Health Informatics Meets eHealth
D. Hayn and G. Schreier (Eds.)

doi:10.3233/978-1-61499-759-7-211

Frequent Treatment Sequence Mining from Medical Databases

Krisztina TÓTH[a], István KÓSA[b] and Ágnes VATHY-FOGARASSY[a,1]
[a]*Department of Computer Science and Systems Technology, University of Pannonia, Hungary*
[b]*Department of Electrical Engineering and Information Systems, University of Pannonia, Hungary*

Abstract. The huge amount of data stored in healthcare databases allows wide range possibilities for data analysis. In this article, we present a novel multilevel analysis methodology to generate and analyze sequential healthcare treatment events. The event sequences can be generated on different abstraction levels automatically from the source data, and so they describe the treatment of patients on different levels of detail. To present applicability of the proposed methodology, we introduce a short case study as well, in which some analysis results are presented arising from the analysis of a group of patients suffering from colorectal cancer.

Keywords. Data Mining, Information Extraction, Clinical Practice Patterns, Medical Oncology, Colorectal Tumors

1. Background

Sequential pattern mining is a widespread data mining technique, which aims to unfold statistically relevant repeating sets of data from sequential databases [1]. Since in the healthcare sector every action, like treatments, diagnosis, prescription of medications, occurs time to time, the available healthcare databases contain a plethora of event sequences. This gives a fundamental basis to various research works in the field of sequential pattern mining in healthcare to find potentially new medical knowledge or even to predict future occurrence of certain events. Accordingly, in medical research reports, we can find numerous application examples for sequential pattern mining. For example, sequential pattern mining has been applied in hepatitis type detection [2], examination of chemotherapy and targeted biologics sequence of metastatic colorectal cancer [3], analysis of influencing factors of glioblastoma survivals [4], and in the prediction of next prescribed medication [5]. The discovered sequences can be compared with the care processes of medical guidelines [6] or with the practice of other hospitals [7].

All these research works aim to discover sequentially repeated data, but the applied methods vary significantly. Some of them analyze similar events [8], while others are based on the analysis of different type of events (e.g. symptoms, test results, treatments) [9,10,11]. The main questions of these methods are: (1) what kind of events are presented in the event sequences, (2) how is the duration of the considered events presented and (3) how are the parallel events (e.g. treatments) handled.

[1] Corresponding Author: Ágnes Vathy-Fogarassy, Department of Computer Science and Systems Technology, University of Pannonia, 2. Egyetem Str., 8200 Veszprém, Hungary, E-Mail: vathy@dcs.uni-pannon.hu.

Event sequences are typically extracted from medical databases by writing complex database quires. A different way to unfold and analyze treatment event sequences is to use process mining techniques. Process mining algorithms (e.g. alpha algorithm or HeuristicMiner) aim to create flow diagrams of processes automatically from log files or databases. As medical treatments can be considered as the events of a process, these algorithms may also provide effective assistance to discover sequences from medical databases. However, if these algorithms are applied, special attention should be paid to their shortcomings and limitations (e.g. the problem of short loops).

The goal of our research is to create a hierarchical data analysis method to generate and discover health care sequences and sequential patterns from medical databases. In our work, only healthcare treatments are considered as events and event sequences are generated from relational databases using database queries. The main contribution of this research work is that the event sequences can be represented at different levels of detail. This is achieved by a hierarchical code system for treatments and a discipline of aggregations, which reflect professional medical practice. As a result, healthcare events can be analyzed on a very detailed low-level, where the diversity is extreme, or they can also be summarized in higher levels, resulting in a considerable reduction in diversity of patient pathways. The highest level of the abstraction emphasizes only the main characteristics of the medical treatments and it can be generated from the lowest level automatically. The proposed method facilitates the comparison of the specificity of treatments of healthcare institutes, and the exploration of frequent or rare treatment sequences for given diseases.

The remaining part of this work is organized as follows. Section 2 introduces the suggested data analysis methodology from theoretical aspects and Section 3 presents a short case study of a possible application. Finally, Section 4 concludes the paper.

2. Methods

The first step of the generation of health care sequences is to determine the cohort of patients to be examined. After the selection of the population, we can generate a detailed event sequence for every patient separately, including all the relevant events of the considered health care. Then, the event sequences can be automatically transformed into a higher, less detailed level. Higher levels contain less specific information about the treatments hereby highlighting their main characteristics. The first two level of the event sequences can be beneficial in clinical practice, while the second two levels are useful in the statistical analysis of health care patterns.

In the next chapters, we introduce these different abstraction levels through a real-life example, which represents the treatment sequence of a patient with colorectal cancer.

2.1. Treatment Event Sequences

2.1.1. Level 1: Detailed Event Sequences

The event sequences in the lowest level of abstraction contain all treatments for the whole examination period, indicated by their International Classification of Procedures in Medicine (ICMP) code. These events are displayed in chronological order, provided with a relative timestamp, which correlates to the appearance of the first event (in this example to the histological confirmation of cancer).

Table 1. Example of a detailed event sequence

Patient	*Event sequence*
p1	0:29000||0:14500||0:16410||16:54551||16:54688||16:55431||24:29000||24:29050||78:70451|| 93:70451||107:70451||126:70451||142:70451||156:70451||171:70451||185:70451||329:X0000||

In Table 1, the detailed event sequence of patient *p1* can be seen. The events are isolated by a separation character (||), and all of them are built up from two parts: the timestamp and the code of the treatment, separated by a colon.

On the zeroth day, the patient had three kinds of treatments: a labor diagnostic procedure (ICMP: 29000: histology) and two clinical diagnostic procedure (14500: Biopsy Intestini Crassi per Colonoscopies and 16410 Colonoscopy). These events are connected to the first detection of the colorectal cancer of the patient. 16 days after the detection, the patient underwent an operation, which was depicted in the code system by three surgical ICMP codes (54551: Haemicolectomia dextra, 54688: Adhaesiolysis interintestinalis and 55431: Resectio omenti maioris). The histological samples from this operation where evaluated on the 24th day, which activity appears in the coding history of the patient with the code of 29000 (Histology) and code 29050 (Immunhistochemy). From the 78th until the 185th day, the patient underwent eight chemotherapies. Finally, the event sequence ends with the death of the patient on the 329th day. As death does not have an ICMP code, we extended the list of ICMP codes with X0000 to represent the death of patients.

2.1.2. Level 2: Typified Event Sequences

To extract typical patterns, the excessive amount of details of the events appearing in the low-level sequences have to be eliminated. This requires to cluster similar events and to organize them into a hierarchy. Clustering assigns the same code to similar events in the event sequence, while a hierarchy of clusters gives a basis to generate different levels of abstraction of treatments automatically. A sample hierarchy used during our research work is presented in Figure 1.

In the hierarchy system, every group gets an identifying character (presented in brackets after the name of the group), and the original ICMP codes are mapped into a four-character typified code system. In this novel code system, every character stands for a unique level of the hierarchy. For example, in the presented example the group of colon surgery events is a third level group, it contains 44 different ICMP codes, and all these ICMP codes are translated into the 4-digit code 'MLB0'.

By applying this code mapping procedure, we can generate the typified event sequence for each patient. If the novel hierarchical code of two treatments, which are on the same day, are equal, than only one hierarchical code will appear in the typified event

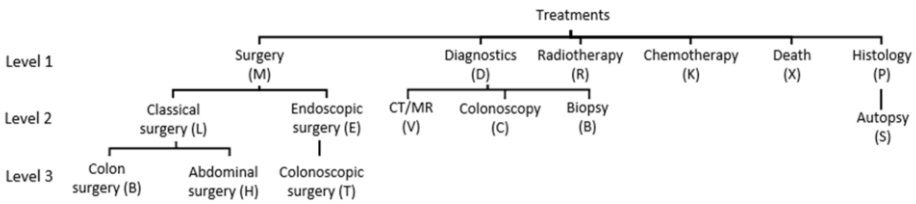

Figure 1. Hierarchy of treatments

Table 2. Example of a typified event sequence

Patient	*Event sequence*
p1	0:P000||0:DC00||16:MLB0||16:MLH0||24:P000||78:K000||93:K000||107:K000||126:K000|| 142:K000||156:K000||171:K000||185:K000||329:X000||

sequence. The result of this process for patient *p1* can be seen in Table 2. On the zeroth day of patient *p1,* we can see a histology, represented with the code P000, and only one procedure, which is coded by DC00. We have to mention, that in this case in the original event sequence there were two clinical diagnostic procedures recorded with different ICMP codes, but both were mapped into DC00. The three surgical procedures on the 16th day are mapped into two groups: colon surgery (MLB0) and abdominal surgery (MLH0). The next two histologies in the original sequence get the same code, thus they appear as one event in the new sequence (P000). Afterward, the chemotherapies appear with the code K000, and finally, the code of death is displayed.

With the help of the typified code system, the variability of the event sequences decreases. For example, in our research work, we have selected 1436 different ICMP codes for the examination of colorectal cancer, and after the code mapping procedure, these codes resulted in 11 different 4-digit codes. Of course, the reduction of the number of codes decreases the variability of event sequences as well.

The typified event sequences give a proper basis to execute aggregation operations, which we present in the next chapter.

2.1.3. Level 3: Aggregated Event Sequences

During the medical care, patients have to undergo the same treatments time to time. Patients diagnosed with carcinoma, usually have a series of chemotherapy or radiotherapy, but the number of these treatments are different from patient to patient. Therefore, these events significantly increase the number of different event sequences. The aggregation of subsequences can help to decrease the number of similar but yet different typified event sequences.

The essence of this aggregation is that from the repeating events only the first one will appear in the event sequence. In the example of Table 2, there were consecutively eight chemotherapies, which can be aggregated into one, thus only the first one appears in the sequence, as shown in Table 3.

Table 3. Example for the aggregation of consecutive events

Patient	*Event sequence*
p1	0:P000||0:DC00||16:MLB0||16:MLH0||24:P000||78:K000||329:X000||

Additionally, we can enhance the representation of the aggregated events with a time interval, which interval shows the number of days between the first and the last aggregated events (Table 4).

Table 4. Example for the enhanced aggregated event sequence

Patient	*Event sequence*
p1	0:P000||0:DC00||16:MLB0||16:MLH0||24:P000||78:K000:107||329:X000||

Table 5. Example for the aggregation of similar events with the same timestamp

Patient	*Event sequence*
p1	0:P000||0:DC00||16:MLB0||24:P000||78:K000:107||329:X000||

Patients can undergo more than one similar event in one day during the medical care. Regularly, these treatments are connected to the same event, thus it is enough if only the most relevant treatment appears in the event sequence. To define which treatment should be the one that appears in the sequence, hierarchy rules have to be defined. If there are more than one event with the same timestamps in the typified event sequence of a patient, only the event with the highest priority will appear in the aggregated event sequence.

For example, if we define a rule for colon surgeries and abdominal surgeries, so that the colon surgery (MLB0) has a higher priority than the abdominal surgery (MLH0), then in the case of patient *p1* only the colon surgery will appear in the sequence, as it can be seen in Table 5.

With the help of these two aggregation methods, the number of the different event sequences are effectively reduced without the loss of valuable information.

2.1.4. Level 4: Health Care Patterns

The previously presented event sequences are still too detailed for different statistics. For this reason, we introduced the health care patterns: the multi-character codes are replaced by a one character code, and the timestamps are omitted from the events. Owing to this replacement, the resulted patterns are not as detailed as the typified and aggregated event sequences, but they give a much more general overview of the care sequence of patients.

In the example below, the health care pattern of patient *p1* can be seen. The pattern was made by keeping the last relevant code from events presented in the aggregated event sequence and timestamp and separation characters were omitted. The pattern in Table 6 describes with a simple code system that the patient had a histology, a colonoscopy, a colon surgery, another histology, chemotherapy and then the patient died.

Table 6. Example of the pattern of patient *p1*

Patient	*Event sequence*
p1	PCBPKX

3. Results

3.1. Financial Database of the Healthcare Sector

To present the applicability of the previously introduced method we have analyzed the healthcare data of oncology patients in Hungary. As the financial healthcare database in Hungary includes data from all healthcare providers, our study was performed on this database. This database contains information about the diagnosis and the medical treatments of patients related to the service providers. The identification of the diagnosis is defined by the International Classification of Diseases (ICD) codes and diagnostic procedures and treatments are recorded by the ICPM code system applied in Hungary.

3.2. Case Study

The cohort of patients was selected as follows. Those cancer patients were selected into the study, which had been diagnosed and treated with primer colorectal carcinoma between 1st January 2009 and 31st December 2014. To identify colorectal carcinoma, we used the ICD blocks of C18, C19 and C20. These blocks were separated to two subgroups: rectum (C20H0) and colon carcinoma (C18-C19). Additionally, the inclusion criteria for the study has specified, that the fact of colorectal cancer had to be verified with histology. The date of verification was set as the origin of the examination (0th day). Patients with recidive tumor were excluded from the analysis. According to these criteria 28817 patients, 16575 men and 12242 women were enrolled into the analysis, whose average age is 70 ± 11 years.

For the analysis, we have developed a user-friendly application in Java programming language. This application provides a wide range of graphical possibilities to generate and analyze the treatment sequences and patterns on different abstraction levels. With the help of this application, the event sequences for each enrolled patients were generated on all four levels of detail automatically. The consecutive radio- and chemotherapies were aggregated with the method described in Section 2.1.3, and for the events with the same timestamp the following hierarchy was defined (Equation 1):

$$colonoscopic\ surgery\ (E) \rightarrow abdominal\ surgery\ (H) \rightarrow colon\ surgery\ (B) \qquad (1)$$

In most cases, the event sequences were examined from the origin until a maximum of 365 days. In our analyses, the event sequences were closed in the case of a 180 days long event-free period. We considered this 180-day-long period as the stabilization of the condition of the patients, and handled the events after this period as they were not involved in the original treatment sequence but happened because of the change in the condition of the patients.

As the financial healthcare database contains information about the healthcare providers, like the code of the institute, the postal code, and medical provider code we were able to make some geographical and institutional analysis as well. The patients and their event sequences were assigned to the institutes based on the place of the record of pathological diagnosis and we considered these institutes as the primary healthcare facilities of the patients.

We analyzed the distribution of the health care patterns by grouping them by institutes and patterns. For example, Table 7 shows the distribution of the ten most frequent patterns for the top ten supplier institutes in case of colon carcinoma. Rows represent the institutes and columns the treatment patterns truncated for the first three treatment events. As it can be seen, colon surgery (B) and chemotherapy after colon surgery (BK) were the most frequent applied therapies in all institutes, but the frequencies of these patterns are different.

To visualize the frequencies, we have generated the box-plot diagram of the patterns (Figure 2). As it can be seen, the frequency of occurrence of the event sequences shows a significant deviation for different institutes. For example, the rate of the event sequences, which contains only chemotherapies, varies between 2.53% and 65.06%.

During our analysis, we have established, that in general, the number of the patients treated in the institutes shows a very weak correlation with the frequency of the different event sequences. It is our suspicion that the frequency of event sequences ending with

Table 7. The distribution of patients' health care patterns diagnosed with colon carcinoma (in percentage)

Institutes/ Patterns	B	BK	K	H	E	BH	EB	EBK	BB	BKE	Sum
1	34,01	42,13	10,41	2,79	1,90	2,92	1,78	1,65	1,65	0,76	100
2	34,83	37,31	5,11	2,94	3,87	3,25	3,72	5,11	2,48	1,39	100
3	22,48	34,87	28,14	5,13	1,77	1,95	1,06	0,53	0,88	3,19	100
4	25,94	39,53	20,21	2,86	1,61	1,61	2,86	3,76	1,07	0,54	100
5	35,52	33,33	8,38	4,55	6,19	2,37	4,19	3,28	0,91	1,28	100
6	36,82	25,98	13,83	2,62	6,92	3,36	5,05	2,24	2,43	0,75	100
7	21,36	28,95	29,16	3,08	4,52	1,23	3,90	4,72	0,62	2,46	100
8	30,90	40,13	9,44	1,50	5,15	2,36	4,08	3,43	1,50	1,50	100
9	46,08	30,18	5,07	4,15	2,53	3,69	2,07	1,84	2,07	2,30	100
10	33,73	27,83	8,73	4,25	6,37	1,65	6,84	6,37	2,59	1,65	100

only chemotherapy (K) is in reverse correlation with the size of the institute. While moving in the same direction, the number of colon surgeries followed by chemotherapies (BK) are slightly increasing. In contrast with this, the examination of the event sequences of the patients, diagnosed with rectum carcinoma shows a clearer correlation. On the one hand, the event sequences containing only a colon surgery (B) were determinative, but this dominancy declines by the increase of the volume of the institute. On the other hand, the rate of all the event sequences containing radiotherapy (RBK, RB and R) increases by the increase of the volume of the institutes.

4. Conclusion

In this paper, we presented a novel hierarchical treatment sequence analysis method. The proposed method allows the generation of treatment sequences from healthcare treatment events on four abstraction levels automatically. These abstraction levels hide the details of the medical treatment processes. The first level contains all specific information about the treatments of patients. The second level classifies the healthcare events into categories and the third level performs two different aggregations on the classified events. Finally, the fourth abstraction level hides all specific information and emphasizes only the main characteristic of the treatment process.

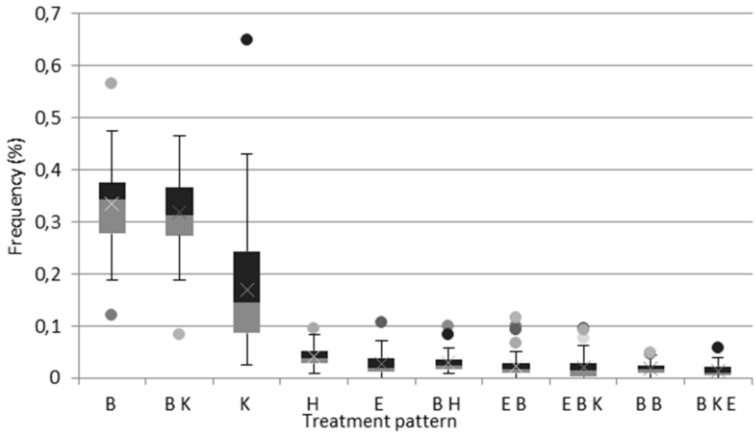

Figure 2. Treatment patterns and frequencies for the colon carcinoma

We have applied this analysis method on a national-wide healthcare database in Hungary. In this study patients suffering from colorectal cancer were analyzed. As the utilization possibility of the hierarchical treatment sequence analysis is very wide, we have introduced only a small part of it. Analysis results confirm the usefulness of the presented method.

Acknowledgement

This publication has been supported by the Hungarian Government through the project VKSZ 12-1-2013-0012 – Competitiveness Grant: Development of the Analytic Healthcare Quality User Information (AHQUI) framework.

5. References

[1] N. R Mabroukeh, C. I. Ezeife, A taxonomy of sequential pattern mining algorithms, ACM Computing Surveys (CSUR) 43(1):3 (2010)

[2] S. Aseervatham, A. Osmani, Mining short sequential patterns for hepatitis type detection, Springer (2005), 6

[3] R.C. Parikh, X.L. Du, R.O. Morgan, D.R. Lairson, Patterns of Treatment Sequences in Chemotherapy and Targeted Biologics for Metastatic Colorectal Cancer: Findings from a Large Community-Based Cohort of Elderly Patients, Drugs - Real World Outcomes 3(1) (2016), 69–82.

[4] K. Malhotra, D. H. Chau, J. Sun, C. Hadjipanayis, S. B. Navathe, Constraint Based Temporal Event Sequence Mining for Glioblastoma Survival Prediction, Journal of Biomedical Informatics 61 (2016), 267-275.

[5] A. P. Wright, A. T. Wright, A. B. McCoy, D. F. Sittig, The use of sequential pattern mining to predict next prescribed medications, Journal of Biomedical Informatics 53 (2015), 73-80.

[6] F. Caron, J. Vanthienen, K. Vanhaecht, E. Van Limbergen, J. Deweerdt, B. Baesens, Monitoring care processes in the gynecologic oncology department, Computers in Biology and Medicine 44 (2014), 88-96

[7] R. Mans, H. Schonenberg, G. Leonardi, S. Panzarasa, A. Cavallini, S. Quaglini, W. van der Aalst, Process mining techniques: an application to stroke care, Studies in Health Technology and Informatics 136 (2008), 573-578

[8] K. Wongsuphasawat, D.H. Gotz, Outflow: Visualizing Patient Flow by Symptoms and Outcome, IEEE VisWeek Workshop on Visual Analytics in Healthcare 3 (2011), 25-28.

[9] E. Roorda, Exploring Patient Data - Getting insight into treatment processes with data mining techniques, Vrije Universiteit Amsterdam, 2009

[10] C. Plaisant, R. Mushlin, A. Snyder, J. Li, D. Heller, B. Shneiderman, LifeLines: Using Visualization to Enhance Navigation and Analysis of Patient Records, Proc AMIA Symposium (1998), 76-80.

[11] K. Wongsuphasawat, J. A. Guerra-Gomez, C. Plaisant, T. D. Wang, B. Shneiderman, LifeFlow: Visualizing an Overview of Event Sequences, CHI '11 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2011), 1747-1756.