# Towards Integrated Analysis of Risk Factors and Diabetes Prevention using Big Data and Natural Language Processing

Dimitar Tcharaktchiev
Medical University Sofia,
University Specialised Hospital
for Active Treatment of Endocrinology
Sofia, Bulgaria
Email: dimitardt@gmail.com

Svetla Boytcheva
American University in Bulgaria
Blagoevgrad, Bulgaria
Email: svetla.boytcheva@gmail.com

Zhivko Angelov
Adiss Lab Ltd.
Sofia, Bulgaria
Email: angelov@adiss-bg.com

Ivelina Nikolova
Institute of Information
and Communication Technologies,
Bulgarian Academy of Sciences
Sofia, Bulgaria
Email: iva@lml.bas.bg

Galia Angelova
Institute of Information
and Communication Technologies,
Bulgarian Academy of Sciences
Sofia, Bulgaria
Email: galia@lml.bas.bg

***Abstract*—This paper demonstrates the potential of natural language processing for extracting patient-related data from very big repositories of semi-structured patient records. We mine 37.9 million outpatient records, extract risk factors and show how to integrate the findings in a system that enables effective diabetes prevention. The findings show the weak points in the organisation of primary care and specialised outpatient care in Bulgaria.**

***Keywords–Diabetes prevention; Discovery of risk factors; Automatic text analysis; Entity extraction; Outpatient record; Data integration, analytics.***

## I. INTRODUCTION

Automatic text analysis is applied to clinical narratives since decades. Many researchers work on this hot topic, developing Natural Language Processing (NLP) algorithms and experimental prototypes in various natural languages. Recently, the NLP tools reached such a level of maturity that one can think about their integration in real software systems, i.e., outside the academic settings, in order to improve knowledge discovery and support decision making. Usually only certain events are extracted from the analysed clinical texts since full "understanding" is hard to achieve; and in general NLP tools deliver a small percentage of erroneous results. But these shortcomings are somehow compensated by the NLP ability to process big repositories of clinical records [1]. The results of automatic information extraction from patient records are relatively good and can be compared with other systems that solve similar tasks, such as CLEF (Clinical E-Science Framework) that retrieves data for cancer patients [2] and AMBIT that retrieves information from biomedical text [3]. Another successful platform is the Mayo Clinic NLP System [4] for structured retrieval of patient records taking into consideration patients' smoking status.

In this way, NLP turns to be an integral component in the initiative for Secondary Use of Electronic Health Records (EHR) data and one of the few technologies that enable tackling of unstructured big data in medicine. In principle, its application is evaluated in terms or precision and recall; these indicators give hints about possible erroneous results.

One of the most important tasks is to extract patient status in structured format (Attribute–Value). The "attributes" can be for instance anatomical organs, major anatomical systems, their characteristics, physician examinations performed during the admission etc. The "values" describe the actual condition of the patient. Thus, structured representation of the patient status can be viewed as a set of "attribute-value" tuples. To solve this task NLP uses Support Vector Machines and related unsupervised and supervised approaches to machine learning. For example [5] presents classifiers based on supervised methods; [6] applies Maximum Entropy classifiers; and [7] uses semi-supervised relation extraction.

Analysing patient records in order to reveal the diabetic patient status or prevent diabetes is a particularly challenging task. NLP tools that extract entities from patient-related documentation can help to optimise the treatment, reduce its costs or deliver early alerts about potential diabetes by identification of high risk factors. For instance, the paper [8] presents prototypes that (*i*) analyse discharge summaries for evidence indicating a presence of diabetes, (*ii*) assess diabetes protocol compliance and (*iii*) identify high risk factors. The tools extract entities like assertion of the disease and associated findings in the text, numerical clinical data and prescribed medications. The classifier analyses reports for the presence and absence of diabetes and recognises the status with accuracy higher than 97%. The evaluation was done for 444 discharge summaries.

 ISBN: 978-1-61208-384-1 

The high risk factor classifier extracts the blood pressure values and cholesterol values; it works with accuracy exceeding 90%. The tools also estimate the record's adherence to quality of care protocols for indicators such as ABCS (i.e., A1c-glycosylated hemoglobin, Blood pressure, Cholesterol, Smoking). This classification is also done with accuracy exceeding 90%.

Here we present results of large-scale NLP applied to a big repository of outpatient records. Big Data pose special requirements to NLP, for instance one cannot inspect manually the corpus that is processed. Our objective in this study is to mine records submitted to the Bulgarian National Health Insurance Fund (NHIF) and, despite their primary accounting orientation, to discover patients with risk factors that are potentially related with development of diabetes. In Bulgaria, there are no established programs for prevention of socially important diseases. An automated system discovering the citizens with increased risk to develop diabetes represents an important contribution for the amelioration of public health system. Having in view all the patients' pathways, including reports of the General Practitioners and all frequently visited specialists, the proposed system can analyse more risk factors than one single doctor, who has no access to all the data obtained in various clinical examinations and reported to NHIF by different specialists.

This paper is organised as follows: Section 2 describes the project objectives and sketches the previously developed modules, which are improved and integrated in the current system. Section 3 discusses data formats and overviews the repository. The results obtained in the present study are reported in Section 4. Sections 5 and 6 contain a brief discussion, the conclusion and further work.

## II. Project Context and Available Tools

The ultimate objective of our project is to accelerate the construction of the Register of diabetic patients in Bulgaria by integration of language technologies and business intelligence tools [9]. Today the growing administrative burdens and multiple registrations are considered as a major obstacle for the development of the Register. However, advanced information technologies would enable to: (*i*) keep the established practice of patient registration without overloading the medical experts with additional paper work; (*ii*) reuse the existing standard records in compliance with all legal requirements for safety and data protection; (*iii*) save time and resources by avoiding multiple patient registrations and disturbance of the diagnostic and treatment process. Practically, once entered in the healthcare system, the patient data might be reused in multiple aspects. A web-interface for self-registration to the Bulgarian Diabetic Register is foreseen as well.

The Register contains 28 indicators of diabetic patients' status, including age, sex, ICD-10 codes of diagnoses of diabetes and its complications, diabetes duration, risk factors, data about compensation, laboratory results, hospitalisations and prescribed medication. Manual collection of data proved to be impractical during the last ten years; in addition there are many diabetic patients who are not formally diagnosed and not treated at all. In the case of diabetes, a progressive chronic disease with serious complications, it is highly desirable to develop a system for early alerts that might signal eventual diabetes symptoms.

It turns out that all the information needed for the Register is available in the outpatient records, collected by the Bulgarian NHIF. There are multiple records stored for the same patient along the months and the years. Thanks to the support of the Bulgarian Ministry of Health and the NHIF, the Medical University - Sofia has received for research purposes a large collection of outpatient records. The data repository currently contains more than 37.9 million pseudonymised reimbursement requests (outpatient records) submitted to the NHIF in 2013 for more than 5 million patients, including 436,000 diabetic ones. In Bulgaria the outpatient records are produced by the General Practitioners (GPs) and the Specialists from Ambulatory Care for every contact with the patient.

We have previous experience in automatic analysis of diabetic patients' discharge letters in Bulgarian language. In 2010-2011 a drug extractor has been developed, based on algorithms using regular expressions to describe linguistic patterns [10]. There are more than 80 different patterns for matching text units which deal with the ATC and NHIF drug codes, medication name, dosage and frequency. Currently, the extractor is elaborated and handles 2,239 drugs names included in the NHIF nomenclatures. Recent extraction evaluation has been performed with large-scale analysis of the outpatient records of 33,641 diabetic patients for 2013. The precision is 95.2% and the recall - 93.7%. This result is slightly better than the accuracy reported in 2011 [10] when the extractor was a (research) prototype dealing with less than 500 drugs. The performance of the module is evaluated manually. The labelled data is split to 20 equal subsets and randomly selected records are evaluated by an expert (about 40% of each subset). The average of the subset evaluation is the final score of the module.

The major reasons for incorrect recognition of drug events are: (*i*) misspelling of drug names; (*ii*) drug names occurring in the contexts of other descriptions; (*iii*) undetected descriptions of drug allergies, sensibility, intolerance and side effects; (*iv*) drug treatment described by (exclusive) OR; (*v*) negations and temporally interconnected events of various kinds: undetected descriptions of cancelled medication events; of changes or replacements in therapy; of insufficient treatment effect and change of therapy. About 30% of the medication events in the test corpus were described without any dosage. Lack of explicit descriptions occurs mostly for treatment of accompanying diseases. After applying the recognition algorithm and default daily dosage, the number of records lacking dosage has been reduced to 15.7% in the final result.

Another extractor that has been developed in 2010-2011 identifies values of clinical tests and lab data in the free text [11]. In 2011, the extractor recognised more than 90 types of laboratory tests, some of them with accuracy higher than 98%. The evaluation was done on 6,200 discharge letters of diabetic patients. Today the extractor is extended to cope with the clinical test values in the NHIF repository of outpatient records. In particular, our current study is focused on the indicators: age, body mass index (BMI), waist circumference, triglycerides, cholesterol, HDL-cholesterol, blood glucose on fasting, and blood glucose on the 120-th minute of the glucose tolerance test.

Moreover, we invested efforts in automatic recognition of temporal markers in discharge letters. In 2011, we proposed an algorithm and tool that recognised drugs taken by the patient at

 ISBN: 978-1-61208-384-1 

the moment of hospitalisation (day 0) [12]. This tool analysed the Case History section of hospital discharge letters. In 2012, we did a more systematic study of the temporal information in diabetic patients' discharge letters and proposed an algorithm for splitting the case history into episodes [13]. The temporal markers, which refer to the absolute or relative moments of time, are identified with precision 87% and recall 68%. The direction of time for the episode events: *backwards* or *forward* (with respect to certain moment orienting the episode) is recognised with precision 74.4%.

To tackle the Repository of outpatient records, we use a Business Intelligence tool (BITool) that processes the database of extracted entities. In principle the BITool can deliver various types of findings to decision makers in order to improve the public health policy and the management of Bulgarian healthcare system. The tool is useful anyway because the Health Insurance Fund data contains a lot of information that is structured using codes of medical classifications and nomenclatures. However, we are interested in the analysis of free text sections and capturing some essential entities described there. By means of NLP techniques integrated with the BITool we discover the potential diabetic patients, which were not formally diagnosed with diabetes. The paper [9] presents the study and more especially, how we classify with precision 91.5% the records according to the hypothesis "*having diabetes*" using only text comments in the sections with unstructured content. The experiment was run on 1,206,276 records of 156,000 patients who are not formally diagnosed with diabetes but the word "diabetes" (in Bulgarian "диабет") occurs in their records. In total, there are 190,189 such records for 156,310 patients in our dataset.

## III. MATERIAL

The outpatient records are semi-structured files with predefined XML-format. Despite their primary accounting purpose they contain sufficient text explanations to summarise the case and to motivate the requested reimbursement. The most important indicators like Age, Gender, Location, Diagnoses are easily seen since they are stored with explicit tags. The Case history is presented quite briefly in the Anamnesis as free text with description of previous treatments, including drugs taken by the patient beyond the ones that are to be reimbursed by the Insurance Fund. Family history and Risk factors are often included in the Anamnesis. Patient status is another section containing free text. It includes a summary of the patient state, symptoms, syndromes, patients' height and weight, body mass index (BMI), blood pressure and other clinical descriptions. The values of Clinical tests and lab data are enumerated in arbitrary order as free text in another section. A special section is dedicated to the Prescribed treatment. Only drugs prescribed by the GPs and reimbursed by the NHIF are coded, using the specific NHIF nomenclatures. All the other medications and treatment procedures are described as free text. In contrast to clinical discharge letters that might discuss treatments in longer past and future periods, the Prescribed treatment section in the outpatient records is more focused on the context at the moment when the record is composed.

The repository given to the Medical University - Sofia is pseudonymised by NHIF which has the keys for mapping the records to the original patients. Our experiments use a completely anonymised data set. Fortunately, the pseudonymised patient identifier helps to track automatically the multiple visits of the same patient to GPs and Ambulatory Care, which is important for a chronicle disease like diabetes.

An outpatient record includes up to 160 tags. The average length of the files is about 1MB. Here, we work with 20-30 tags and consider the unstructured content of four sections.

Our findings are obtained after the extraction of the patient age and seven risk factors. Our source repository for the present study consists of 1,206,276 records of 156,310 patients who are not formally diagnosed with diabetes but the word 'diabetes' occurs in at least one of their outpatient records.

## IV. RESULTS AND FINDINGS

We studied the inter-dependences of the age (40+ years), the seven risk factors, drugs taken by the patients and phrases suggesting diabetes in the free text of the outpatient records.

### A. Extraction of Lab Test Values

At first we extracted automatically the values of all tests related to the risk indicators enumerated on Table I. The total number of patients having two and more risk factors (being older than 40 years is one of the risk factors) is 68,681. The records of these patients are mined for extracting the values of BMI, waist circumference, triglycerides, cholesterol and the blood glucose tests. The number of patients having the different risk factors is given on Table I. We consider only the records of patients who have at least one risk factor.

TABLE I. NUMBER OF PATIENTS WITH DIFFERENT RISK FACTORS.

| Age $>$ 40 years: 68,681 patients with Risk Factors | |
|---|---|
| Indicators | Number |
| Body mass index (BMI) a | 49483 |
| Waist circumference | 47921 |
| Triglycerides | 6834 |
| Cholesterol | 9063 |
| HDL-cholesterol | 1226 |
| Fasting blood glucose | 8020 |
| Combination of 2 factors | 27773 |
| Combination of 3 factors | 31009 |
| Combination of 4 factors | 7173 |
| Combination of 5 factors | 2271 |
| Combination of 6 factors | 437 |
| Combination of 7 factors | 18 |

When the results of fasting blood glucose are uncertain the diagnostic of diabetes should be confirmed by the oral glucose tolerance test (OGGT). The fasting blood glucose for 8,020 patients, aged 40+, exceeded 6,1 mmol/l and the diabetes was not registered in their records. So, we searched for patients having more than 2 risk factors who have accomplished a OGGT. We discovered that only 1,103 OGTTs were executed, in 687 cases the patient age exceeded 40 years. In 102 cases the diagnostic of diabetes was confirmed. We note that 687 cases represent only 8.56% of the 8,020 patients with risk factors.

### B. Analytics on the Outpatient Records' Database

A key finding that is easily seen in the NHIF repository using our Business Intelligence tool is the heterogeneous source of the submitted outpatient records. An outpatient record is created for each visit to the GP or to Specialists from Ambulatory Care, in doctor's office or patient home. Figure 1 summarises the number of visits of the "risky" patients to 24 types of medical experts in the primary and the specialised

outpatient care. Column 2 at Figure 1 means that 10,878 citizens had only one visit to the GPs (row 1), 3,217 citizens had a single visit to Gynecologists (row 2), 313 citizens visited an Allergist only once (row 3) and so on. Column 3 presents the number of citizens who visited the corresponding medical experts 2-5 times; Column 6 - the number of citizens who visited the respective medical specialists 16-42 times.

Figure 1 shows that the citizens aged 40+ visited often their GPs (38.72% of all visits) but also had consultations with other specialists. In Bulgaria one specialist can send a patient to another specialist without obligation to inform the GP about this. The clinical information systems of the GPs and the specialists cannot exchange any information among them and this is not required by the health authorities. The only obligation of the specialists is to provide information in xml-format to NHIF. Therefore the GPs, in general, have no access to all the information concerning the patient. In this way, the collection of all relevant documents in the GP's archives depends only on the good will of the patient, who needs to bring the documentation to his/her GP as paper copy (and therefore, the GP has to store a paper archive). Thus, the GPs have a rather partial view to the patient status.

| | Visits_ | | | | | |
|---|---|---|---|---|---|---|
| Specialist | 1 | 2-5 | 6-10 | 11-15 | 16+ | Total |
| 00 GP | 10878 | 26828 | 17429 | 8635 | 2929 | 66699 |
| 01 Gynecologist | 3217 | 2789 | 92 | 10 | | 6108 |
| 02 Allergist | 313 | 362 | 10 | | | 685 |
| 03 Gastroenterologist | 1858 | 1554 | 24 | 3 | | 3439 |
| 04 Dermatologist | 1931 | 1686 | 26 | 3 | | 3646 |
| 05 Endocrinologist | 2440 | 3567 | 70 | 1 | | 6078 |
| 06 Internist | 1022 | 710 | 57 | 9 | | 1798 |
| 07 Infectionist | 92 | 56 | 1 | | | 149 |
| 08 Cardiologist | 10526 | 11353 | 419 | 8 | | 22306 |
| 10 Neurologist | 5016 | 5802 | 272 | 20 | 6 | 11116 |
| 11 Nephrologist | 869 | 1086 | 22 | | | 1977 |
| 12 Oncologist | 187 | 71 | | | | 258 |
| 14 Otolaryngologist | 3041 | 2239 | 20 | 3 | | 5303 |
| 15 Ophtalmologist | 9949 | 5241 | 111 | 4 | | 15305 |
| 16 Parasitologist | 8 | 10 | | | | 18 |
| 18 Psychiatrist | 846 | 1108 | 145 | 11 | | 2110 |
| 19 Pulmonologist | 1658 | 1823 | 18 | | | 3499 |
| 20 Rheumatologist | 718 | 596 | 11 | | | 1325 |
| 22 Urologist | 1886 | 1631 | 46 | 4 | | 3567 |
| 23 Physiotherapist | 9 | 3059 | 121 | 4 | 3 | 3196 |
| 24 Hematologist | 289 | 286 | 15 | 2 | | 592 |
| 25 Surgeon | 2771 | 2219 | 98 | 12 | 4 | 5104 |
| 26 Anaesthetist | 60 | 2 | | | | 62 |
| 29 Neurosurgeon | 118 | 61 | 1 | | | 180 |
| Total | 59702 | 74139 | 19008 | 8729 | 2942 | 164520 |

Figure 1. Number of visits (1, 2-5, 6-10, 11-15, 16-42) of citizens to specialists

## C. Extraction of Drugs

We have extracted the drugs from the outpatient records that are considered in the present experiment. Please note that the records contain a variety of drug presentation formats: from free narrative in the text to fully structured information in XML, from partial to complete details, in Bulgarian language only or as a mixture of English and Bulgarian language. To illustrate the varieties we present here three examples:

1) Drug names in Cyrillic immediately followed by the dosage daily scheme:
НовоРапид 20/19/ +15+12+18 и Левемир 10+22 E
(in English: NovoRapid 20/19/ +15+12+18 and Levemir 10+22 E)
2) Drug name in English language with information about the NHIF registration code for this drug - AF433, and description in Bulgarian language about the dosage 20E and the scheme:
АФ433 Левемир (300 UI x10) - доза: 20E в 22 часа, подкожно
(in English: AF433 Levemir Penfill (300 UI x10) - dosage: 20E at 22 o'clock, subcutaneously)
3) Drug information about Levemir Penfill structured in XML format with NHIF registration code:

```
<DrugCode>AF433</DrugCode>
<DrugICD10>E10.9</DrugICD10>
<Quantity>4</Quantity>
<Day>30</Day>
```

Some 30,486 patients (out of the 68,681 patients we deal with) admit drugs. Drug names occur in 117,798 outpatient records. The 7-digit ATC codes, identified in the outpatient records, are 306. The identified 3-digit ATC codes are 47. The drug trade names are 356. Having in mind the risks for diabetes developments, we have analysed the number of patients that use glucocorticoides (H02), hydrochlorothiazide (C03AA03) and combinations of thiazides with other diuretics (C03EA01-hydrochlorothiazide and potassium-sparing agents).

Analysing the prescription of thiazides and their combinations at Figure 2, we notice that the percentage of patients taking thiazides (2%) is higher than the number of outpatient records containing thiazides (1%). This means that in general, the prescription of thiazides is not multiple so the medication is taken in short periods only. It can be assumed that the doctors are careful when prescribing thiazides to patients with risk factor for diabetes.

In Figure 3, we present the inter-dependencies of the risk factors and the number of patients aged 40+ who use glucocorticoids and thiazides. We note that the C03AA03 hydrochlorothiazide is most often prescribed to patients with high BMI and high waist circumference. The combination of hydrochlorothiazide and potassium-sparing agents C03EA01 are also often prescribed for these patients. Figure 3 supports the finding in Figure 2 that thiazides are prescribed relatively rare.

## D. Mining Free Text for Opinions of Medical Experts

In addition to the extraction of Lab data and Drugs, we process the doctors' utterances to discover phrases that potentially signal risk factors. To confirm/reject the hypotheses of (*i*) **having diabetes** and (*ii*) **family heredity**, we apply a hybrid approach of rough rule-based pre-filtering followed by training of machine learning models. For testing both hypotheses we use text chunks extracted from a concordancer around the string "диабет" (diabetes). The data set consists of 67,904 distinct chunks extracted from the records of 156,310 patients who are not formally diagnosed with diabetes. Each chunk contains the string "диабет" (diabetes) and a 6-token window of its left and right context. The text is only tokenised (i.e., split to words and meaningful strings) and stemmed (word endings are deleted). Figure 4 shows 2 examples from the manually

 ISBN: 978-1-61208-384-1 

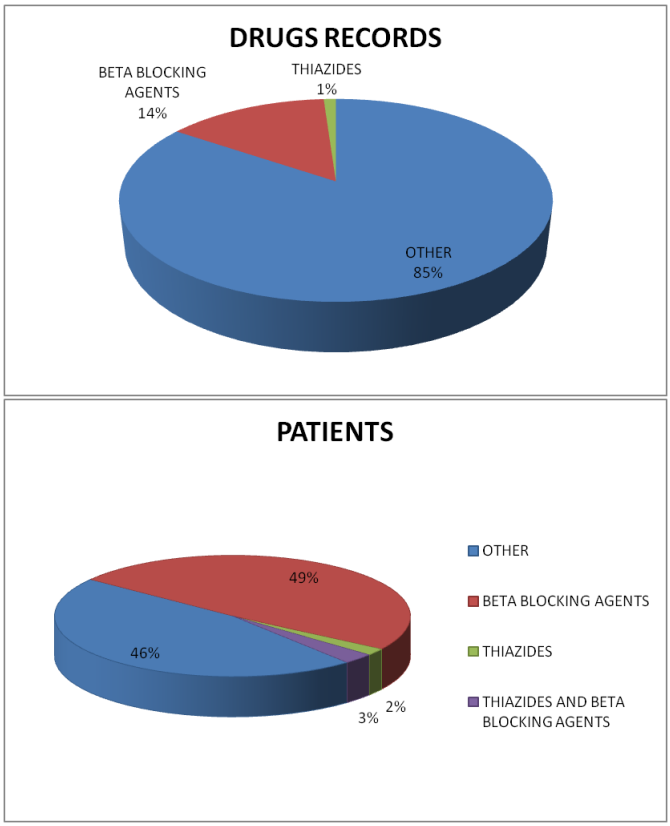


Figure 2. Statistics about using Thiazides, Beta Blocking Agents and other drugs

| Risk | Atc C03AA03 | C03EA01 | H02 | Total |
|---|---|---|---|---|
| Body mass index | 980 | 447 | 122 | 1549 |
| Cholesterol | 191 | 73 | 28 | 292 |
| Fasting blood glucose | 208 | 101 | 21 | 330 |
| HDL cholesterol | 20 | 8 | 1 | 29 |
| Triglycerides | 160 | 65 | 14 | 239 |
| Waist circuference | 946 | 444 | 139 | 1529 |
| Total | 2505 | 1138 | 325 | 3968 |

| Conbination Of | Atc C03AA03 | C03EA01 | H02 | Total |
|---|---|---|---|---|
| 2 risk factors | 468 | 226 | 86 | 780 |
| 3 risk factors | 604 | 290 | 85 | 979 |
| 4 risk factors | 193 | 74 | 15 | 282 |
| 5 risk factors | 57 | 20 | 5 | 82 |
| 6 risk factors | 5 | 6 | 1 | 12 |
| Total | 1327 | 616 | 192 | 2135 |

Figure 3. Drug Prescription to Patients Aged 40+ with Risk Factors

(*i*) NEG-diabetes; POS-familyHeredity
Фамилност - обременен/а - диабетици по майчина линия.
Family - heredity - diabetic on maternal line.

(*iii*) NEG-diabetes; NEG-familyHeredity
Липсва фамилна обремененост за захарен диабет.
No family heredity for diabetes mellitus.

Figure 4. NEGative and POSitive chunks examples from free text descriptions

annotated training data prepared in order to train the classifier. The last example is negative to both hypotheses whereas the first one is positive to family heredity and negative to having diabetes. These chunks come from records of patients who are not formally diagnosed with diabetes so, in principle there are more negative examples to our hypotheses.

*1) Rule-based rough pre-filtering:* In first place we do rough pre-filtering of the data in order to reduce its size. The expressions we used for filtering were manually selected by analysing the data, they are such as ”no evidence about diabetes”, ”no diabetes in the family” etc., and for the second experiment ”no family heredity”, ”no heredity” etc. After applying the rules in the first experiment the data reduced to almost 1/3 of its initial size while in the second case about 7% were reduced.

*2) Supervised classification of positive/negative examples:* In the second phase we apply a number of machine learning techniques on the reduced datasets. We create datasets of randomly extracted records from the full set and manually annotate them. The dataset used for testing the hypothesis “*having diabetes*” has two subsets – one of 282 documents and one of 1,000 documents. The first one is our development set, it is used for selecting features and for initial tests. It contains 74 positive and 208 negative examples whereas the second one (our test set) contains 187 positive and 813 negative examples respectively. By using various features and classification algorithms we check the applicability of machine learning to the automatic extraction of records referring to ”*having diabetes*” (similarly to [4], [12], [14]) and set a reasonable baseline for this task. We tried several algorithms on the same dataset: NaiveBayes, J48, SMO and JRip and MaxEnt, all with boolean features. JRip, J48 and MaxEnt performed best and MaxEnt algorithm outperformed all. We measure the performance of the models in terms of $precision$ (percentage of true positive examples in all extracted examples), $recall$ (percentage of the true positive examples in all available positive examples) and their harmonic mean $f - measure$.

The features we used were - words’ stems, bigrams and trigrams. The classification with MaxEnt reached 91.5% $precision$ on the positive examples and 88.55% on average (positive/negative). The $recall$ was comparatively low (52.1% on average) but for us precision is of major importance here because we want to select potential diabetic patients with high certainty. Some 37-42 phrases (depending on the selected model) were extracted as positive examples out of 1000 randomly selected test documents with $precision$ higher than 91%. This suggests that several thousands of positive assertions would be found in the free texts of the original data set. We consider this number significant given that the patients we deal with were not diagnosed with diabetes.

The dataset for classification of records according to the hypothesis “*family heredity*” has the following subsets: development subset - 600 documents (300 positive and 300 negative examples) and test set with 1,727 documents (915 positive and 812 negative examples). As features we use the words stems and bigrams and trigrams available in the development set (including punctuation). We use the same classification algorithms as in the first experiment, with boolean vectors. The best performance was achieved with MaxEnt algorithm using all features without feature selection - 93.8% $f - measure$. This means that out of 600 records, 276 were selected as

 ISBN: 978-1-61208-384-1 

approving the risk factor "*family heredity*" with precision over 95% (we keep in mind that the class distribution in the development and test sets may differ from the distribution in the original data). For comparison in [15] is achieved 100% $f - measure$ on extracting the "*experiencer*" feature of an event (whether the event is experienced by the patient or by other person). The experiment which is closest to ours is done on discharge letters and the authors report that there was insufficient data for thorough testing of the "*experiencer*" feature extraction.

## V. Discussion

The results presented here clearly confirm the lack of prevention-orientated thinking of the general practitioners and the specialists. The organisation of primary care and the specialised outpatient care in Bulgaria do not stimulate the doctors to do the prevention. The preventive measures are not systematic and are not adequately reimbursed. In addition the specialists have no access to the complete adequate information concerning all Lab data, clinical examinations and consultations made by the patients.

## VI. Conclusion

Information extraction from clinical texts matures only recently but its performance gradually improves and often exceeds 90% [16]. Our experience shows that in a rapid development process, one can achieve good performance in separate extraction tasks within 2-3 years. The review [16] however states that "current applications are rarely applied outside of the laboratories they have been developed in, mostly because of scalability and generalisability issues". We would add here that the negative results are partly due to the inconsistency, incompleteness and fragmentariness of the medical documentation per se; these shortcomings become obvious in the computer age when ambitious goals like the Secondary Use of EHR data are set.

The erroneous results that might include also overgeneration (false positive indications) are an inevitable part of the NLP technologies. They might be dangerous for further use unless the applications are based on very large resources. In these cases the small percentage of false positive entities is statistically insignificant and practically negligible. Human recognition of entities might also include some erroneous choices. We test carefully our extractors and integrate them only in scenarios where their role is to deliver primarily supporting evidence.

From the social medicine perspective the present system is an important achievement. It enables to develop a diabetes prevention procedure after the analysis of the risk factors and delivering the codes of the "risky patients" to the National Health Insurance Fund. In this way the National authorities can send alerts to the GPs and also to the patients, informing them electronically about risk factors and the necessity to implement active prophylactic measures.

## Acknowledgment

The research work presented in this paper is partially supported by the FP7 grant 316087 AComIn "Advanced Computing for Innovation", funded by the European Commission in the FP7 Capacity Programme in 2012–2016, the project BG161PO003-1.1.06-0023-C0001 "Analysing and identifying dependencies in big data repositories - application for economic and technological analyses" funded by the Competitiveness Operational Programme in 2012–2015 and the project D01-192/2014 funded by the Bulgarian Ministry of Education and Science. The authors also acknowledge the support of the Bulgarian Health Insurance Fund, the Bulgarian Ministry of Health and the Medical University - Sofia.

## References

[1] D. Demner-Fushman, W. W. Chapman, and C. J. McDonald, "Methodological Review: What can natural language processing do for clinical decision support?" Journal of Biomedical Informatics, vol. 42(5), 2009, pp. 760–772.

[2] H. Harkema, A. Setzer, R. Gaizauskas, M. Hepple, R. Power, and J. Rogers, "Mining and modelling temporal clinical data," in Proceedings of the 4th UK e-Science All Hands Meeting, S. Cox, Ed., Nottingham, UK, 2005, pp. 507–514.

[3] R. Gaizauskas, M. Hepple, N. Davis, Y. Guo, H. Harkema, A. Roberts, and I. Roberts, "AMBIT: Acquiring Medical and Biological Information from Text," in S.J. Cox (Ed.) Proceedings of the 2nd UK e-Science All Hands Meeting, 2003, pp. 370–373.

[4] G. Savova, P. Ogren, P. Duffy, J. Buntrock, and C. Chute, "Mayo Clinic NLP System for Patient Smoking Status Identification," Journal of American Medical Informatics Association, vol. 15(1), 2008, pp. 25–28.

[5] M. Poesio and A. Almuhareb, "Identifying concept attributes using a classifier," in Proceedings of the ACL Workshop on Deep Lexical Semantics, 2005, pp. 18–27.

[6] B. Kedar, P. P. Talukdar, G. Kumaran, O. Pereira, M. Liberman, A. Mccallum, and M. Dredze, "Lightly Supervised Attribute Extraction," in The SelectedWorks of Andrew McCallum.

[7] Y. W. Wong and K. N. D. Widdows, T. Lokovic, "Scalable Attribute-Value Extraction from Semi-Structured Text," in Proceedings of ICDM Workshop on Large-scale Data Mining: Theory and Applications, 2009.

[8] N. K. Mishra, R. Y. Son, and J. J. Arnzen, "Towards Automatic Diabetes Case Detection and ABCS Protocol Compliance Assessment," Clin. Med. Res., vol. 10(3), August 2012, pp. 106–121.

[9] I. Nikolova, D. Tcharaktchiev, S. Boytcheva, Z. Angelov, and G. Angelova, "Applying Language Technologies on Healthcare Patient records for Better Treatment of Diabetic Patients," in Proceedings of AIMSA 2014. Springer, Lecture Notes in Computer Science Vol. 8722, 2014, pp. 92–103.

[10] S. Boytcheva, "Shallow Medication Extraction from Hospital Patient Records," in Studies in Health Technology and Informatics series. IOS Press, 2011, pp. 119–128, Koutkias, V., J. Nies, S. Jensen, N. Maglaveras, and R. Beuscart (Eds.).

[11] D. Tcharaktchiev, G. Angelova, S. Boytcheva, Z. Angelov, and S. Zacharieva, "Completion of Structured Patient Descriptions by Semantic Mining," in Studies in Health Technology and Informatics series. IOS Press, 2011, pp. 260–269, Koutkias, V., J. Nies, S. Jensen, N. Maglaveras, and R. Beuscart (Eds.).

[12] S. Boytcheva, D. Tcharaktchiev, and G. Angelova, "Contextualization in automatic extraction of drugs from Hospital Patient Records," in Studies in Health Technology and Informatics series. IOS Press, 2011, pp. 527–531, a. Moen at al. (Eds).

[13] S. Boytcheva, G. Angelova, and I. Nikolova, "Automatic Analysis of Patient History Episodes in Bulgarian Hospital Discharge Letters," in Proceedings of the Demonstrations at EACL 2012, 2012, pp. 77–81.

[14] I. Nikolova, "Unified Extraction of Health Condition Descriptions," in Proceedings of the North America ACL HLT 2012 Student Research Workshop, 2012, pp. 23–28.

[15] H. Harkema, J. Dowling, T. Thornblade, and W. Chapman, "ConText: An Algorithm for Determining Negation, Experiencer, and Temporal Status from Clinical Reports," J Biomed Inf, vol. 42(5), 2009, pp. 839–851.

[16] S. Meystre, G. Savova, K. Kipper-Schuler, and J. F. Hurdle, "Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research," in Yearbook of Medical Informatics, 2008, pp. 128–144.

 ISBN: 978-1-61208-384-1 