BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **15**, No 4

Sofia • 2015

Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.1515/cait-2015-0055

Text Mining and Big Data Analytics for Retrospective Analysis of Clinical Texts from Outpatient Care

Svetla Boytcheva[1], *Galia Angelova*[1], *Zhivko Angelov*[2], *Dimitar Tcharaktchiev*[3]

[1]*Institute of Information and Communication Technologies, BAS, Sofia, Bulgaria*
[2]*Adiss Lab Ltd., Sofia, Bulgaria*
[3]*Medical University Sofia, University Specialised Hospital for Active Treatment of Endocrinology, Sofia, Bulgaria*
Emails: *svetla.boytcheva@gmail.com* *galia@lml.bas.bg* *angelov@adiss-bg.com* *dimitardt@gmail.com*

Abstract: *This paper presents the results of an on-going research project for knowledge extraction from large corpora of clinical narratives in Bulgarian language, approximately 100 million of outpatient care notes. Entities with numerical values are mined in the free text and the extracted information is stored in a structured format. The Algorithms for retrospective analyses and big data analytics are applied for studying the treatment and evaluating the diabetes compensation and control of arterial blood pressure.*

Keywords: *Medical Informatics, Big Data, Text Mining, Information Extraction, Natural Language Processing.*

1. Introduction

During the last decade several Information Extraction (IE) system for analysis of clinical texts were developed – for diagnosis extraction, drugs and dosage identification, recognition of complaints and related events, risk factors, etc., [1]. Despite the achievements in this area these systems are difficult to (re)use because most of them, including the associated linguistic resources, are specific language (mainly for English language) and cannot be easily adapted for other languages. Moreover, they are developed either as academic research projects or as commercial software. Usually their results are evaluated on annotated corpora manually tuned to specific tasks, so that the performance assessment is difficult as well.

Our project aims at the design of methods and technologies for automatic extraction and structuring of patient-related entities from a Big Data collection of outpatient care notes, which is a quite challenging task. We need to scale the existing methods designed for "toy" data to process approximately 100 million patient records provided from the General Practitioners (GP) and specialists to the Bulgarian National Health Insurance Fund (NHIF). This motivates us to start developing a system especially designed for processing outpatient records in Bulgarian language that is able to cope with noisy and missing data still providing reliable results.

The outpatient records describe dozens of important examinations: values of indicators like Body Mass Index (BMI), Weight (W), Blood Pressure (BP), glycated hemoglobin (HbA1c), Blood Glucose (GLU), etc.; the latter are among the key risk factors for development of cardio-vascular diseases (high BP) and diabetes mellitus (levels of HbA1c and GLU). However, these values occur in the free text, presented in a huge variety of formats, so their automatic identification is a challenging task. Extraction of such data with high confidence is essential for retrospective analyses in the secondary use of clinical texts. In our project we need to extract the current patient status, including BP and other numerical values, in order to monitor hypertension and assess diabetes treatment and risks. We propose a hybrid method for automatic rules generation for IE from clinical data. The experiments are made and evaluated over approximately 9.5 million of outpatient records.

This paper is structured as follows. Section 2 overviews related research. Section 3 briefly describes the available tools and materials we use. Section 4 presents the proposed methods for text mining including pre-processing of large collections of clinical narratives. Section 5 reports the experiments, the evaluation results and applications built for the assessment of diabetes compensation and control of arterial BP. Section 6 concludes with discussion of some scalability issues and briefly sketches further work.

2. Related work

Recently IE from clinical texts is a hot research topic [1-4]. The annotation tasks related to competitions like i2b2 and SemEval [5] consolidate the efforts for the development of adequate linguistic resources. Research progresses also for some low-resourced languages, e.g., [6] presents extraction of the entities for Disorder, Finding, Pharmaceutical Drug and Body Structure from Swedish health records.

Z h o u et al. [7] describe the MEDical IE (MedIE) system that mines free-text clinical records of patients with breast complaints. They propose three techniques to solve different IE tasks: a graph-based approach which uses the parsing results of a link-grammar parser for identification of relations; an ontology-based approach to extract the medical terms of interest; and a NLP-based feature extraction method coupled with an ID3-based decision tree to perform text classification.

V o o r h a m and D e n i g [8] present a character sequence algorithm for recognition of measurement labels for systolic and diastolic BP, weight, height, serum glucose, glycated hemoglobin (HbA1c), as well as several measures of

serum, triglycerides, and serum creatinine.

Turchin et al. [9] use regular expressions for extraction of BP values documented in the text and identification of anti-hypertensive medication modification. Murtaugh et al. [10] apply a similar approach for bodyweight values extraction from clinical texts.

Jiang et al. [11] apply machine learning for extracting clinical entities including medical problems, tests, and treatments, as well as their status from hospital discharge summaries.

Patrick et al. [12] use a cascade method in a pipeline system that performs different tasks for extractions and classifications of clinical data: concept annotation, assertion classification and relation classification.

Bigeard et al. [13] propose CRF supervised categorisation for detection of segments (themes, numerical sequences and units) and a rules-based approach for associating these segments in order to build semantically meaningful sequences.

3. Project context, available tools and materials

The ultimate objective of our project is to extract healthcare information about patients from medical documents in a semi-structured format and to evaluate the quality of patient treatment using language technologies and business intelligence tools. The initial processing starts from the raw XML-data that is transformed to structured representations of extracted entities – in databases and comma separated format, using a Business Intelligence Tool (BITool) together with some Natural Language Processing (NLP) tools (Fig. 1). Further analytics can deliver various types of findings to decision makers in order to improve the public health policy and the management of Bulgarian healthcare system [14].

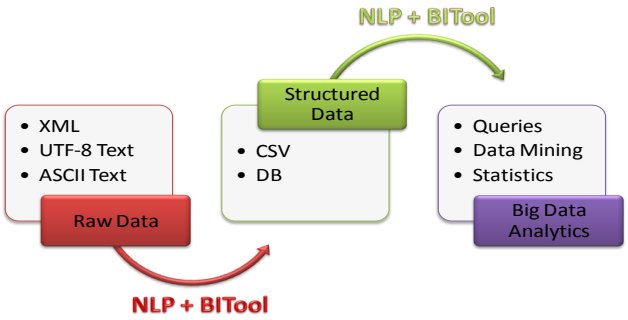

Fig. 1. Repository of outpatient records

We deal with a repository of anonymous outpatient records provided by the Bulgarian NHIF in XML format. Most data needed for health management are structured using nomenclatures, such as ICD [15] but the records also contain paragraphs of unstructured text in the fields “Anamnesis”, “Status”, “Examine”, etc.

The text style is telegraphic usually with no punctuation and a lot of noise (some words are concatenated; there are many typos, syntax errors, etc.). Due to the lack of resources and annotated corpora we cannot use dictionaries and traditional

methods for text analysis. All texts are in Bulgarian but contain a variety of terms in Latin (in Latin alphabet) or Latin terms transliterated in Cyrillic alphabet.

Using previously developed NLP tools [16] we extract from the records' text:

- *Diagnoses* – diagnoses with their codes according to the International Classification of Diseases, 10th Revision (ICD-10) [15]; this tool is applied to the field "Anamnesis" where diseases names might occur in the free text;
- *Examinations* – values of BMI, weight, HbA1c, Waist circumference, Triglycerides, Cholesterol, HDL-cholesterol, and Fasting blood glucose; also values for high BP and data about current patient status, blood sugar and values of Lab tests needed to monitor hypertension treatment, diabetes compensations and risk;
- *Treatment* – drugs, dosage, frequency and route, mainly included in the "Therapy" section but sometimes the "Anamnesis" contains sentences that discuss the current or previous treatment. In 2010-2011 we developed a drug extractor, based on algorithms using regular expressions to describe linguistic patterns [17]. There are more than 80 different patterns for matching text units to ATC drug names/codes [18] and NHIF drug codes, medication name, dosage and frequency. Currently, the extractor is elaborated and handles 2,239 drug names included in the NHIF nomenclatures;

The XML structure of the records enables direct identification of:

- *Personal Data* – age and gender of (anonymous) patients;
- *Visit-related information* – doctors' code, region of practice (RZOK), sub-region (ZdrRajon), code of doctors' medical specialty from 00 to 56, etc.;

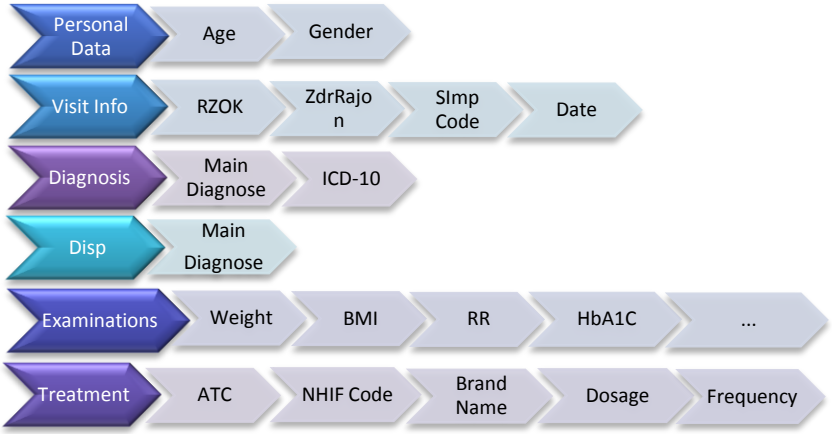

Fig. 2. Combining IE and XML-tagged content to structure information about patient visits

- *Diagnoses* – the principal and other diagnoses encoded in ICD-10;
- *Disp* – a special code is used for "*Disp*" in case the patient needs special monitoring (called "Dispanserization" in Bulgaria, it is a system of measures to support patients in worsened health condition).

Table 1 presents the two collections of clinical texts used for training and tests in our experiments, containing data for patients suffering from Diabetes mellitus (ICD-10 codes E10-E15) and Hypertension (ICD-10 codes I10-I15).

Table 1. Clinical text collections

Set	Outpatient Records	Patients	Diagnose ICD-10	Period	Size (GB)
T1	6,327,503	435,953	E10-E14	1 year	18
T2	50,840,676	5,299,588	I10-I15	3 years	212

4. Methods

4.1. Pre-processing

Here we discuss in more detail the construction of corpora needed for the development of the numeric value extractor. Our aim is to construct carefully the corpus avoiding any potentially abnormal high frequency of some phrases and notations. The particular example is recognition of BP values for diabetic patients.

Initially 10,000 patients suffering from Diabetes Mellitus Type 2 were chosen from the collection T1 for 1 year period – with the highest number of visits 182,726 in total, to ensure many occurrences of the desired data. The GPs' language is similar in most outpatient records but the notes written by specialists are different as they do not do usual examinations (e.g., Ophthalmologists rarely measure the BP).

We split the collection of all outpatient records into collections depending on the specialists who wrote them. There are in total 57 different specialist codes in the collections. For the pre-processing phase 20 of them are considered. Table 2 contains the top 6 types of specialties, for whom the total number of numerical data occurrences is the highest one.

For the 10,000 patients under consideration, a total of 287,394 occurrences of numerical data are identified for all kinds of specialists with 7.93 values in average per outpatient record and 28.74 per patient. Numerical values referring to TIMEX3 tags [19], Protocol Numbers, Codes for Lab Exams and others are excluded.

Table 2. Distribution of numerical data in the clinical text from different specialists

Expert code	Specialty	Number of records	Total numerical values	Average per record	Average per patient
00	General Practitioner	123,247	403,370	3.27	40.34
05	Endocrinology, Metabolic disorders	14,753	157,158	10.65	15.72
08	Cardiology	15,069	93,372	6.20	9.34
15	Ophthalmology	5,109	25,982	5.09	2.60
03	Gastroenterology	1,395	5,285	3.79	0.53
19	Pneumology and Physical Therapy	1,348	4,640	3.44	0.46

We use a collocation extractor developed at the Chulalongkorn University [20] to collect all words, as well as the 2-grams and 3-grams surrounding the strings of numerical data. Chi-square, Log Likehood and Mutual Information statistical methods are applied for the collocation extraction. The best 2,000 candidates with *p*-value of 0.005 are chosen for each method; after manual selection by experts, a golden standard with stop words is prepared to indicate different types of clinical examinations. Reference values are set for each item, to define the possible range of the numeric data.

It is seen that the majority of numerical data for BP are reported in the "Status" field and just few instances are available in the fields "Anamnesis" and "Examine".

We also notice that the outpatient records written by some types of specialists contain no BP measurements: e.g., Gynecologist, Orthopedist, Otorhinola-ryngologist, Ophthalmologist, Urologist and doctors in Physical Medicine and Rehabilitation, as well as Vascular Surgery. Thus we can immediately generalise the rule that for such specialists no BP data occur in the general case, because of the limitations of our small training excerpt. Moreover, some kinds of specialists are not represented in this excerpt collection.

Numerous patterns for BP values were identified, Table 3 lists some examples. The main difference between them, besides the variety of delimiters, are the abbreviations in Bulgarian and Latin for BP like RR (Riva Roci), PP (Рива Рочи / RR in Cyrillic), Кръвно налягане (blood pressure) and abbreviations like АН (Артериално Налягане – arterial pressure), АКН (Артериално Кръвно Налягане – arterial blood pressure), etc.

Table 3. Variety of recording measurements of blood pressure

Кръвно налягане: 125 75	RR 12585	АН- 140/90
АКН 140/ 90	кр. нал.-135 /80	РР между 130-14080
PP. 130 . 80	RR: 120/70	PP140-85
PP- 110 /70	RR.130 90	RR 125/80mm/Hg

Negative examples are also very important because the complexity of clinical data extraction requires deep study and detailed description of all possible cases. After combining positive and negative examples some 364 different patterns for BP values extraction were identified.

4.2. Clinical exam data language $\mathcal{L}$

We define the clinical exam data language $\mathcal{L}$ by a Context-Free Grammar (CFG) $\Gamma = \langle N,\ \Sigma,\ P,\ S \rangle$, where:

- N is a finite set of nonterminal symbols;
- Σ is the alphabet - a finite set of terminal symbols, $N \cap \Sigma = \varnothing$;
- P is a finite set of production rules, $P \subseteq N \times (N \cup \Sigma)^*$;
- $S \in N$ is the start (sentence) symbol.

There are dozens of nonterminals in the grammar. Due to the limited space we give examples to illustrate the major grammar constituents. The meaning of some more important nonterminal symbols is explained below:

S – describes the clinical exam data string (sentence) in the language $\mathcal{L}$;

G – describes a group marker - the string preceding the examination data;

D – represents delimiters like blank spaces, special characters, etc.;

R_i – denotes the examination result;

E – is the exam marker – (optional) marker for sequence of examination results. Often used as "Изследване" or its abbreviation „Изсл." (Examination);

H – examination code according to the NHIF nomenclatures;

C – class of the examination;

M – result marker – an (optional) marker for sequence of examination results. Often used as "Резултат" or its abbreviation „Рез." (Result);

F – specification – (optional) specification of special conditions in which the examination was done/taken;
T – type of the examination/ Theme;
I – individual marker;
A – data;
O – (optional) period – date, time;
L – (optional) limit – e.g., *around*, *above*, *below*, *in average*;
U – numerical data can be single or multiple (Fig. 3);
K – (optional) units.
Some production rules in P are listed in Table 4, where ε is the empty word.

Table 4. Some production rules for clinical exam data language $\mathcal{L}$

$S \to GD^*R_1 \mid R_2$	$I \to TD^*F \mid FD^*T \mid T$
$G \to G_1D^*M \mid G_1$	$\mathrm{A} \to \mathrm{O}_1\mathrm{L}_1\mathrm{UK}_1$
$G_1 \to EDG_2 \mid G_2$	$\mathrm{O}_1 \to \mathrm{O}D^* \mid \varepsilon$
$G_2 \to H \mid HD^*G_3 \mid G_3$	$\mathrm{L}_1 \to \mathrm{LD} \mid \varepsilon$
$G_3 \to C \mid FD^*C$	$\mathrm{K}_1 \to D^*\mathrm{K} \mid \varepsilon$
$R_1 \to R_2 \mid A$	Е → Изследване|изследване|Изсл. |изсл.
$R_2 \to ID^*A$	М → Резултат|Рез. |резултат|рез.

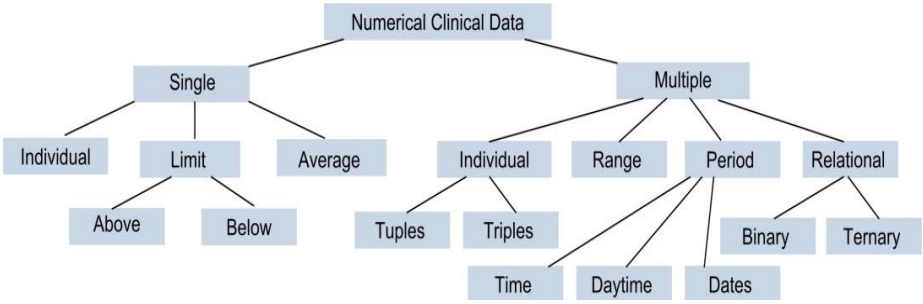

Fig. 3. Numerical data classification using structural and clinical features

It is useful to consider categories of numeric data (Fig. 3) since this facilitates the design of grammar rules. For example, strings with one occurrence of a numeric value are a "*single*" type. The string "Ръст 168 см" (Height 168 cm) is a "*single → individual*", RR 120/80 is "*multiple → individual → tuple*"; the phrase „Поддържа кр. налягане 130-140/85-90." (Sustainable blood pressure 130-140/85-90) belongs to the type "*multiple → range*"; the string "ВОД=0.2 с +1.25дсф.=0.6" (Vis OD=0.2 s+1.25dsf.= 0.6) is type "*multiple → relational → ternary*", etc.

The numerical data can be also classified by numerical data formats, e.g.:

- Integers:
  - Signed: "01.08, result: bld 3+ pro 3+ leu 3+", "v.ok.sin= +1.0 sph -0.5"
  - Unsigned: "Ръст 168 см.; Тегло 60 кг." (Height 168 cm, Weight 60 kg)
- Floating point with comma or dot delimiter:
  - Signed: "ВОД=0.2 с +1.25дсф.=0.6" (VisOD=0.2 s+1.25dsf.= 0.6)
  - Unsigned: "Vis OD=0,3/0,4 Vis OS=0,3/0,4", "Vis OS 1.0 Vis OD 1.0"

Table 5. Some production rules for different types of patterns for data combinations

Type	Production Rules	Example
Individual data in the text (S)	$Text_1\ S\ Text_2$, where $Text_1 \notin \mathcal{L}$ and $Text_2 \notin \mathcal{L}$	*Наднормено тегло-109 кг.-обезитет -IIIст. (Excess weight- 109 kg, Obesity – gr.III)*
Sequential (Q)	$Q \rightarrow Q_1 \mid Q_3$ $Q_1 \rightarrow SQ_2$ $Q_2 \rightarrow DSQ_2 \mid \varepsilon$ $Q_3 \rightarrow GQ_4$ $Q_4 \rightarrow DR_2Q_4 \mid \varepsilon$	*Ръст:154 см, Тегло:80 кг (Height 154 cm, Weight 80 kg)* *Ехокардиография – ЗСЛК=13,5; ЛК=43/29; (Echocardiography – BWLV =13,5, LV=43/29)*
Grouped (W)	$W \rightarrow GDR_2W_1$ $W_1 \rightarrow DR_1W_1 \mid \varepsilon$	*Изсл.: 01.12, резултат: 7ч-11.82 12ч-11.24 15ч-9.75 (Lab.test:01.12, result: 7:00 -11.82 12:00 – 11.24 15:00 - 9.75)*
Nested (V)	$V \rightarrow GDR_2DA$	*Сърце-РСД; СЧ-72/мин; 125/90 (Heart – NHR; FR-*72/min; 125/90)

Moreover, there are several patterns for presentation and combination of multiple Clinical Exam Data in the "Status", "Examine" and "Anamnesis" fields:

- *Individual* – the exam data in the text are surrounded by non-exam data;
- *Sequential* – different exam data are listed sequentially. All exam results R_2 have different types of the examinations T, which must be explicitly specified. There are two types of sequential presentations – with and without group marker.
- *Grouped* – multiple data for single examination. All exam results R_i, $i = 1, 2$, must have the same type of the examination T that must be specified explicitly in R_2 and is optional for R_1.
- *Nested* – This is a mixture between *grouped* and *sequential* patterns. The group of the examinations is specified explicitly but their types differ like in the *sequential* pattern and are usually omitted. The type can be implicitly inferred by the units or the type of numerical data and their referent values.

Table 5 contains examples for the four types of patterns listed above. The consideration of these structural patterns also facilitates the grammar construction.

4.3. Parsing and parsing trees

We propose a hybrid method for parsing clinical examination data. The method is inspired by dependency grammars, constituents and Government and binding theory. Similarly to Voorham and Denig [8], and Bhatia et al. [21], our algorithm is triggered by numeric values, but they use methods that initially extract all numeric values and later search for the closest to them words referring to the examinations, and finally associate them to these numeric values. The obtained result is a list of attribute-value pairs. There are only extracted numerical data that are "*single→individual*" and "*multiple→individual→tuples*". In the proposed original algorithm, similarly to dependency grammars, we initially test some portion of the text and generate constituents for "phrases" in the clinical exam data language $\mathcal{L}$, where the "core words" are numerical values. Later these constituents are combined in parsing trees. This enables the identification of complex relations and structures involving multiple values and concepts. Moreover, in the proposed method several types of numerical values are extracted (see Fig. 3). Additional heuristics for numerical data validation are applied afterwards in order to cope with noisy data.

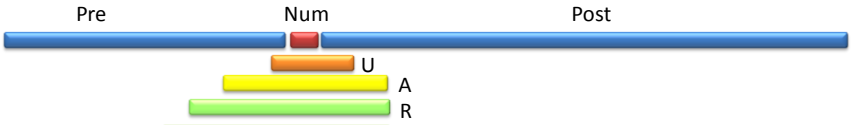

Fig. 4. Intermediate results of the parsing algorithm

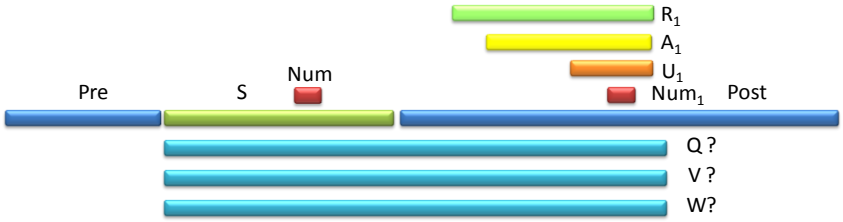

Fig. 5. Construction of the patterns from S for sequential and group patterns

The chart on Fig. 4 illustrates the bottom-up parsing process – from the identification of the first numerical value Num, through matching U (*single*/*multiple* data type), A (data), R (examination result) and S (sentence). The next step is to test the rules for generation of different patterns – Q, W or V in order to parse the maximal possible fragment that corresponds to a string belonging to the language $\mathcal{L}$. This is illustrated in Fig. 5 where the parser tries to combine results in a more complex pattern: sequential (Q), grouped (W) or nested one (V).

Fig. 6 contains an example for *Individual data*, see the first row of Table 5.

Fig. 7 and Fig. 8 illustrate *Sequential exam data*. Fig. 9 shows a parse tree for a *Grouped* data.

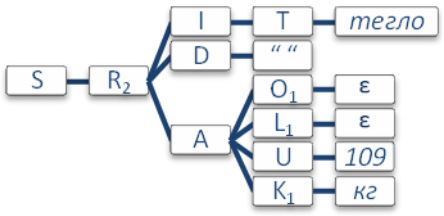

Fig. 6. Parsing tree for individual data: "тегло-109 кг" (weight-109 kg)

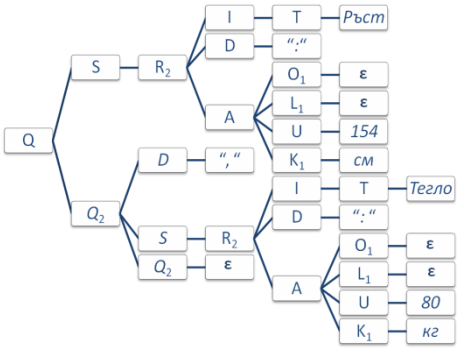

Fig. 7. Parsing tree for a sequential pattern "Ръст: 154 см, Тегло: 80 кг" (height: 154 cm, weight: 80 kg)

The proposed method for clinical data extraction is language independent; collecting the appropriate phrases and terms in the pre-processing phase will allow its application for different languages. In general, the context-free grammar will not change significantly for various natural languages because there are almost standard notations for clinical examination data reports. The bottleneck of the grammar design is the initial selection of corpora for the pre-processing phase and their proper processing and analysis.

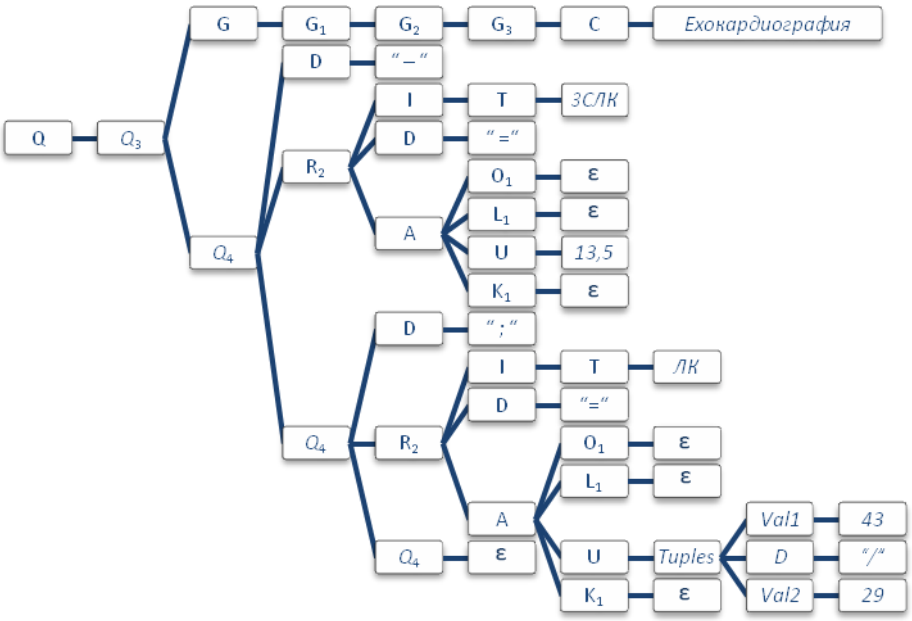

Fig. 8. Parsing tree for a sequential pattern – "Ехокардиография – ЗСЛК=13,5; ЛК=43/29"
(Echocardiography – BWLV=13,5, LV=43/29)

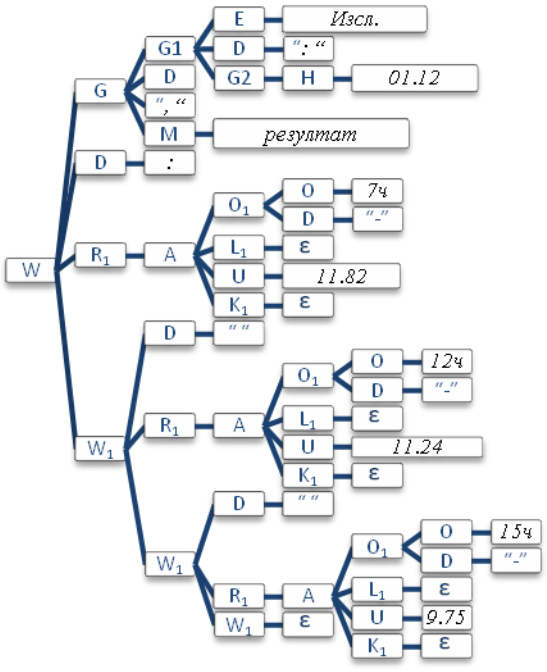

Fig. 9. Parsing tree for a grouped pattern – "Изсл.: 01.12, резултат: 7ч-11.82 12ч-11.24 15ч-9.75"
(LabTest: 01.12, result: 7h-11.82 12h-11.24 15h-9.75)

5. Experiments, results and applications

5.1. Experiments and results

The proposed method is applied for the extraction of clinical exam results in the collections T1-T2 (Table 1). Extracted entities are used for various purposes.

One of the first experiments was carried out using a subset of T1 – all 89,352 records of 3,500 patients with Diabetes Mellitus type 2 who had an incretin based treatment with Glucagon-Like Peptide-1 (GLP-1). The aim was to check whether the information contained in the NHIF archive is sufficient and reliable for statistically-significant conclusions and medical decision making. We note that extracting the mere values is not enough because in this case we also need temporal orientation to be able to monitor the treatment effect. For these reasons the patients were split into two groups: initial treatment (1st phase) and continuing treatment (2nd phase).

This experiment involves extraction of HbA1c, BMI and Weight values. The results for the HbA1c value extraction are summarised in Table 6. In total 10,037 outpatient records contain HbA1c values for 3,469 patients. From all 13,583 mentions of HbA1c, only for 9,357 a particular value is written explicitly; these values are found in 7,830 outpatient records for 3,301 patients. There are 158 different notations for HbA1c values. The amounts of extracted HbA1c values per patient are shown on Fig. 10: there are values of HbA1c examination for almost all patients, and 45% of the patients (1,583 out of 3,500) have 2 measurements.

Table 6. Test for HbA1c indicator

Records/Indicators	Patients	Outpatient Records	Total occurrences
Test set	3,500	89,352	–
HbA1c indicator mentioned	3,469	10,037	13,583
HbA1c with explicit value	3,301	7,830	9,357

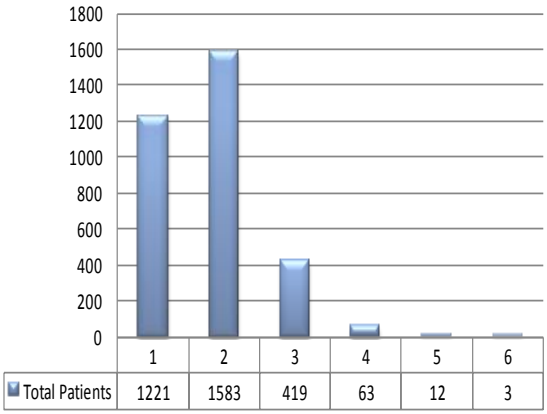

Fig. 10. Frequency of HbA1c measurements per patients in 1-year period

Information about the BMI indicator was available for 1,358 patients. The patients with initial treatment (1st phase) are 1,072, and with continuing treatment (2nd phase) – 286. Fig. 11 presents the distributions of patients according to their

BMI before and after the 1st phase of treatment with GLP-1. Fig. 12 shows the same distributions after the 1st and 2nd phase of treatment. The distribution of the Weight indicator has the same tendencies in the two groups as the indicator BMI.

These analytics results comply to the findings reported in the literature concerning the treatment with drugs belonging to the group GLP-1. This supports our assumption that the NHIF repositories contain sufficiently reliable and detailed information so building decision making strategies on them is feasible.

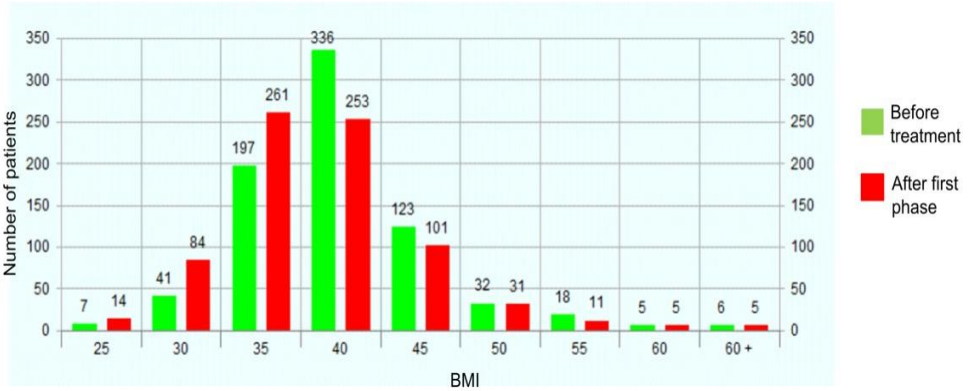

Fig. 11. Effects of initial GLP-1 treatment

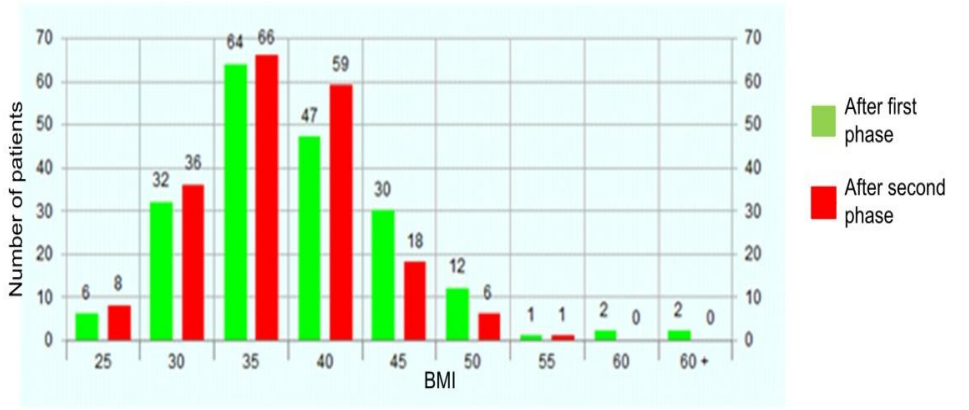

Fig. 12. Effects of continuing GLP-1 treatment

Another cycle of experiments concerns the BP extraction from the outpatient records of about 1,800,000 patients with arterial hypertension for 3 year period from the test set T2. The extracted values are about 38,300,000 with precision 92% and recall 98%. The distribution of data among the XML fields “Anamnesis”, “Status” and “Examine” is presented on Fig. 13, it is counted for a period of 1 year only. We note that the grammar sketched in Section 4 is designed iteratively, in several cycles, with the intention to ensure broadest possible coverage and recall. The experimental results support the belief that the suggested IE method is an adequate technology to attack large repositories of clinical texts.

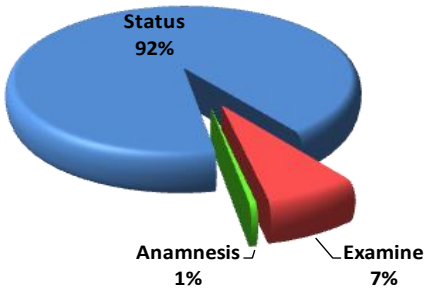

Fig. 13. Distribution of BP values per XML fields of the outpatient records

5.2. Error analysis

The outpatient records in general contain a huge variety of numerical data that do not present values of clinical examinations, like:

- *Date and time* – e.g., the duration of complains "Шум и бръмчене в ушите от около 10-15 дни" (Noise and buzzing in the ears from 10-15 days);
- *Status* – "Hepar -3-4 см.под ребр. дъга" (Liver -3-4 cm below the rib arc); "Бъбреци - паренх. зона 15-16мм" (Kidneys - parenchym 15-16 mm); "в десния киста с д 36мм ..." (Cyst in right kidney with 36 mm length); "На 1 пръст на дясната ръка" (On the 1st finger of the right hand);
- *Drugs and treatment detail*s – "Издаден протокол за Gabanevral 300" (Written protocol for Gabanevral 300); "на лечение с Амарил 3мг" (treatment with Amaril 3 mg); "Провежда лечение с Дигликал 80 млг 2 по 1" (Treatment with Diglical 80 mg 1 tablet 2 times per day);
- *Protocols* – "Издаден нов протокол №1929 / 21.08.2012г." (Written a new protocol No1929 / 21.08.2012)
- *ICD-10 codes of diagnoses* – "Състояние: Новооткрито [L65.8]" (Condition: newly identified [L65.8]); "Състояние: Рецидив [L50]" (Condition: Relapse [L50]).

The major reasons for failures to identify numerical clinical data, even when they occur in the outpatient records, can be summarised as follows:

- *Incorrect data* – numerical values outside the range of the standard values for the particular clinical examination type. The parsing tree is generated successfully, but the numerical data do not pass the test for correctness;
- *Typo errors* – the parser fails to generate parsing trees due to syntax/typo errors in one or more of the markers;
- *Abbreviations* – some abbreviations are used, that are not included in the grammar developed at the pre-processing stage. But this phenomenon is easy to address as there is an option for grammar update and new abbreviations can be taken into consideration for further processing.

After careful analysis of all problems mentioned above, we constantly improve the recall of our IE modules.

5.3. Applications and big data analytics

Big Data Analytics is applied after the IE tools transform the semi-structured outpatient records to a DB repository. The promising evaluation results of the proposed approaches for text mining of clinical texts encourage us to develop more complex applications for medical management support. Thus we can fully benefit from the main advantage of secondary use of clinical data – it requires no additional efforts of medical practitioners since the input documents are results of their usual daily activities and there is no need of additional data collection tasks.

In the project context we developed several applications using the data for 3-years period. The initial data repository contained about 500,000 files with total size exceeding 212 GB, with more than 112 million of outpatient records for about 6,300,000 patients. Since the NHIF collections are pseudonymised, it is possible to trace multiple visits of the same patient and monitor the patient history.

Application 1. The first application is the automatic generation of anonymous Register of Diabetes Mellitus patients (Fig. 14, a screenshot of the Diabetic Register interface). At the beginning the Register content was extracted from the T1 collection; later all available data for a 3-years period were added. The Register contains current patient data, the chronological timeline of visits, the values of relevant indicators, as well as the texts of the outpatient records corresponding to the visits (from where the indicators were extracted).

Our intention is to apply the NLP tools not only at the initial stage of Register generation, but also for the full cycle of its maintenance – update, extension, data aggregation, data analytics and so on.

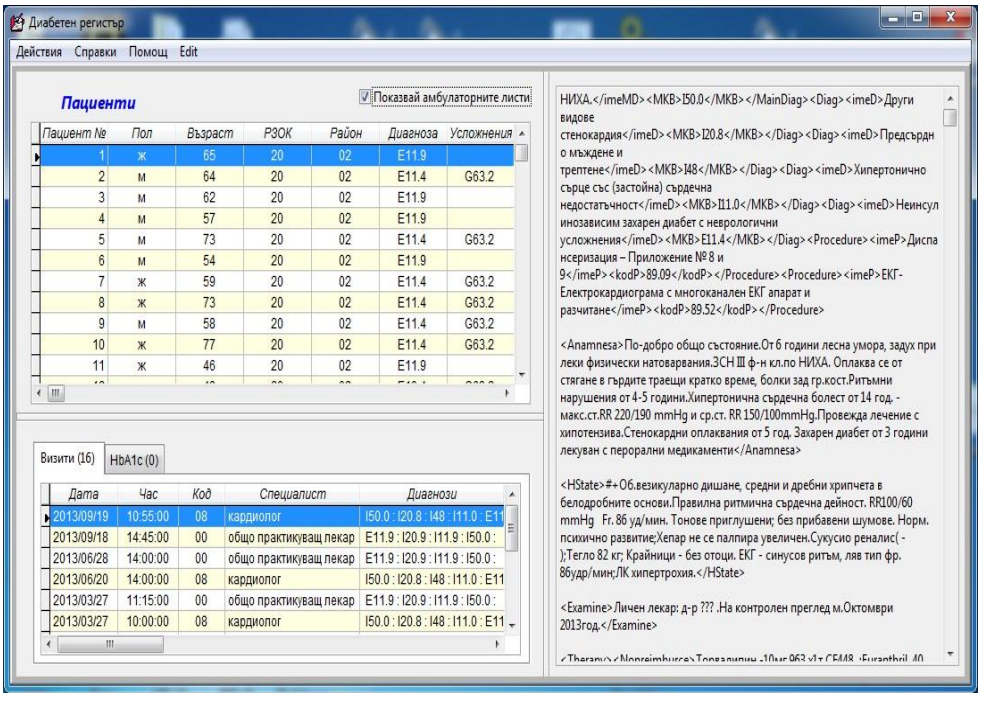

Fig. 14. The interface of the Diabetic Register

The information contained in the Register enables personal control of the health status of particular patients. Fig. 15 shows the changes of HbA1c values for 13 pre-selected patients, compared for two consecutive measurements. For two of the patients (#843 and #1815) the HbA1c values are higher at the second visit. For patient #1815 the change is not significant but patient #843 is clearly decompensated and probably needs special attention.

We notice that the structured information in the Register can be easily aggregated, which allows for observations of more complex phenomena like diabetes compensation – subject to our next application.

Application 2. Using the Diabetes Register, we can evaluate the compensation of diabetic patients in relation with their dispanserization status. The criterion for separation of patients into compensated/decompensated groups is the HbA1c level: "compensated" below 7% and "decompensated" above 7%. When data about HbA1c level is not available in the free text of the outpatient record, we analyze the level of fasting blood glucose or the blood glucose two hours after eating. In this case the criterion for compensation is: for fasting blood glucose – below 7.8 mmol/l and for postprandial blood glucose – below 10 mmol/l.

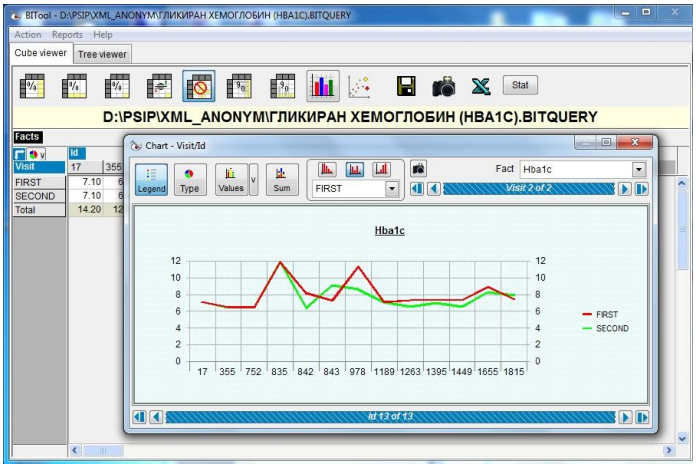

Fig. 15. Monitoring HbA1c levels for selected patients

Since the assessment period for diabetes compensation is long (we deal with a repository for a 3-years period), the patient status can change several times. Therefore, the patient compensation is considered as a percentage: the period during which the patient was compensated divided to the entire observation period. We take into account the fact that the HbA1c level provides information about compensation for a period of 3 months prior to the test. Then the patient is considered "compensated" for the subsequent period until the next measurement.

It is important to analyze the compensation of the diabetic patients as a function of their dispanserization status in order to assess the complex of diagnostic and treatment activities carried out at the dispensary monitoring. The patients were divided into two groups – (*i*) with dispanserization for diagnoses with ICD-10 type E and (*ii*) without dispanserization. The analysis included 359,682 patients monitored in 3 years period, for whom the necessary data about measurements of

blood sugar and HbA1c were available. In this framework, we group the diabetic patients into categories: those who were "compensated" most of the time, i.e., above 50% in the 3-years period and who were "decompensated" most of the time, i.e., above 50% in the 3-years period. These categories are mapped to the dispanserization status of the respective patients. The results are shown in Fig. 16: the percentage of patients with "compensated" Diabetes Mellitus in the dispensary care is 23.48% while the percentage of the "compensated" diabetic patients outside it is higher, reaching 34.66%.

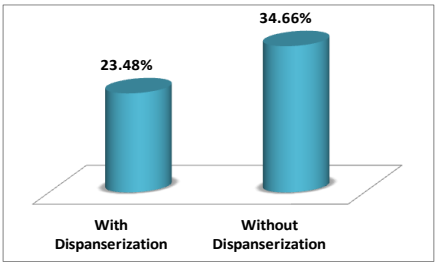

Fig. 16. Diabetes Mellitus Type 2 compensation for 3-years period

This result provides arguments to consider the efficiency of the diagnostic and treatment activities within the dispensary system (taken into account the fact that the dispensary care deals with patients who are seriously ill). Processing the NHIF archive, we can make calculations for each general practitioner separately and help the healthcare authorities in their monitoring tasks.

Application 3. Another important application is to assess therapy efficiency. We trace major clinical findings as HbA1c levels and BMI for patients suffering from Diabetes Mellitus who have an incretin-based treatment with GLP-1. Fig. 17 presents the significant decrease of the HbA1c levels after the incretin-based therapy (1st phase). Using statistical methods like the *t*-test of Student's *t*-distribution we check the validity of the null hypothesis (that the provided treatment does not influence any changes of HbA1c levels and BMI) and can reject it.

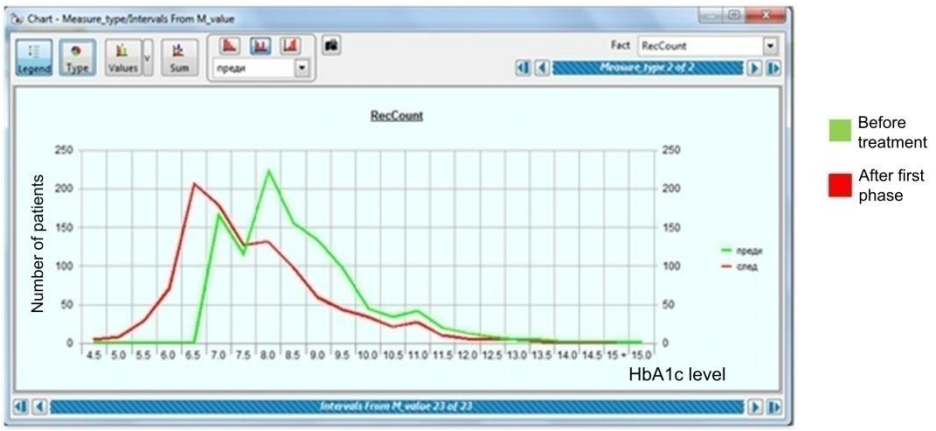

Fig. 17. HbA1c levels for 220 patients before and after the incretin based therapy

Application 4. Using the extractor of BP data values we developed an application for monitoring patients with Arterial Hypertension – a subset of the collection T2 for one year period. The criterion for efficient control of hypertension is keeping blood pressure below 140 mm Hg systolic and 90 mm Hg diastolic pressures in a sitting position at rest. Since the assessment period of hypertension control is long (one year), we designed an algorithm to calculate the period during which the patient's BP was under control (compensated) as a percentage of the entire observation period. The approach is similar to the assessment of diabetes compensation. Once a normal BP value is measured, the patient is considered to be "compensated" for the subsequent period until the next measurement. The scale of compensation levels is divided into 10 intervals: 0-9%, 10-19%, 20-29%, and, etc., because the BP values vary considerably. For example, a patient was examined in January, July and December, and the measured BP values were: in January – RR 130/80, in July – RR 150/95 and in December – RR 145/100. Then the patient is viewed as "compensated" only for the first six months of the year, which is 50% of the entire observation period. This patient is related to the compensation interval 50-59% of the entire observation period.

Summarising the data, we can divide the patients into two groups – those who most of the time had stable control of hypertension (more than 50% of the period) and the ones who could not achieve BP normalisation most of the time during the one year period. Fig. 18 shows that the percentage of patients with arterial hypertension who had adequately controlled blood pressure is 72.96%, and the patients with poor control were 27.04%.

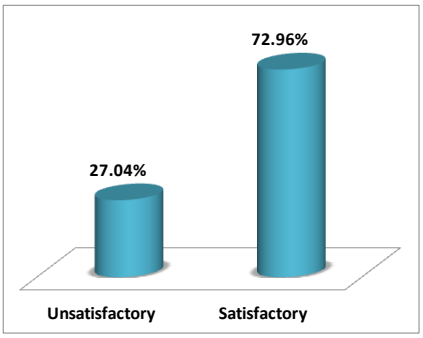

Fig. 18. Arterial hypertension control

6. Conclusion

Today more than 80% of the patient-related clinical information is stored as free text in the Electronic Health Record (EHR) systems. Applying reliable language technologies for text analysis, based on suitable nomenclatures like, e.g., SNOMED CT, is the only way to extract entities and facts that constitute the core of standardised EHR repositories and structures like Archetypes (EN ISO 13606-2) and HL7 CDA (Clinical Document Architecture). These enhanced EHR systems will enable the semantic interoperability between the national health systems and will increase the quality and efficiency of the cross-country healthcare services.

The NLP tools presented in this article enable the data structuring and aggregation, as well as analytics based on various criteria for monitoring of important features including control at individual level. We started to develop applications with impact at national impact, e.g., the anonymous Register of patients with Diabetes Mellitus. Accumulating more data for longer periods we shall be able to address further challenging tasks of major importance for the society, such as assessing the efficiency of dispensarization care for diabetic patients. The in-depth analysis will support the implementation of adequate health policy based on evidence.

It probably becomes clear that scalability is the bottleneck of large scale text analysis. The variety of recording formats and explanations written by thousands of medical professionals require constant evaluation of the grammar coverage and the extraction accuracy in general. One of the main advantages of the proposed method, despite its performance and precision in text mining, is that it is modular, extensible, scalable and flexible. As we above mentioned the technology can be easily adapted for other languages. The modularity of all suggested techniques allows combining them in different applications as it was shown in Section 5. In addition, each of these methods is functionally scalable and can be augmented with additional features for the purposes of particular applications.

All software modules are vertically and horizontally scalable. The distributed processes can be performed only in the local network of an organisation with accredited rights for data access. Several experiments were done with multi-core CPUs (4 vs 8 core) to evaluate the performance. Our future plans include test with multiple CPUs to evaluate weak vs strong performance scaling. Due to some limitations in the access to healthcare documentation, administrative and geographical scalability is not applicable, on this stage although all documents are pseudo-anonymised.

Diabetes Mellitus and Arterial Hypertension are major chronic diseases. Due to this reason we focus our attention on the extraction of numerical values that are central for the monitoring of patient status during the entire episode of ambulatory and hospital care, and making informed decisions based on evidence. We also enable control of the health status of the particular patient and give hints to medical authorities to propose changes in the treatment plan. Our next aim is to integrate the drug extractor in the analytics tasks in order to trace the developments of diabetes complications.

***Acknowledgement*:** The research work presented in this paper is partially supported by FP7 Grant 316087 AComIn "Advanced Computing for Innovation", funded by the European Commission in the FP7 Capacity Programme in 2012-2016. The authors also acknowledge the support of the Bulgarian Health Insurance Fund, the Bulgarian Ministry of Health and the Medical University – Sofia. The results were presented at the Workshop "Big Data in Education and Digital Collections" held on 29 June 2015 at the Institute of Information and Communication Technologies, Bulgarian Academy of Sciences.

References

1. Meystre, S. G. Savova, K. C. Kipper-Schuler, 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.
2. Jensen, P. B., L. J. Jensen, S. Brunak. Mining Electronic Health Records: Towards Better Research Applications and Clinical Care. – Nature Reviews Genetics, Vol. **13**, 2012, No 6, pp. 395-405.
3. Raghupathi1, W., V. Raghupathi. Big Data Analytics in Healthcare: Promise and Potential. – Health Information Science and Systems, Vol. **2**, 2014, No 3, doi:10.1186/2047-2501-2-3.
4. Champion, H., N. Pizzi, R. Krishnamoorthy. Tactical Clinical Text Mining for Improved Patient Characterization. – In: Proc. of IEEE International Congress on Big Data, 2014, pp. 683-690. doi= 10.1109/BigData.Congress.2014.101**.**
5. Sun, W., A. Rumshisky, O. Uzuner. Evaluating Temporal Relations in Clinical Text: 2012 i2b2 Challenge. – Journal of American Medical Informatics Association, Vol. **20**, 2013, pp. 806-813. doi:10.1136/amiajnl-2013-001628.
6. Skeppstedt, M., M. Kvist, G. H. Nilsson, H. Dalianis. Automatic Recognition of Disorders, Findings, Pharmaceuticals and Body Structures from Clinical Text: An Annotation and Machine Learning Study. – Journal of Biomedical Informatics, Vol. **49**, 2014, pp. 148-158.
7. Zhou, X., H. Han, I. Chankai, A. Prestrud, A. Brooks. Approaches to Text Mining for Clinical Medical Records. – In: Proc. of 21st Annual ACM Symposium on Applied Computing 2006, Technical Tracks on Computer Applications in Health Care (CAHC'2006), Dijon, France, 2006, pp. 235-239.
8. Voorham, J., P. Denig. Computerized Extraction of Information on the Quality of Diabetes Care from Free Text in Electronic Patient Records of General Practitioners. – Journal of American Medical Informatics Association, Vol. **14**, May-Jun 2007, No 3, pp. 349-354. doi: 10.1197/jamia.M2128.
9. Turchin, A., N. S. Kolatkar, R. W. Grant, E. C. Makhni, M. L. Pendergrass, J. S. Einbinder. Using Regular Expressions to Abstract Blood Pressure and Treatment Intensification Information from the Text of Physician Notes. – Journal of the American Medical Informatics Association, Vol. **13**, November-December 2006, No 6, pp. 691-695.
10. Murtaugh, M., B. S. Gibson, D. Redd, Q. Zeng-Treitler. Regular Expression-Based Learning to Extract Bodyweight Values from Clinical Notes. – J. Biomed Inform., Vol. **54**, April 2015, pp. 186-90. doi: 10.1016/j.jbi.2015.02.009. Epub 2015 Mar 5.
11. Min, J., Y. Chen, M. Liu, S. T. Rosenbloom, S. Mani, J. C. Denny, H. Xu. A Study of Machine-Learning-Based Approaches to Extract Clinical Entities and Their Assertions from Discharge Summaries. – J. Am Med Inform Assoc., Vol. **18**, September-October 2011, No 5, pp. 601-6. doi: 10.1136/amiajnl-2011-000163. Epub 2011 Apr 20.
12. Patrick, J. D., D. H. M. Nguyen, Y. Wang, M. Li. A Knowledge Discovery and Reuse Pipeline for Information Extraction in Clinical Notes. – J. Am. Med. Inform. Assoc., Vol. **18**, September-October 2011, No 5, pp. 574-9. doi: 10.1136/amiajnl-2011-000302. Epub 2011 Jul 7.
13. Bigeard, E., V. Jouhet, F. Mougin, F. Thiessard, N. Grabar. Automatic Extraction of Numerical Values from Unstructured Data. – In: EHRs. Digital Healthcare Empowering Europeans, R. Cornet et al., Eds. 2015, European Federation for Medical Informatics (EFMI). doi:10.3233/978-1-61499-512-8-50.
14. Nikolova, I., D. Tcharaktchiev, S. Boytcheva, Z. Angelov, G. Angelova. Applying Language Technologies on Healthcare Patient Records for Better Treatment of Bulgarian Diabetic Patients. – In: G. Agre et al., Eds. Artificial Intelligence: Methodology, Systems, and Applications, Lecture Notes in Artificial Intelligence, Vol. **8722**, Springer, 2014, pp. 92-103.
15. International Classification of Diseases and Related Health Problems 10th Revision. **http://apps.who.int/classifications/icd10/browse/2015/en**

16. A n g e l o v a, G., D. T c h a r a k t c h i e v, S. B o y t c h e v a, I. N i k o l o v a, H. D i m i t r o v, Z. A n g e l o v. From Individual EHR Maintenance to Generalised Findings: Experiments for Application of NLP to Patient-Related Texts. – In: R. Kountchev and B. Iantovics, Eds. Advances in Intelligent Analysis of Medical Data and Decision Support Systems, Studies in Computational Intelligence, Vol. **473**, Springer, 2013, pp. 203-212.
17. B o y t c h e v a, S. Shallow Medication Extraction from Hospital Patient Records. – In: V. Koutkias, J. Nies, S. Jensen, N. Maglaveras, and R. Beuscart, Eds. Studies in Health Technology and Informatics, Vol. **166**, IOS Press, pp. 119-128.
18. Anatomical Therapeutic Chemical (ATC) Classification System. **http://atc.thedrugsinfo.com/**
19. Time ML Markup Language for Temporal and Event Expressions, Version 1.2.1, October 2005. **http://www.timeml.org/site/index.html**
20. Wirote Aroonmanakun, 2000. Collocation Extract. V. 3.06. **http://pioneer.chula.ac.th/~awirote/resources/collocation-extract.html**
21. B h a t i a, R. S., A. G r a y s t o n e, R. A. D a v i e s, S. M c C l i n t o n, J. M o r i n, R. F. D a v i e s. Extracting Information for Generating A Diabetes Report Card from Free Text in Physicians Notes. – In: Proc. of NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents, Association for Computational Linguistics, 2010, pp. 8-14. **http://www.aclweb.org/anthology/W10-1102**