Mining Clinical Events to Reveal Patterns and Sequences

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

[1] Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Bulgaria
[2]Adiss Lab Ltd., Sofia, Bulgaria
[3]Medical University Sofia, University Specialised Hospital for Active Treatment of Endocrinology, Bulgaria

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

Abstract.
This paper presents results of ongoing project for discovering complex temporal relations between disorders and their treatment. We propose a cascade data mining approach for frequent pattern and sequence mining. The main difference from the classical methods is that instead of applying separately each method we reuse and extend the result prefix tree from the previous step thus reducing the search space for the next task. Also we apply separately search for diagnosis and treatment and combine the results in more complex relations. Another constraint is that items in sequences are distinct and we have also parallel episodes and different time constraints. All experiments are provided on structured data extracted by text mining techniques from approx. 8 million outpatient records in Bulgarian language. Experiments are applied for 3 collections of data for windows with size 1 and 3 months, and without limitations. We describe in more details findings for Schizophrenia, Diabetes Mellitus Type 2 and Hyperprolactinemia association.

Keywords: Medical Informatics, Big Data, Text mining, Temporal Information, Data mining

1 Motivation

Analyzing relations between temporal events in clinical narratives has high importance for proving different hypotheses in healthcare: in risk factors analysis, treatment effect assessment, comparative analysis of treatment with different medications and dosage; monitoring of disease complications as well as in epidemiology for identifying complex relations between different disorders and causes for their co-existence – so called comorbidity. A lot of research efforts were reported in the area of electronic health records (EHR) visualization and analysis of periodical data for single patient or searching patterns for a cohort of patients [1, 2, 3, 4, 5]. The work [6] proposes a method for temporal event matrix representation and a learning framework that discovers complex latent event patterns or Diabetes Mellitus complications.

adfa, p. 1, 2011.

Patnaik et al. [1, 2] report one of the first attempts for mining patients' history in big data scope – processing over 1.6 million of patient histories. They demonstrate a system called EMRView for mining the precedence relationships to identify and visualize partially ordered information. Three tasks are addressed in their research: mining parallel episodes, tracking serial extensions, and learning partial orders.

Mining frequent event patterns is a major task in data mining. It filters events with similar importance and features; this relationship can be specified by temporal constraints. There are two major tasks in data mining related to the temporal events analysis: *(i)* frequent patterns mining and *(ii)* frequent sequence mining. The difference between them is that in the first case the event order does not matter.

In *frequent patterns mining* the events are considered as sets – collections of objects called itemsets. We investigate how often two or more objects co-occur. Usually they are considered as a database of transactions presented like tuples $(transaction, itemset)$, the sets of transaction identifiers are called tidsets. Several methods are proposed for solving this task that vary from the naive BruteForce and Apriori algorithms, where the search space is organized as a prefix tree, to Eclat Algorithm that uses tidsets directly for support computation, by processing prefix equivalence classes [7]. An improvement of Eclat is dEclat, it reduces the space by keeping only the differences in the tidsets as opposed to the full tidsets. Another efficient algorithm is Frequent Pattern Tree Approach – FPGrowth Algorithm. Using the generated frequent patterns by all these methods we can later generate association rules.

For *frequent sequence mining* the order does matter [7]. The Level-wise generalized sequential pattern (GSP) mining algorithm searches the prefix tree using breadth-first search. SPADE algorithm applies vertical sequence mining, by recording for each symbol the position at which it occurs. PrefixSpan algorithm uses projection-based sequence mining by storing only the suffix after the first occurrence of each symbol and removing infrequent symbols from the suffix. This algorithm uses depth-first search only for the individual symbols in the projection database.

There are different mining approaches for temporal events, for instance we can consider sequences leading to certain target event [8]. Gyet and Quiniou [9] propose recursive depth-first algorithm QTIPrefixSpan that explores the extensions of temporal patterns. Further they extract temporal sequences with quantitative temporal intervals with different models using a hyper-cube representation and develop a version of EM algorithm for candidates' generation [10]. Patnaik et al. present the streaming algorithm for mining frequent episodes over a window of recent events in the stream [2]. Monroe et al. [11] presents a system with visual tools that allows the user to narrow iteratively the process for mining patterns to the desired target with application in EHRs. Yang et al. [12] describe another application of temporal event sequence mining for mining patient histories. They develop a model-based method for discovering common progression stages in general event sequences.

2 Project Setup

The main goal of our research is to examine comorbidity of diseases and their relationship/causality with different treatment, i.e. how the treatment of a disease can affect the co-existing other disorders. This is a quite challenging task, because the number of diagnoses (more than 10,000) and of medications (approx. 6,500) is huge. Thus the theoretically possible variations of diagnoses and corresponding treatments are above 10^{500} for one patient. That is why we shall examine separately chronic vs. acute diseases [13] and afterwards shall combine the patterns into more complex ones. Chronic diseases constitute a major cause of mortality according to the World Health Organization (WHO) reports and their study is of higher importance for healthcare.

In order to solve this challenging task we split it down into two subtasks:

- **Task 1:** To find frequent patterns of distinct chronic diseases. Afterwards for each frequent pattern of chronic diseases – to find frequent patterns of treatment and investigate their relationship in order to identify complex relations.
- **Task 2:** To search for causality and risk factors for chronic diseases by sequence mining. In this task several experiments are explored – with no limitations for the distance between events, only the order matters, and with different window limitations between events – 1 month, 3 months, etc. In this task also more complex sequences are considered like parallel (simultaneous) episodes/disorders.

Each of these tasks needs to be investigated in the general case and also for gender and age specific conditions.

3 Materials

We deal with a repository of pseudoanonymous Outpatient Records (OR) in Bulgarian language provided by the Bulgarian National Health Insurance Fund (NHIF) in XML format. The majority of data necessary for the health management are structured in fields with XML tags, but there are still some free-text fields that contain important explanations about the patient: “Anamnesis”, “Status”, “Clinical examinations” and “Therapy”.

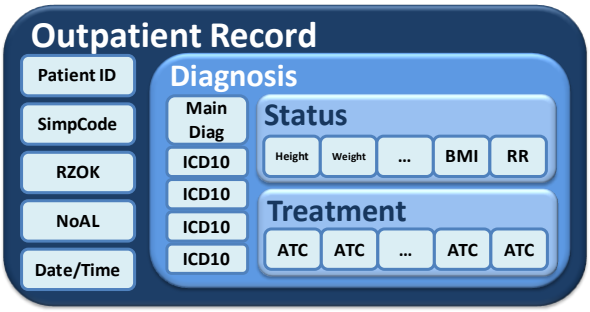


Fig. 1. Structured event data

From the XML fields with corresponding tags we know the Patient ID, the code of doctors' medical specialty (SimpCode), region of practice (RZOK), Date/Time and ID of the outpatient record (NoAL). XML tags also point to the main diagnose and additional diagnoses with their codes according to the International Classification of Diseases, 10th Revision (ICD-10) [14]. Each OR contains a main diagnosis ICD-10 code and up to 4 additional disorders ICD-10 codes, i.e. in total from 1 to 5 ICD-10 codes. ORs describe events that can be represented in structured format (Fig. 1).

Our experiments for pattern search are made on three collections of outpatient records that are used as training and test corpora, see Table 1. They contain data about patients suffering from Schizophrenia (ICD-10 code F20), Hyperprolactinaemia (ICD-10 code E22.1), and Diabetes Mellitus (ICD-10 codes E10-E15). These collections are of primary interest for our project because they contain cases with high diversity of chronic disorders. Schizophrenia and Diabetes Mellitus are chronic diseases with a variety of complications that are also chronic diseases. The collections are extracted by using a Business Intelligence tool (BITool) [15] from the repository of about 120 millions ORs for approx. 5 million patients for 3 years period. The size of the repository is approx. 212 GB.

Table 1. Characteristics of data collections

Collections / Characteristics	S1	S2	S3
Outpatient Records	1,682,429	288,977	6,327,503
Patients	45,945	9,777	435,953
Main Diagnose ICD-10	F20	E22.1	E10-E15
Period	3 years	3 years	1 year
Size	4 GB	1 GB	18 GB

4 Methods

We designed a system for exploring complex relationships between disorders and treatments, see Fig. 2. It contains two main modules - for text mining and for data mining, and two repositories – for XML documents (ORs) and for structured data (temporal events sequences).

Text mining module is responsible for the conversion of the raw text data concerning treatment and status to structured event data and in addition for the "translation" of the structured data in the XML document to event data. For extraction of information about the treatment we use a Text mining tool because the ORs contain free texts discussing drugs, dosage, frequency and route mainly in the "Therapy" section. Sometimes the "Anamnesis" also contains sentences that discuss the current or previous treatment. We developed a drug extractor using regular expressions to describe linguistic patterns [16]. There are more than 80 different patterns for matching text units to ATC drug names/codes [17] 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. For extraction of clinical examination

data we designed a Numerical value extractor [18] that processes lab and test results described in "Anamnesis", "Status", and "Clinical examinations" – for instance body mass index (BMI), weight (W), blood pressure (Riva Roci - RR), etc.

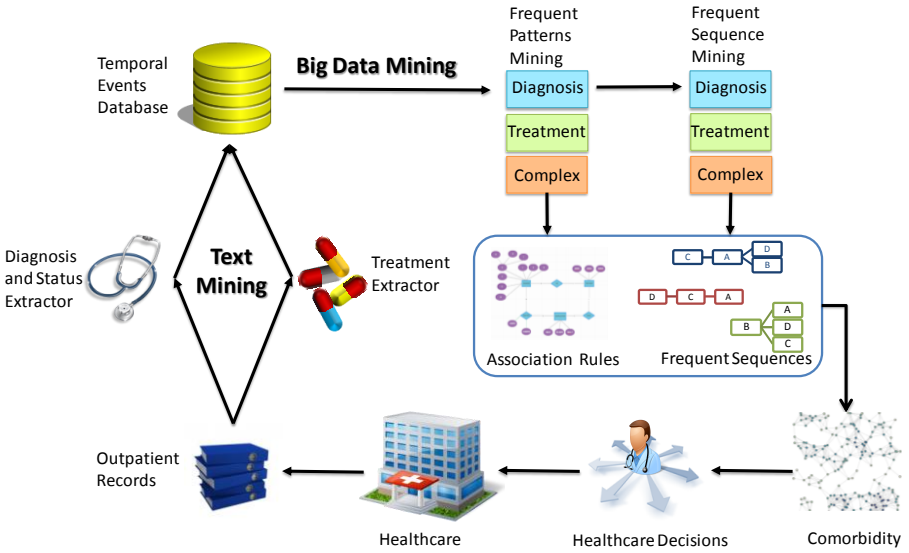


Fig. 2. System architecture

Data mining module uses a cascade approach for solving the two main tasks. The process will be shown in more details in the following subsections. Briefly the idea is that Task 1 can be solved by applying modification of the classical frequent itemsets mining algorithms and association rules generation. The solution of Task 2 is based on frequent sequence mining expanding the prefix tree generated as a result from the previous task solution. There are a lot of efficient algorithms for solving each task separately. However, in our project we are interested in a single algorithm that solves both tasks. That is why we propose a cascade method that uses the obtained results from the previous task and in such a way reduces the search space for the next task. Another constraint is that items in sequences in Task 2 are distinct and we have also parallel (simultaneous) episodes and different time constraints.

4.1 Formal Presentation

Each collection $S \in \{S1, S2, S3\}$ is processed independently from the other two collections. For the collection S the set of all different patient identifiers $P = \{p_1, p_2, \dots, p_N\}$ is extracted. This set corresponds to transaction identifiers (tids) and we call them *pids* (patient identifiers). For each patient $p_i \in P$ events sequence of tuples $\langle event, timestamp \rangle$ is generated: $E(p_i) = \left(\langle e_1, t_1 \rangle, \langle e_2, t_2 \rangle, \dots, \langle e_{k_i}, t_{k_i} \rangle\right)$, $i = \overline{1, N}$ where timestamps $t_{n-1} \leq t_n$, $n = \overline{2, N}$. Let $\mathcal{E}$ be the set of all possible events and $\mathcal{T}$ be the set of all possible timestamps. Let $\mathcal{C} = \{c_1, c_2, \dots, c_p\}$ be the set of

all chronic diseases [13], which we call *items*. Each subset of $X \subseteq \mathcal{C}$ is called an *itemset*. We define a projection function $\pi: (\mathcal{E} \times \mathcal{T})^n \rightarrow (\mathcal{C} \times \mathcal{T})^n$:

$$\pi\big(E(p_i)\big) = C(p_i) = \big(\langle c_1, t_{1i}\rangle, \langle c_2, t_{2i}\rangle, \dots, \langle c_{m_i}, t_{m_i}\rangle\big)$$

such that for each patient $p_i \in P$ the projected time sequence contains only the first occurrence (onset) of each chronic disorder of the patient p_i that is recorded in $E(p_i)$ in the format $\langle chronic\ disease, timestamp\rangle$.

Table 2. Example database for chronic diseases with timestamps

pid				
1.	$\langle A, 12/01/2012\rangle$	$\langle B, 12/01/2012\rangle$	$\langle C, 01/02/2012\rangle$	$\langle D, 16/05/2012\rangle$
2.	$\langle B, 27/01/2012\rangle$	$\langle C, 06/02/2012\rangle$	$\langle D, 13/09/2012\rangle$	–
3.	$\langle B, 03/02/2012\rangle$	$\langle C, 08/02/2012\rangle$	$\langle A, 08/02/2012\rangle$	$\langle D, 27/06/2012\rangle$
4.	$\langle B, 10/02/2012\rangle$	$\langle A, 10/02/2012\rangle$	$\langle D, 19/06/2012\rangle$	–
5.	$\langle A, 22/02/2012\rangle$	$\langle C, 22/02/2012\rangle$	$\langle D, 20/08/2012\rangle$	–
6.	$\langle B, 28/02/2012\rangle$	$\langle A, 14/03/2012\rangle$	$\langle C, 14/03/2012\rangle$	$\langle D, 25/09/2012\rangle$

To investigate both cases: with no limitations and with different window limitations of the distance between events, we store two versions of the temporal event sequences database for each collection.

Table 3. Example database for chronic diseases with normalized timestamps

pid	(a)			
1.	$\langle A, 0\rangle$	$\langle B, 0\rangle$	$\langle C, 1\rangle$	$\langle D, 2\rangle$
2.	$\langle B, 0\rangle$	$\langle C, 1\rangle$	$\langle D, 2\rangle$	–
3.	$\langle B, 0\rangle$	$\langle C, 1\rangle$	$\langle A, 1\rangle$	$\langle D, 2\rangle$
4.	$\langle B, 0\rangle$	$\langle A, 0\rangle$	$\langle D, 1\rangle$	–
5.	$\langle A, 0\rangle$	$\langle C, 0\rangle$	$\langle D, 1\rangle$	–
6.	$\langle B, 0\rangle$	$\langle A, 1\rangle$	$\langle C, 1\rangle$	$\langle D, 2\rangle$

pid	(b)			
1.	$\langle A, 0\rangle$	$\langle B, 0\rangle$	$\langle C, 20\rangle$	$\langle D, 125\rangle$
2.	$\langle B, 0\rangle$	$\langle C, 10\rangle$	$\langle D, 230\rangle$	–
3.	$\langle B, 0\rangle$	$\langle C, 5\rangle$	$\langle A, 5\rangle$	$\langle D, 145\rangle$
4.	$\langle B, 0\rangle$	$\langle A, 0\rangle$	$\langle D, 130\rangle$	–
5.	$\langle A, 0\rangle$	$\langle C, 0\rangle$	$\langle D, 180\rangle$	–
6.	$\langle B, 0\rangle$	$\langle A, 15\rangle$	$\langle C, 15\rangle$	$\langle D, 210\rangle$

In the first version all timestamps are substituted with consecutive numbers starting from 0. In this case the particular dates of events do not matter, only the order matters. In this section we introduce a simple synthetic example to illustrate the proposed method (Table 2). For the example on Table 2, the corresponding database with normalized timestamps is shown on Table 3(a). In the second version all timestamps are replaced with relative time – to the first event in the sequence we assign time 0, and for all other events the timestamp is converted to the number of days distance from the first event. In this case the distance between events does matter. The corresponding database with normalized timestamps is shown on Table 3(b). Note that ORs always contain dates.

Additionally we generate time sequences for treatment. For each medication we can find the corresponding diseases with which treatment it is associated. Similarly to disorders we define a projection function that, applied over the event sequence, results a treatment sequence.

4.2 Frequent Patterns Mining

This module applies a modification of the classical Eclat algorithm [7] and generates a prefix-based search tree for chronic diseases itemsets (Fig. 4). In our task frequent patterns are itemsets. Then the main difference is that instead of using pids intersection we apply projection of items in the database and at each level we merge the projection vectors. Let $x \in \mathcal{C}$ is an item and $\mathcal{D}$ is the database. We define a vector projection $v(x, \mathcal{D}) = \langle x_1, \dots, x_n \rangle$, where:

$$x_i = \begin{cases} k & , \exists k: \langle p_i, X \rangle \in \mathcal{D} \; and \; \langle x, k \rangle \in X \\ -1 & , otherwise \end{cases}$$

For the database on Table 3(a) we obtain the projection vectors shown in Fig. 3.

	1	**2**	**3**	**4**	**5**	**6**
A	<0,	-1,	1,	0,	0,	1>
B	<0,	0,	0,	0,	-1,	0>
C	<1,	1,	1,	-1,	0,	1>
D	<2,	2,	2,	1,	1,	2>

Fig. 3. Projection vectors of the chronic diseases A, B, C, D (Table 3a)

Each itemset is considered as a sequence of items in lexicographical order, according to the ICD-10 indices. Let $a = p + d', b = p + d''$ be two sequences of chronic diseases with the same prefix p and corresponding vectors $\langle a_1, \dots, a_n \rangle$ and $\langle b_1, \dots, b_n \rangle$. If $d' < d''$, then we generate the result sequence $ab := p + d' + d'' = ad''$. Let the vector for item d'' be $\langle d_1'', \dots, d_n'' \rangle$. For the sequence ab with length k we generate the vector of ordered k-tuples $\langle m(a_1, d_1''), \dots, m(a_n, d_n'') \rangle$, where the function m is defined as follows:

$$m(a_i, d_i'') = \begin{cases} -1 & , a_i = -1 \; or \; d_i'' = -1 \\ \langle a_i, d_i'' \rangle & , a_i \neq -1 \; and \; d_i'' \neq -1 \end{cases}$$

We define a support for sequence a with vector $\langle a_1, \dots, a_n \rangle$ as $sup(a) = |\{a_i | a_i \neq -1, i = \overline{1,n}\}|$. We are looking for itemsets with frequency above given $minsup$.

From the collection $\mathcal{F}$ of frequent itemsets, using the classic algorithm we generate association rules in the form: $\alpha \overset{s,c}{\rightarrow} \beta$, where $\alpha, \beta, \alpha\beta \in \mathcal{F}, s = sup\,(\alpha\beta) \geq minsup$, and the confidence c is defined as follows:

$$c = \frac{sup\,(\alpha\beta)}{sup\,(\alpha)}$$

We are looking for strong rules, i.e. with confidence above a given minimum confidence value $minconf$.

Although the collection $\mathcal{F}$ gives us some information about the chronic disorders coexistence necessary for Task 1 solution, association rules show more complex relations between disorders association. In Fig. 5 are shown association rules generated from the prefix tree in Fig. 4.

For example from association rules in Fig. 5 we can conclude that existence of anyone or two of the diseases A, B, C is a risk factor for presence of disorder D as well.

In addition the coexistence of diseases A, and D is a risk factor for presence also either of disorder B, or of disorder C.

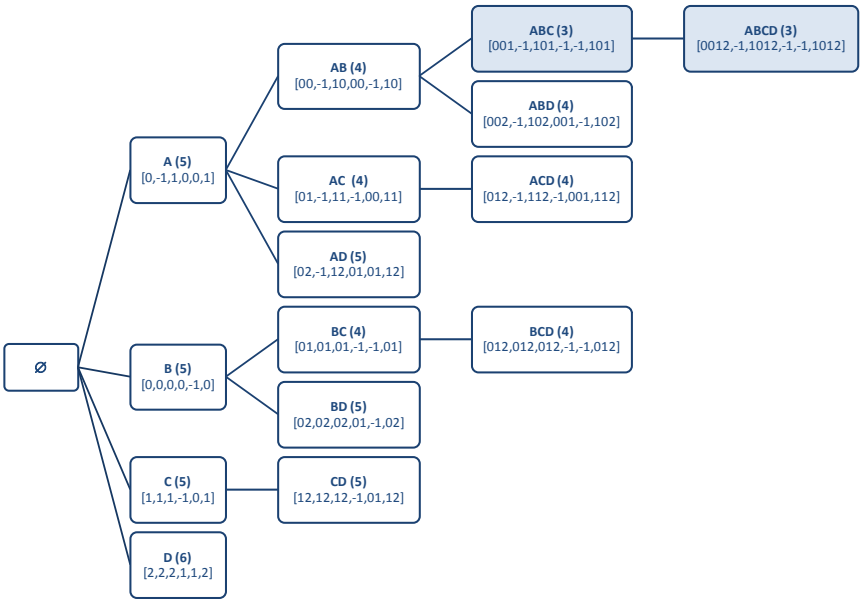


Fig. 4. Prefix tree with projection vectors for $minsup = 4$. Shaded boxes indicate infrequent itemsets. For simplicity parentheses are omitted. Next to the itemset is shown its support.

Association rule	confidence
$AB \rightarrow D$	1.0
$AC \rightarrow D$	1.0
$AD \rightarrow B$	1.0
$AD \rightarrow C$	1.0
$BC \rightarrow D$	1.0
$A \rightarrow D$	1.0
$B \rightarrow D$	1.0
$C \rightarrow D$	1.0

Fig. 5. Association rules generated with $minconf = 0.9$

4.3 Frequent Sequence Mining

For Task 2 we need to study the sequences of chronic disorders and their treatment. We apply breadth-first search in the prefix-tree generated from the previous module and map on each node a substitution σ that converts vectors of k-tuples to vectors of patterns. Let $v = \langle v_1, v_2, \dots, v_n \rangle$ be a vector for some frequent itemset $X \in \mathcal{F}$, and v_i=$\langle n_{i1}, n_{i2}, \dots, n_{ik} \rangle$, where $k = |X|$, and $n_{ij}, j = \overline{1,k}$ are numeric values corresponding to the positions of items in the database event sequences. Let the sorted sequence

of distinct values in v_i be $n_{ia_0} < n_{ia_1} < \cdots < n_{ia_l}$, where $l \leq k-1$. We apply normalization of v_i by substitution σ of all occurrences of n_{ia_q} by q, $0 \leq q \leq l$.

The prefix tree in Fig. 4 is transformed to the tree in Fig. 6 after the normalization. For the itemset $\{A, B\}$ we have support 2 only for both patterns [0,0] and [1,0], hence this set will be pruned. The resulting frequent patterns are listed in Fig. 7, where parallel (simultaneous) events are presented in brackets and the leftmost event is the start, and rightmost event is the final event in the sequence.

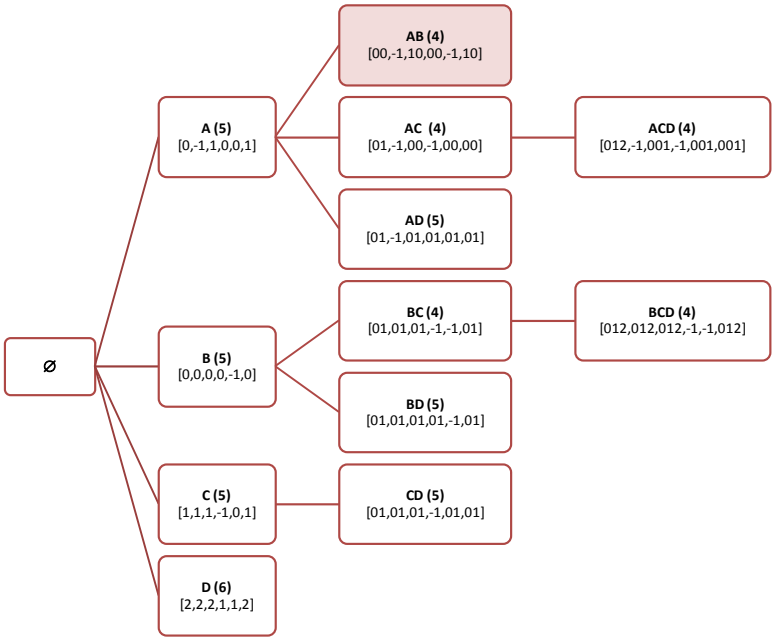


Fig. 6. Prefix tree with patterns mapped vectors for $minsup = 3$. Shaded boxes indicate infrequent itemsets.

Sequence pattern	support
$[AC]$	3
$[AC]D$	3
AD	5
BC	4
BCD	4
BD	5
CD	5

Fig. 7. Sequence patterns. Parallel events are presented in brackets.

5 Experiments and Results

To cope with big data, we use tabulation in the implementation of both methods. Thus each level of the tree is stored in separate table as a file. During the tree generation we deal only with the last generated table and the table corresponding to the level 1. For the projection we need to store in the memory only the vectors for three nodes.

We applied both methods for each collection separately and obtained the following results for chronic diseases and treatment (Table 4). These methods are applied also for age and gender specific constraints for each collections. The figures below present the extracted frequent patterns for chronic diseases in total and for age 15-44 years, with $minsup = 80$ for S1 (Fig. 8), and $minsup = 45$ for S2 (Fig. 9). For both collections the minimal support is chosen as 0.5% of the number of patients, respectively.

Table 4. Characteristics of structured event data

Collections / Characteristics	S1		S2	
	Total	**Age 15-44**	**Total**	**Age 15-44**
Total number of extracted ICD-10 codes	782,448	288,625	248,067	198,588
Total number of distinct ICD-10 codes	5,790	4,530	4,697	4,240
Total number of extracted chronic diseases	107,789	36,180	31,151	22,194
Total number of distinct chronic diseases	227	216	228	215
Total number of patient with chronic diseases	37,921	16,059	8,414	6,933
Avg length *k* of chronic diseases sequence	2.843	2.253	3.702	3.201

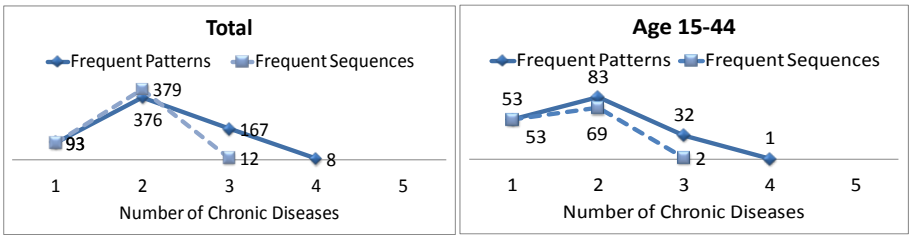


Fig. 8. Total number of extracted frequent patterns and sequences for collection S1

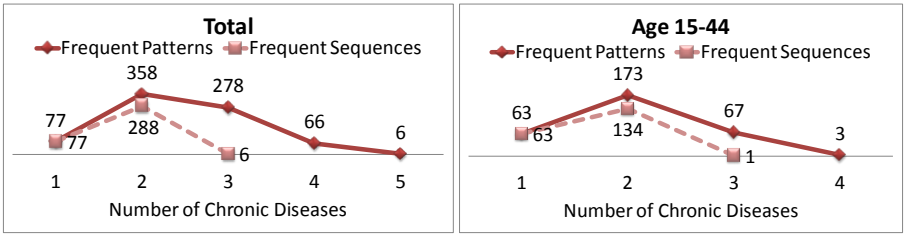


Fig. 9. Total number of extracted frequent patterns and sequences for collection S2

We also experimented with different window lengths – 1 month and 3 months. These intervals are with high importance for healthcare, because they denote the minimal and the optimal interval for which the medication treatment can show effect. In our experiments for window 1 month the results are almost similar to the results for window 0, because the frequency of patient visits usually exceeds 1 month. For window 3 months the obtained results decline dramatically, because of the rather short period of our retrospective study – only 3 years. And therefore the maximal size of the frequent sequence is 2 chronic diseases only.

Together with a bunch of well known associations in each collection we found some patterns that are not well studied in the medical literature and further study will be needed to investigate their significance in healthcare management.

5.1 Case Study

Initially we started processing the collection $S3$ for patients with Diabetes Mellitus Type 2 (T2D – E11) and extracted relatively high number of frequent patterns containing different mental disorders – ICD10 codes F00-F99. This result motivated us to process collection $S1$ for patients with Schizophrenia (SCZ – F20). Mining complex itemsets including both chronic and acute diseases we obtained as a result high frequency of some acute diseases as well. And we included in our study collection $S2$ for patients with Hyperprolactinemia (E22.1) to investigate in more details the association of these three diseases (Table 1).

On the other side SCZ develops in relatively young age 18-35 years and the peak ages of onset are 25 years for males and 27 years for females. In contrast T2D develops relatively later and the peak ages of onset are 45-65 years. This motivates us to study in more details age specific event sequences (Table 5). These co-morbidity facts together with the administrated medications were interpreted as temporal events and the event sequences were processed by the mining tool for pattern search. The data are extracted from more than 8 millions ORs (Table 1).

Table 5. Statistics for collections S1, S2 and S3

Year	**2012**		**2013**		**2014**	
ICD10	**Total**	**Age 15-44**	**Total**	**Age 15-44**	**Total**	**Age 15-44**
Total	5,048,165	1,583,980	5,132,403	1,598,595	5,147,648	1,585,024
F20	38,560	16,226	38,464	15,871	37,921	15,241
E11	403,048	18,704	419,237	19,304	432,705	19,200
E22.1	3,663	3,018	5,273	4,269	5,347	4,267
E11+F20	2,775	505	3,093	538	3,209	576
F20+E22.1	158	73	251	107	237	183
E11+E22.1	271	120	472	206	534	231
F20+E22.1 +E11	19	7	33	18	30	20

It is well known that patients with SCZ are at an increased risk of T2D [19], therefore a better understanding of the factors contributing to T2D is needed. SCZ is often

treated with antipsychotic agents but the use of antipsychotics has been associated with Hyperprolactinemia, or elevated prolactin levels (a serious hormonal abnormality). Thus, given the large repository of ORs, that covers more than 5 million citizens of Bulgaria, it is interesting to study associations and dependencies among SCZ, T2D and Hyperprolactinemia in the context of the treatment prescribed to the patients.

Regarding the treatment it is well known that the classical antipsychotics, blocking D2 dopamine receptors, lead to extrapyramidal effects related to antagonism in the nigrostriatal pathway, and Hyperprolactinemia is due to antagonism in the tuberoinfundibular pathway. In the early 1990s a new class of antipsychotics was introduced in the clinical practice with the alleged advantage of causing no or minimal extrapyramidal side effects, and the resulting potential to increase treatment adherence. However, there are data, that some of these antipsychotics can induce Diabetes, Hyperlipidaemia and weight gain.

Our study considers the presence of:

- Hyperprolactinemia in the patients with Schizophrenia,
- T2D and Schizophrenia in the patients with Hyperprolactinemia,
- T2D and Hyperprolactinemia in the patients with Schizophrenia and
- T2D in the patients with Schizophrenia and Hyperprolactinemia.

We also combine diagnosis patterns with treatment sequences patterns that include first generation of antipsychotics (ATC codes: N05AF01, N05AB02, N05AD01, N05AF05) and/or second generation of antipsychotics (ATC codes: N05AL05, N05AX12, N05AH02, N05AH03, N05AX13, N05AH04, N05AX08, N05AE03, N05AE04).

Table 6. Extracted sequence patterns with $minsup = 3$ from S1

Sequence Pattern	**Total**	**Age 15-44 years**
$F20 \rightarrow E11 \rightarrow E22.1$	19	6
$F20 \rightarrow E22.1 \rightarrow E11$	11	6
$F20 \rightarrow [E22.1, E11]$	21	8
$[F20, E11] \rightarrow E22.1$	8	4
$[F20, E11, E22.1]$	5	3
$F20 \rightarrow E22.1$	314	264
$[F20, E22.1]$	25	22
$E22.1 \rightarrow F20$	27	23
$E11 \rightarrow F20$	785	142
$[F20, E11]$	1,512	231
$F20 \rightarrow E11$	2,201	491

We found an increased rate of Hyperprolactinemia and T2D in patients with Schizophrenia, compared to presence of these diseases in patients without Schizophrenia. Table 6 presents extracted frequent sequence patterns for SCZ (F20), T2D (E11) and Hyperprolactinemia (E22.1). We can observe that in these sequences dominates the relation SCZ (F20) precedes T2D (E11). In Fig. 10 is presented the temporal relation

between the onset of T2D (E11) after the onset of SCZ (F20) measured in months. Table 7 shows the prevalence of T2D (E11) in the entire population and among patients suffering from SCZ (F20). Although there is no significant difference for the total collection, the statistics for age 15-44 years show that for patients with SCZ (F20) there is about three times higher risk for development of T2D (E11). The finding is explicated in relation with the administrated treatment.

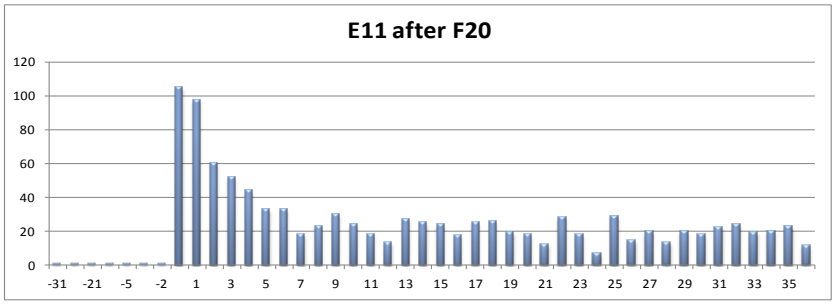
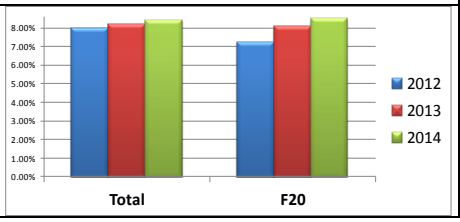
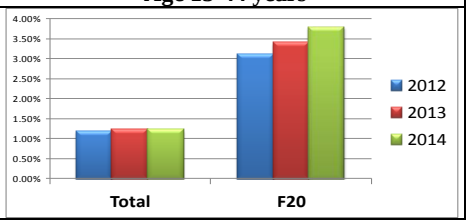


Fig. 10. Total number of patients with E11 after diagnosis F20, grouped by period measured in months

Table 7. Prevalence of E11

Total	Age 15-44 years
	

6 Conclusion and Further Work

The proposed approach presents an elegant and efficient solution for the task of frequent patterns and sequences mining in the scope of big data applied over real medical datasets. Due to the constraints for distinct diagnoses in our patterns the proposed solution allows to reuse the frequent patterns tree generated for the Task 1 for frequent sequences mining. This is due to the fact that the tree from Task 1 can only shrink on Task 2, i.e. not all possible permutations of the itemsets have sufficient support above the *minsup*. It also allows checking different hypotheses for treatment effect and risk factors, by using different size for search window. In contrast to the other classical methods for frequent sequence mining the proposed solution allows parallel (simultaneous) events in sequences to be recognized and grouped. This ap-

proach can be applied also for other domains where heterogeneous events need to be mined.

In contrast to other approaches for temporal events mining in EHR the proposed approach is not so expensive like EMRView [1, 2] which requires offline preliminary generation of frequent patterns over which later the system applies filtering and projection. Also we process the entire collection with no limitation like interactive systems [5] that reduce the task by initially selecting events of interest for analysis and visualization queries or sentinel events [3].

The main advantage of the proposed approach is that it takes into account the data specificity and enables flexible parameterization of *(i)* the set of the diagnoses in the research interest, and *(ii)* the time window size.

The contribution of the paper is that it demonstrates a powerful, sufficiently generalized technology for discovering correlations in the medical domain.

Our future plans include experiments with collections of ORs for patients with Cardio-vascular Disorders and Malignant neoplasm. We also plan to process the complete sequence of diseases – including both chronic and acute diseases to investigate more complex causalities.

Integrating mining of status data for patients will further elucidate the risk factors and causality for some acute and chronic disorders. Moreover, changing patient states brings new events in the whole picture and hopefully longer sequences will be revealed that will point to new interesting medical facts.

Acknowledgements

The research work presented in this paper is 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 authors also acknowledge the support of the Bulgarian Health Insurance Fund, the Bulgarian Ministry of Health and Medical University – Sofia for providing the experimental data.

7 References

1. Patnaik, D., L. Parida, P. Butler, B. J. Keller, N. Ramakrishnan, and D. A. Hanauer. Experiences with Mining Temporal Event Sequences from Electronic Medical Records: Initial Successes and Some Challenges. In *Proc. 17th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'11)*, San Diego, August 2011, pp. 360-368.
2. Patnaik, D., S. Laxman, B. Chandramouli, and N. Ramakrishnan. Efficient Episode Mining of Dynamic Event Streams, in *Proc. of the IEEE Int. Conf. on Data Mining (ICDM'12)*, Brussels, Belgium, December 2012, pp. 605-614.
3. Wang, T., C. Plaisant, A. J. Quinn, R. Stanchak, S. Murphy, and B. Shneiderman. Aligning temporal data by sentinel events: discovering patterns in electronic health records. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '08), ACM, New York, NY, USA, 2008, pp. 457-466.

4. Gotz, D., Fei Wang, and A. Perer. A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data. *Journal of Biomedical Informatics*, Vol. 48, April 2014, pp. 148-159.
5. Rind, A., Taowei David Wang, W. Aigner, S. Miksch, K. Wongsuphasawat, C. Plaisant, and B. Shneiderman. Interactive Information Visualization to Explore and Query Electronic Health Records, *Journal of Foundations and Trends® in Human–Computer Interaction* 5(3), 2013, pp. 207-298.
6. Lee, N., A.F. Laine, Jianying Hu, Fei Wang, Jimeng Sun, and S. Ebadollahi. Mining electronic medical records to explore the linkage between healthcare resource utilization and disease severity in diabetic patients. *Proc. First IEEE Int. Conf. on Healthcare Informatics, Imaging and Systems Biology (HISB)*, 2011, pp. 250 – 257.
7. Zaki, M. J., and Meira Wagner Jr. *Data Mining and Analysis: Fundamental Concepts and Algorithms.* Cambridge University Press, 2014.
8. Sun, X., M. Orlowska, and X. Zhou. Finding Event-Oriented Patterns in Long Temporal Sequences, in *Proc. 7th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2003)*, Seoul, Korea, April 2003, Springer LNCS 2637, pp. 15-26.
9. Guyet, T., and R. Quiniou. Mining temporal patterns with quantitative intervals. In Zighed D., Z. Ras, and S. Tsumoto (Editors): *Proc. of the 4th Int. Workshop on Mining Complex Data, IEEE ICDM Workshop*, 2008, pp. 218-227.
10. Guyet, T., and R. Quiniou. Extracting temporal patterns from interval-based sequences. In *Proc. 22nd Int. Joint Conference on Artificial Intelligence*, 2011, pp. 1306-1311.
11. Monroe, M., Rongjian Lan, Hanseung Lee, C. Plaisant, and B. Shneiderman. Temporal Event Sequence Simplification, in *IEEE Transactions on Visualisation and Computer Graphics*, 19(12), December 2013, pp. 2227-2236.
12. Yang, J., J. McAuley, J. Leskovec, P. LePendu, and N. Shah. Finding progression stages in time-evolving event sequences. In *Proc. of the 23rd international conference on World wide web* (WWW '14). ACM, New York, NY, USA, 2014, pp. 783-794.
13. Chronic diseases, WHO, http://www.who.int/topics/chronic_diseases/en/
14. International Classification of Diseases and Related Health Problems 10th Revision. http://apps.who.int/classifications/icd10/browse/2015/en
15. Nikolova, I., D. Tcharaktchiev, S. Boytcheva, Z. Angelov, and 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* 8722, Springer,2014, pp. 92–103.
16. Boytcheva, S. Shallow Medication Extraction from Hospital Patient Records. In: Koutkias, V., J. Nies, S. Jensen, N. Maglaveras, and R. Beuscart (Eds.), *Studies in Health Technology and Informatics*, Vol. 166, IOS Press, 2011, pp. 119-128.
17. Anatomical Therapeutic Chemical (ATC) Classification System, http://atc.thedrugsinfo.com/
18. Boytcheva, S., G. Angelova, Z. Angelov, and D. Tcharaktchiev. Text Mining and Big Data Analytics for Retrospective Analysis of Clinical Texts from Outpatient Care. *Cybernetics and Information Technologies*. Volume 15, Issue 4, November 2015, pp. 58–77.
19. Marder, S., S. Essock, A. Miller, R. Buchanan, D. Casey, J. Davis, J. Kane, J. Lieberman, N. Schooler, N. Covell, S. Stroup, E. Weissman, D. Wirshing, C. Hall, L. Pogach, X. Pi-Sunyer, JT Jr Bigger, A. Friedman, D. Kleinberg, S. Yevich, B. Davis, and S. Shon. Physical health monitoring of patients with schizophrenia. *American Journal of Psychiatry*. 161(8), August 2004, pp. 1334-49.