of abstraction. They present an application in the marketing domain based on clustering of frequent patterns of customers depending on their age, gender, etc. in contrast to the classic FSM methods. Ziembiński [15] proposes a new FSM approach for extracting small contextual models from smaller collections of data that are summarized later in generalized models using information from contextual models with common information. A metrics for measuring distance of context models is applied. Huang et al. [14] present one of the first approaches for contextual FPM in EHRs for adverse drug effect monitoring. Two algorithms are proposed: semantic hypergraph-based k-itemset generation and ontology-based k-itemset enrichment. These methods identify some complex patterns which are usually skipped by other FPM algorithms and prove to be very useful in health informatics.

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

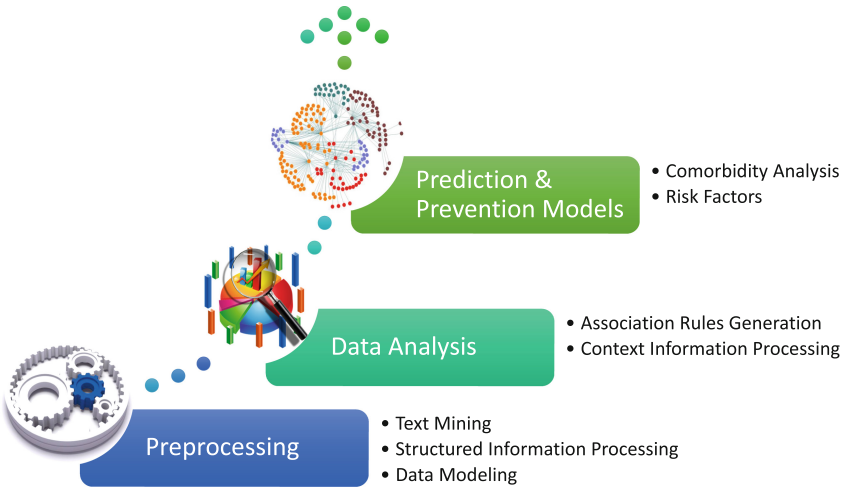

Fig. 4. System architecture

We define a set of attributes of interest $A = \{a_1, a_2, \ldots, a_k\}$. Context Q for some patient $p_i \in P$ is defined as the set of attribute-value pairs from patient profile information: $Q(p_i) = \{\langle a_1, q_1\rangle, \langle a_2, q_2\rangle, \ldots, \langle a_k, q_k\rangle\}$.

From $Q(p_i)$ we generate a feature vector $v(p_i) = (v_{1i}, v_{2i}, \ldots, v_{mi})$, where each attribute $a_j \in A$ with N_j possible values is represented by N_j consecutive positions in the vector. For a set of MFI $\mathcal{M}$ 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 as we deal with a big database (millions of ORs) and a large group of comorbidity classes (ICD-10 contains approx. 12,000 four-sign codes of diagnoses). We use Support Vector Machines and optimization based on block minimization method described by Yu et al. [16].

For searching diseases comorbidities we apply the MixCO algorithm for searching MFI. We propose a cascade data mining approach for MFI enriched with context information. MixCO is a tabular method using a vertical database, depth-first traversal as well as set intersection and diffsets [11].

The architecture of the experimental workbench is shown in Fig. 4. We start with preprocessing by gathering context data and diagnosis codes for FPM. Then we provide data analysis by applying MIxCO and context based analysis. The post-processing identifies the importance of different attributes for each MFI. To study the nature of comorbidities we need to investigate the context in which they occur.

The preprocessing modules combine structured OR data (age, gender, and demographic region, clinic visits and hospitalizations, ATC codes of drugs that are reimbursed by NHIF) and perform free text analysis in order to deliver additional context attributes beyond the structured information about the patients. Text mining tools [9] extract vital parameters (BMI, blood pressure – Riva Rocci), lab tests values (HbA1c, Blood Glucose levels, etc.), and some prescribed therapy (ATC codes of drugs beyond the ones that are reimbursed by NHIF). Due to the huge number of possible distinct

attribute values some aggregation is needed. WHO provides some standard aggregated categories like standard age groups, BMI classification[2] - *underweight*, *normal weight*, *overweight*, *obesity*. An approach for generalization of attributes related to geolocations is presented in [11], it helps for identify associations between patient attributes and the location where they live. For the status and lab test data we take the worst value for the period, according to the risk factors definition.

4.2 Experiments and Results

We discuss experimental results for patients with DM2 onset in 2015. We excerpted from the Diabetes Register the ORs of these patients for 2013–2014 when, as we assume, they were in a prediabetes condition. The idea is to check whether we can successfully discover risk factors for these patients looking only at their ORs in 2013 and 2014. Then, mapping our hypotheses to the real data in 2015, we test whether our approach is feasible (due to the 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, contains up to four 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 analyze individually two collections: the original one, which is more specific (with 4-sign codes) and we generalize also all diagnoses to more general classes (with 3-sign codes). The result of data analysis for patients with DM2 onset in 2015 are shown in Table 1.

Table 1. Data analysis results for patients with prediabetes in 2013–2014

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

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

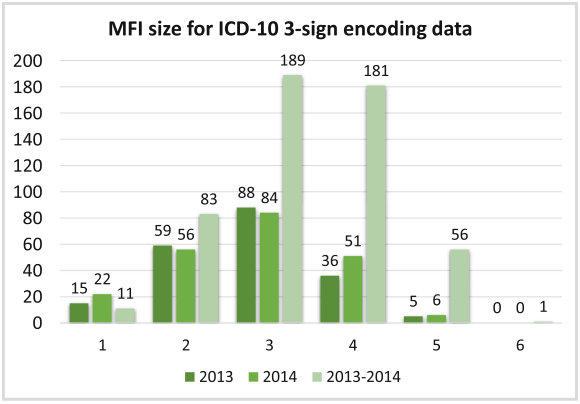

Fig. 5. Distribution of MFI by size for three OR collections with ICD-10 3-sign encodings

The distribution of MFIs by size for three collections with ICD-10 3-sign and 4-sign encodings is shown correspondingly in Figs. 5 and 6.

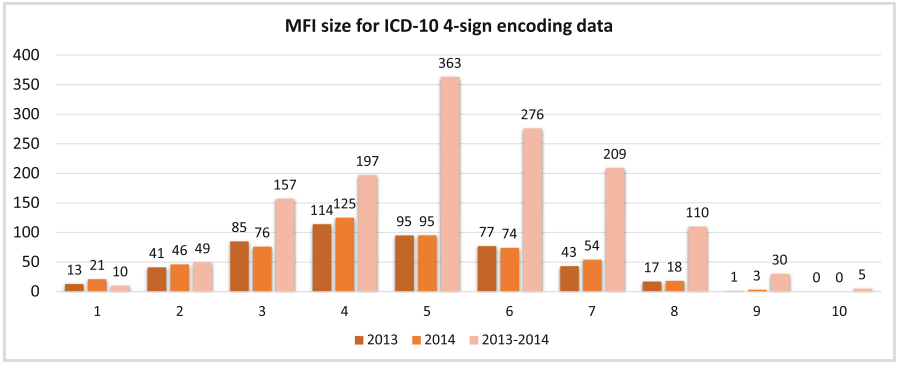

Fig. 6. Distribution of MFI by size for three OR collections with ICD-10 4- sign encodings

The top three strongest (with maximal support) MFI found by the algorithm are shown in Table 2 (ICD-10 3-sign encodings) and Table 3 (ICD-10 4-sign encodings), where the support value is denoted by #S.

Table 2. The top 3 MFI for data collection with 3-sign ICD-10 encodings

2013	2014	2013–2014
I11 I20 H25 #S:671	I11 I20 M51 #S:755	I11 I20 E78 #S:931
I10 H52 #S:667	I11 I20 H25 #SUP:748	I11 I20 J20 #S:847
I11 I20 M51 #S:628	I10 H52 #S:722	I11 I20 K29 #S:831

Table 3. The top 3 MFI for data collection with 4-sign ICD-10 encoding

2013	2014	2013–2014
I11.9 I20.8 I48 #S:583	I11.9 I20.8 I20.9 #S:740	I11.9 I20.8 I11.0 I50.0 I48 #S:400
I11.9 E04 #S:555	I11.9 I20.8 I69.8 #S:736	I11.9 I20.8 I20.9 I11.0 I50.0 #S:372
I10 I20.9 #S: 512	I11.9 I20.8 M17.0 #S:567	I11.9 I20.8 H52.4 H35.0 #S:357

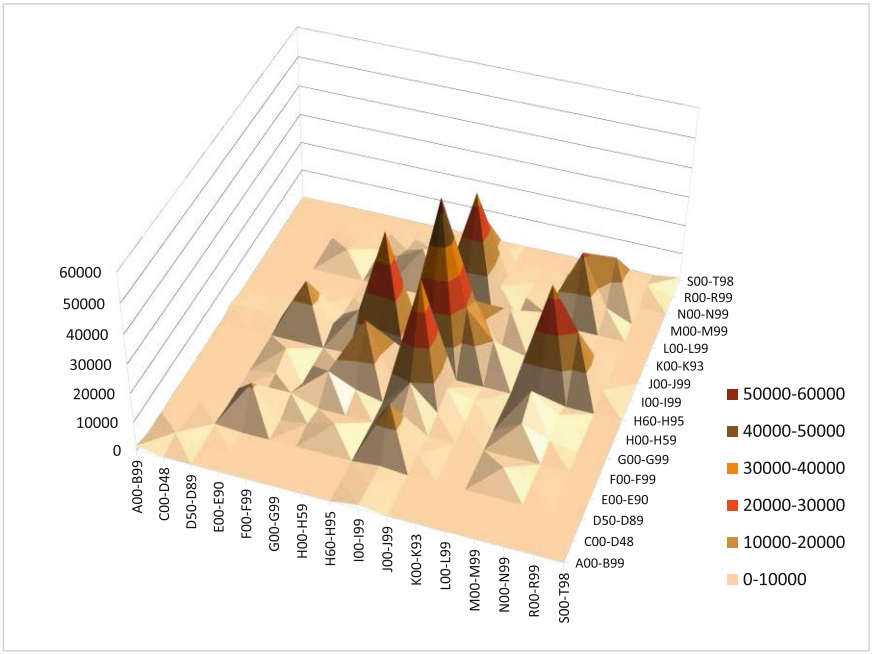

Fig. 7. Comorbidities for 2013-2014 collection of ORs grouped by classes in ICD-10

Now we need to explain why the diagnoses in the MFIs appear together. It is not surprising that the strongest top 3 MFIs in Tables 2 and 3 contain different diseases of the circulatory system, like Hypertensive diseases (I10-I15), Ischaemic heart diseases (I20-I25), Atrial fibrillation and flutter (I48), Cerebrovascular diseases (I60-I69), and other forms of heart disease (I30-I52). It is well known that diseases of the circulatory system are primary risk factors for DM2. These can be seen also as highest peaks in Fig. 7 which presents the comorbidities of different ICD-10 classes for 2013–2014 in ORs of patients with DM2 onset in 2015.

Further classes of diseases with higher frequency in the MFIs are shown in Fig. 7: Diseases of the eye and adnexa (H00-H59), Diseases of the musculoskeletal system and connective tissue (M00-M99), Diseases of the nervous system (G00-G99), Acute bronchitis (J20), and Gastritis and duodenitis (K29), all of them typical for prediabetes.

One unusual finding is the frequency of Malignant neoplasm of breast (C50) that was also identified as a maximal frequent itemset MFI#149 with a single diagnose only in all three collections. Figure 8 shows the demographic information for prevalence of

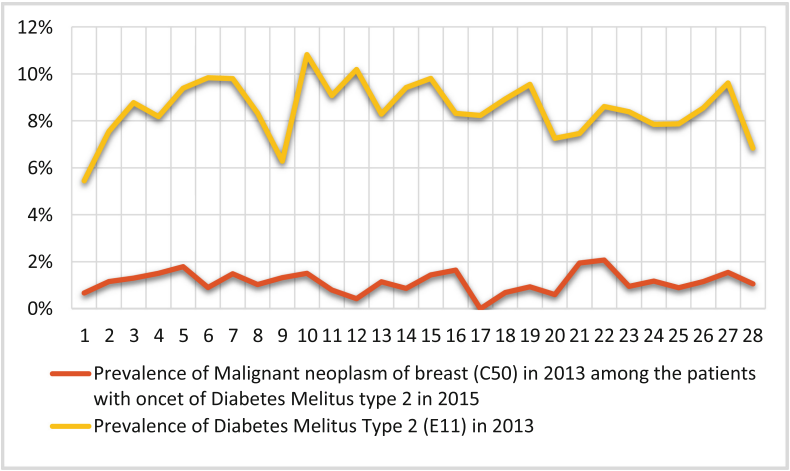

Fig. 8. Demographic data about the prevalence of C50 in prediabetes condition and E11

Maligna neoplasm of breast (C50) in prediabetes condition and DM2, with ICD-10 code E11, in 28 Bulgarian regions. There is a strong correlation of 0.93 between these two diagnoses except for two regions with ID#9 and ID#17. The latter finding is unexpected and needs further clarification; the Register shows that there are less registered diabetic patients in region #9 but this is insufficient to motivate the correlation shown in Fig. 8.

Figure 9 shows the distribution of patients in the support of "MFI#149" according to their age. The gender value in the support of "MFI#149" is female, with one exception for a male, for whom this diagnose is considered as a rare disease. The age values show that these are mainly female patients in menopause which is considered as a period with high risk for breast cancer. From the context information in the support of "MFI#149" for BMI and blood pressure we can also observe that most patients in this support set have higher risk of DM2 due to the presence of multiple risk factors as obesity (ICD-10 code I66) and hypertension (ICD-10 codes I10-I11).

Usually the association between Malignant neoplasm of breast (C50) and DM2 is studied in the opposite direction, considering the diabetes treatment as a risk factor for breast cancer [17]. However recently the association of breast cancer as a risk factor in prediabetes condition was in focus as well [18]. We note that in general the ICD-10 diagnose C50 is not considered risky for diabetes. But our algorithm reveals this unknown and latent interrelationship so it needs deeper analysis by medical experts.

Finally we briefly discuss the data quality issue and how we deal with it in our data mining approach. It is well known that missing data in medical documentation is inevitable. There are many patients for whom the available ORs contain no information about certain context attributes. Thus some attribute values are replaced by the value NA, which is considered as the most general value.

Data about HbA1c (glycated hemoglobin) are available only for 15% of patients, that is why we consider this attribute as a more general value ANY. But we note that

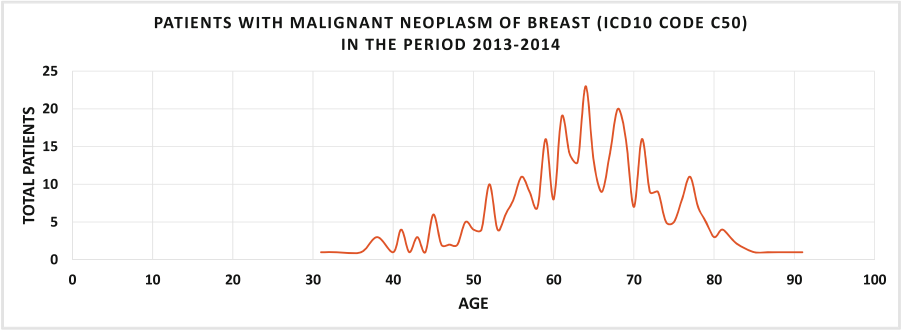

Fig. 9. Age of the patients in the support set of "MFI#149"

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 45% of these patient and for 30% of them the values were high.

5 Conclusion and Future Work

In this paper we present the national Diabetes Register, automatically generated using semi-structured patient records in Bulgarian language, and show how the stepwise integration of modern data processing technologies turn the Register to an environment for monitoring, prediction, issuing alerts, and discovery of specific risk factors. Application of automatic NLP in large scale is a real novelty in this area. Perhaps one of the most important achievement is the demonstration that reuse of available medical documentation leads to new quality when modern ICT is integrated as an enabling tool. The authors show this achievement to the national healthcare authorities whenever possible and officially propose to reuse existing clinical texts in the implementation of the Electronic Health Record (EHR) system in Bulgaria.

Future work includes further elaboration of specific algorithms that take into consideration temporal sequences of events. Developing more efficient knowledge discovery tools 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 socially-important chronic diseases. In this way it will become possible to identify the Bulgarian citizens who have predisposition to various serious diseases.

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

References

1. WHO Diabetes Fact Sheets, November 2017. http://www.who.int/mediacentre/factsheets/fs312/en/. Accessed 20 Jan 2018
2. WHO Global Report on Diabetes (2016). http://apps.who.int/iris/bitstream/10665/204871/1/9789241565257_eng.pdf?ua=1. Accessed 20 Jan 2018. ISBN 978 924 156525 7
3. Richardson, E., (ed.): National Diabetes Plans in Europe: what lessons are there for the prevention and control of chronic diseases in Europe? Policy Brief of the Joint Action on Chronic Diseases and Promoting Healthy Ageing across the Life Cycle, WHO Regional Office for Europe (2016). ISSN 1997-8065
4. Garrofé, B., Björnberg, A., Phang, A.Y.: Euro Diabetes Index 2014. Health Consumer Powerhouse Ltd., (2014). ISBN 978-91-980687-4-0
5. Boytcheva, S., Angelova, G., Angelov, Z., Tcharaktchiev, D.: Integrating Data Analysis Tools for Better Treatment of Diabetic Patients. In: Kalinichenko, L., Manolopoulos, Y., Skvortsov, N., Sukhomlin, V. (eds.) Selected Papers of the XIX International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2017), CEUR Workshop Proceedings, vol. 2022, pp. 230–237 (2017). http://ceur-ws.org/Vol-2022/. Accessed 20 Jan 2018
6. European Best Information through Regional Outcomes in Diabetes (EUBIROD) homepage. http://www.eubirod.eu/. Accessed 20 Jan 2018
7. Hallgren Elfgren, I.M., Törnvall, E., Grodzinsky, E.: The process of implementation of the diabetes register in primary health care. Int. J. Qual. Health Care **24**(4), 419–424 (2012)
8. Tcharaktchiev, D., Zacharieva, S., Angelova, G., Boytcheva, S., Angelov, Z., et al.: Building a bulgarian national registry of patients with diabetes mellitus. J. Soc. Med. **2**, 19–21 (2015). ISSN 1310-1757 (in Bulgarian Language)
9. Boytcheva, S., et al.: Obtaining status descriptions via automatic analysis of hospital patient records. Informatica **34**, 269–278 (2010)
10. Boytcheva, S., Angelova, G., Angelov, Z., Tcharaktchiev, D.: Text mining and big data analytics for retrospective analysis of clinical texts from outpatient care. Cybern. Inf. Technol. **15**(4), 58–77 (2015). https://doi.org/10.1515/cait-2015-0055
11. Boytcheva, S., Angelova, G., Angelov, Z., Tcharaktchiev, D.: Mining comorbidity patterns using retrospective analysis of big collection of outpatient records. Health Inf. Sci. Syst. **5**(1), 3 (2017). https://doi.org/10.1007/s13755-017-0024-y
12. Aggarwal, C., Bhuiyan, M., Hasan, M.: Frequent pattern mining algorithms: a survey. In: Aggarwal, C., Han, J. (eds.) Frequent pattern mining, pp. 19–64. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07821-2_2
13. Rabatel, J., Bringay, S., Poncelet, P.: Mining sequential patterns: a context-aware approach. In: Guillet, F., Pinaud, B., Venturini, G., Zighed, D. (eds.) Advances in Knowledge Discovery and Management, pp. 23–41. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-35855-5_2
14. Huang, J., Huan, J., Tropsha, A., Dang, J., Zhang, H., Xiong, M.: Semantics-driven frequent data pattern mining on electronic health records for effective adverse drug event monitoring. In: 2013 IEEE International Conference on Bioinformatics and Biomedicine BIBM, pp. 608–611. IEEE (2013). https://doi.org/10.1109/bibm.2013.6732567
15. Ziembiński, R.Z.: Accuracy of generalized context patterns in the context based sequential patterns mining. Control Cybern. **40**(3), 585–603 (2011). http://yadda.icm.edu.pl/baztech/element/bwmeta1.element.baztech-article-BATC-0009-0001/c/httpwww_bg_utp_edu_plartcc2011ziembinski.pdf. Accessed 20 Jan 2018

16. Yu, H.F., Hsieh, C.J., Chang, K.W., Lin, C.J.: Large linear classification when data cannot fit in memory. ACM Trans. Knowl. Discov. Data **5**(4), 23 (2012). https://doi.org/10.1145/2086737.2086743
17. Pan, X.F., He, M., Yu, C., Lv, J., Guo, Y., Bian, Z., et al.: Type 2 Diabetes and risk of incident cancer in China: a prospective study among 0.5 million Chinese adults. Am. J. Epidemiol., kwx376 (2018). https://doi.org/10.1093/aje/kwx376
18. Onitilo, A.A., Stankowski, R.V., Berg, R.L., Engel, J.M., Glurich, I., Williams, G.M., Doi, S.A.: Breast cancer incidence before and after diagnosis of type 2 diabetes mellitus in women: increased risk in the prediabetes phase. Eur. J. Cancer Prev. **23**(2), 76–83 (2014). https://doi.org/10.1097/CEJ.0b013e32836162aa