2019 IEEE 21st Conference on Business Informatics (CBI)

Application of Modern Data Analysis Methods to Cluster the Clinical Pathways in Urban Medical Facilities

Elizaveta S. Prokofyeva
Faculty of Business and Management
National Research University Higher School of Economics
Moscow, Russia
eprokofeva@hse.ru

Roman D. Zaitsev
Department of aerophysics and space research
Moscow Institute of Physics and Technology, MIPT
Moscow, Russia
roman.zaitsev@phytech.edu

Svetlana V. Maltseva
Faculty of Business and Management
National Research University Higher School of Economics
Moscow, Russia
smaltseva@hse.ru

***Abstract*— Patient flow modeling in healthcare plays a large role in understanding the operation of the system and its characteristics. Besides, modeling techniques can significantly improve the effectiveness of the medical facilities. The existing level of automation in these facilities enables the accumulation of large amounts of various data. Therefore, the collected data might be considered as the resource of new valuable knowledge. A novel approach to automatically identify the groups of similar clinical pathways based on event hospital data is presented in the paper. More specifically, the approach summarizes the most frequent pathways by implementing hard and soft clustering algorithms in order to describe the behavior patterns. The obtained clusters of clinical pathways serve as a starting point for the development of a personalized approach in modelling the heterogeneous patient flow in urban medical facilities. The results indicate the suitability of multidimensional time series clustering and Additive Regularization of Topic Models (ARTM) for the clinical event data.**

Keywords*— *data analysis, hierarchical clustering, clinical pathways, topic models, process mining, healthcare

I. Introduction

Currently, electronic healthcare systems are considered by many medical institutions as the main data source. The understanding of the patient's trajectories between medical units and specialists within these units allows the management team to plan the resource utilization, to ensure a high level of accessibility of services, and to optimize the work of their organization.

One of the examples of the healthcare systems is "United Medical Information and Analytical System" of Moscow (EMIAS), which was launched in 2011. EMIAS information system includes booking the hospital visits, EHR (Electronic Health Record) management, attachment the patient's personal record to the polyclinic, and electronic prescription services. The main purpose of the system is to increase the quality and access of medical services in public health clinics. In order to achieve the high level of availability, it is important to make a comprehensive analysis of patient traces in Moscow.

EMIAS information system frees doctors from filling endless papers and certificates. Besides, local doctors are trained to work on the computer and system equipped workplace with modern technology. What is more, this information system ensures the doctor access to patient information and help to make informed decisions on the appointment of treatment. The collected data in information systems such as EMIAS could bring more insights for a better decision making.

Processes in the healthcare sector are stochastic in nature [1], therefore, the long-term resource planning in medical facilities is a challenging task. In addition, each patient trace is a unique set of events due to individual patient history, demographic characteristics, and other relevant factors.

The variability of the patient traces is also a specific characteristic of medical care in urban medical facilities, where numerous specialists are available for appointing procedure. In order to reduce the level of uncertainty about the processes in the facilities, healthcare specialists have developed a clinical pathways tool. Generally, clinical pathway means the trajectory of the patient's movements when receiving medical services in the relevant institutions. According to the source [2], clinical pathways have been introduced internationally since the 1980s. The development of clinical pathways based on real datasets of health facilities is taking into account the personal characteristics of patients and their movement between numerous specialists, reflecting the high variability of routing. The researches emphasize that clinical pathways are focused on the patient rather than on the medical facility unit, enabling more personalized care [3].

The clinical pathways include a sequence of events that correspond to the profile of the medical institution: an appointment for the therapist, laboratory tests, consultation with a specialist, and others. The authors [4] state that the identification of the patterns of clinical pathways allows one to potentially complement information about the intentions and behavior of the patient and can also serve as a basis for further analysis of patient movements. According to [5], the introduction of clinical pathways contributes to the adaptation of accepted standards of healthcare in the protocols of the selected institution. Thus, clinical pathways are valuable in solving the problems of healthcare standardization.

The researchers [6] indicated that attempts to integrate the developed clinical paths with the electronic document flow of medical institutions were made. However, the rapid growth of available data volumes revealed the necessity to automatically determine the clinical pathways of the patient based on these data. The technology of process mining [16], data mining [11], machine learning algorithms [12] and others became the solution for automatic identification of a personal clinical path.

The paper addresses the issue of automatic modeling of clinical pathways, based on available large data

978-1-7281-0650-2/19/$31.00 ©2019 IEEE
DOI 10.1109/CBI.2019.00016

volumes in medical information systems, to improve the management of heterogeneous patient flows in urban medical institutions.

Dedicated clusters will serve as a reference point for improved prediction, or for a potential recommender service for specific diagnosis.

To this end, clustering methods are applied to discover the main patterns of patients and to provide the probability distribution of each defined group for a patient. The resulting groups of patients reflect the main trajectories for further analysis and process improvement.

The paper is organized as follows. Next section "Related work" is focused on researches in the related area. Section "Methodology" represents notation and terminology, clustering approaches to explore the clinical pathways. Section "Experiments and results" is summarizing the output of the applied data analysis methods, and also proving the directions for the future works. The "Conclusion" section is the overview of the present study.

II. Related Work

The considerable amount of research papers has been devoted to the analysis of clinical pathways based on the data of medical institutions. The methodological basis of the probability theory, statistics, data mining, graph theory, semantic technologies and process mining can be applied to model the clinical pathways of patients.

A probabilistic model of the patient's clinical pathways was developed in the study [14] for the automatic detection of patterns of treatment of unstable angina and some types of cancer. The authors applied a probabilistic topic model, Latent Dirichlet Allocation (LDA) to discover clinical pathways patterns.The obtained probability distribution from LDA application allowed to extract the most meaningful features of clinical pathways patterns and to reveal objective hidden patterns of patient routes based on event log data [14].

In [1], the author conducts an extensive review of the use of other probabilistic models of clinical pathways. In particular, the author describes in detail the capabilities of the Markov chains used to model possible system states and transitions between these states.

The study [15] demonstrated the use of weighted oriented graphs for modeling the processes of various enterprises. Clinical patient paths can be conveniently represented as a set of nodes (patient states) and a set of oriented ribs (patient movements). In the source [16], the clinical pathway model is represented by a directed graph, where the nodes to the departments in the hospital. For instance, operating room, high care and a ward- a separate hospital room for inpatient treatment and care of the chosen scheme.

The research [1] also describes the use of data mining technologies to simulate patient routes to identify patterns in medical event sequences and to predict the next possible route steps. The developed forecast models promote the improvement of the resources usage and other key indicators upgrading. For example, the decision tree model was developed to predict outcomes after cardiac arrest in patients [16].

Process mining is another powerful method to constructing the clinical pathways. The purpose of process mining is to extract new knowledge about the processes from the event logs [17]. Thus, process mining discipline lies at the intersection of machine learning, data mining and business process modelling. Process mining enables the development of clinical pathways model on the real behavior of patients, their routes and the main characteristics that influence the choice of a particular trajectory.

The event log serves as the initial data type for process mining techniques. In this context, the event log is a set of patient traces, where each trace is a sequence of activities ordered by time. It is assumed that the event log contains data related to one process under study.

However, the implementation of process mining techniques on large volumes of medical event logs may produce complicated models [18], which are problematic for the interpretation. In this regard, a novel approach to cluster the clinical pathways to detect the patterns has been introduced. This approach is based on the soft topic clustering, meaning that the pathway may belong to multiple clusters. In particular, Additive Regularization of Topic Models (ARTM) is applied on healthcare event log. In terms of topic modeling, the patient trace with a sequence of various activities (registration, operation, discharge) is analogic to a document with a sequence of words [4]. The obtained clusters by ARTM are similar to clinical pathways patterns. Besides, the multidimensional time series hard clustering method is performed as the alternative way to discover patterns. The developed clusters or latent topics may serve as a support tool for the healthcare specialists to define the risk groups and to provide the take the possible preventive measures.

III. Methodology

The specific structure of clinical pathways allows the application of cluster analysis methods for time series. One way to significantly improve the accuracy of the model is the segmentation of the initial dataset into subgroups of similar objects (clusters) and the construction of separate "personalized" models for each of the groups.

Notation and terminology

The study is aimed at determining clusters of clinical pathways from event logs of medical facilities. In particular, hard and soft clustering approaches are used. Based on theoretical background in this field [23], we define the main concepts for the approach.

The set of patient's *activities* is A, while a patient *trace* is denoted as a non-empty set of activities performed by an actual patient: $\sigma = <a_1, a_2, ..., a_n>$, where $a_i \in A$ $(1 \leqslant i \leqslant n)$ is an activity executed by a patient. Let R be a set of various *resources* in medical facility, which also may be used as the grouping label of activity. For instance, a group of different activities may belong to one resource variable.

According to soft clustering method, activities are *words* in the topic model, and patient trace is considered to be a bag of activities, or a *document*. Therefore, a pattern of patient's clinical pathway is a *topic* of the model. One of the main data sources for the event logs with such data is healthcare information system.

Hard clustering approach

Ward's hierarchical agglomerative clustering method [19] was chosen as the first approach to group the clinical pathways. As the authors [19] describe Ward's method, they state that this method is based on a classical sum-of-squares criterion. The method was proposed in 1963 by Ward to minimize losses associated with the formation of clusters. To estimate the optimal number of clusters, the Silhouette coefficient was used [20]. The highest silhouette coefficient among clustering models obtained using the same distance d can be used as an optimality criterion for choosing the preferred number of clusters N, and the preferred clustering algorithm.

At the initial stage, all patient traces are encoded by replacing activities with letters of the English alphabet in order. After the initial analysis of the path length distribution, abnormally long paths were rejected to improve the quality of clustering. The upper limit of the length of the path under consideration was set to 42 events, or Q50 + 3 * (Q75-Q50), where Q50 is the median, Q75 corresponds to 75% of the quantile.

Next, a distance matrix is constructed based on the limited Damerau-Levenshtein distance, or the optimal arrangement of rows [21].

For the subsequent hierarchical cluster analysis, the Ward method was chosen, for which initially only one object is included in each cluster [22]. The method is producing groups that minimize within-group dispersion at each binary fusion [19].

The silhouette coefficient from 2 to 40 clusters was evaluated to find the optimal number of clusters [20] after the previous clustering of the clinical pathways. Suppose that the the distance *d* on the clustered set is given, and some method is applied to obtain the clusterization model. Let for each sample object *i* belonging to the cluster C_i, the value of a(i) is equal to the average distance from *i* to each of the objects *j* of the same cluster:

$$a(i) = \frac{1}{|C_i|}\sum_{j\neq i} d(i,j)\,, j \in C_i \qquad (1)$$

This value indirectly indicates how much object *i* is similar to its cluster. Further, we define a cluster C' from the set of all clusters C adjacent for a point *i,* if:

$$C' = argmin_{C_k \in C \setminus C_i}\left(\frac{1}{|C_k|}\sum_{j\in C_k} d(i,j)\right) \qquad (2)$$

Also we denote the average distance from point *i* to a neighboring cluster as b(i):

Figure 2. Silhouette coefficient depending on the number of clusters

$$b(i) = \frac{1}{|C'|}\sum_{j\in C'} d(i,j) \qquad (3)$$

Then the silhouette coefficient of the object *i* in the resulting model is determined as follows:

$$s(i) = \frac{b(i)-a(i)}{max\{a(i);b(i)\}} \qquad (4)$$

The silhouette coefficient for each object varies in the range [-1; 1] and shows how much closer the element is to its cluster than to the nearest neighbor. By averaging the silhouette coefficients of the elements, one can obtain the silhouettes of individual clusters s(C_k).

$$s(C_k) = \frac{1}{|C_k|}\sum_{j\in C_k} s(j)) \qquad (5)$$

and the overall silhouette of the clustering model s(C)

$$s(C) = \frac{1}{|C|}\sum_{C_k\in C} s(C_k) \qquad (6)$$

Soft clustering approach

Probabilistic "soft" clustering of traces on clusters of clinical patterns allows to develop more flexible approach to group the patients, providing them a probability to belong to several clusters. The approach is based on topic modelling, which is a model that represent the observed conditional distribution *p(w|d)* of words *w* in documents *d* of collection D:

$$p(w|d) = \sum_{t\in T} p(w|t)p(t|d) \qquad (7)$$

where T — set of topics, $\varphi_{wt} = p\ (w|t)$ is distribution of words in topic *t,* $\theta_{td} = p(t|d)$ is distribution of topics in document. The parameters of the topic model - the matrices $\Phi = (\varphi_{wt})$ and $\Theta = (\theta_{td})$ are found by solving the likelihood maximization problem [26]:

$$\sum_{d\in D}\sum_{w\in d} n_{dw}\log\sum_{t\in T}\varphi_{wt}\theta_{td} \to \underset{\Phi,\,\Theta}{max}, \qquad (8)$$

with normalization and non-negativity constraints,

$$\sum_w \varphi_{wt} = 1, \sum_w \theta_{td} = 1, \varphi_{wt} \geq 0, \theta_{td} \geq 0,$$
(9)

where n_{dw} is the number of occurrences of the word w in the document d. One of the most popular topic models is Latent Dirichlet Allocation (LDA). In [23], LDA method to model the clinical pathways of patients was first described. The authors suggested that LDA would allow to present hidden patterns of patient treatment as probabilistic combinations of initiating events. The choice of a probabilistic approach to modeling was due to the complexity of medical processes and the high variability of patient behavior during treatment.

The Latent Dirichlet Allocation is the generative hierarchical probabilistic model described in 2003 in a study [24] and originally developed to characterize text documents. The parameters of this model are generated from the Dirichlet a priori distribution, and methods of the Bayesian approach are used for training the model [24]. The document in LDA model is represented by distribution by hidden (latent) topics, each of which is characterized by word distribution.

In probabilistic generative models, in particular, LDA, the available data are considered as the result of a generating process involving hidden variables [24] . The authors also note that the generating process determined the joint probability distribution over the observed and hidden random variables. As a result, this joint distribution is used to calculate the conditional probability of hidden variables with observables, or a posteriori probability. In the Bayesian approach, one of the options for estimating the parameters of the LDA model is to maximize the a posteriori probability.

The method is characterized by a priori Dirichlet distribution on the parameters Φ and Θ. The parameter Φ contains discrete distributions on the set of words *w* for each topic *t*: $\varphi_{wt} = p\ (w|t)$. The parameter Θ contains probability distributions on a variety of topics for each document θ_{td} =

$p(t|d)$. For the patient's clinical path model based on the LDA method, it is necessary to consider the Dirichlet distribution, or a continuous multidimensional distribution on the simplex:

$$Dir(\theta|\alpha) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}, \qquad (10)$$

where $\alpha_k > 0, \forall k = 1, \dots, K$ – parameters. The application of LDA method allows for each patient to choose a set of clinical pathway patterns with different emphasis on the significance of these patterns. However, LDA chooses one of the possible solutions without providing an opportunity for the researcher to compare and to choose the best solution for a particular task.

Therefore, the alternative approach was developed [27], called Additive Regularization of Topic Models (ARTM). The model is based on the idea of multi-criteria *regularization,* which serves to set the desired properties of the topic model in the form of optimization criteria. Regularization is used to solve the problem of non-uniqueness and instability. Additional restrictions are imposed on the desired solution. For instance, there are regularizes that increase the difference in topics and provide the maximum possible sparseness of the solution. The ARTM model allows to develop models that satisfy many constraints at the same time. Each constraint is formalized as a regularizer, the optimization criterion $R_i(\Phi, \Theta) \to \max$, depending on the parameters of the model. Weighted total of all such criteria:

$$R_i(\Phi, \Theta) = \sum_{i=1}^{k} \tau_i R_i(\Phi, \Theta) \qquad (11)$$

To solve the regularized likelihood problem, an EM algorithm with modified M-step formulas is used:

$$\phi_{wt} \propto \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}}\right)_+,$$
$$\theta_{td} \propto \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}}\right)_+. \qquad (12)$$

In this study the following regularizes are applied:

- Decorrelation of topics in Φ, in order to increase the difference of topics (patterns).
- Smooth/Sparse Θ, to smooth or to sparse subsets of topics in Θ matrix.
- Smooth/Sparse Φ, to smooth or to sparse subsets of topics in Φ matrix.

To evaluate the quality of modelling the following metrics are used:

- Perplexity- one of the main scores, measuring the convergence of the model. This score has been used to define the optimal number of topics [24]. In context of this study, the perplexity score determines the number of underlying patient's patterns in data [23]:

$$Perplexity = exp\left[-\frac{\sum_{c\in L} \log p(c\sigma|L)}{\sum_{c\in L}|\sigma|}\right] \qquad (13)$$

where $|\sigma|$ is the number of patient's activities in σ, L is the event log of the medical facility.

- Sparsity Θ- a metric to smooth or to sparse subsets of topics in Θ matrix.
- Sparsity Φ- a metric to smooth or to sparse subsets of topics in Φ matrix.

IV. EXPERIMENTS AND RESULTS

For the experiment of the proposed approach the real event journal of the hospital, hosted by the University of Eindhoven University of Technology in 2016, was considered. This event log contains 846 patients, 16 activities and 15214 events recorded in the ERP (Enterprise Resource Planning) system of the medical institution. The presented routes describe the paths of patients with a diagnosis of sepsis, a severe inflammatory disease. The choice of the source is due to the fact that the databases contain complete and open information necessary for research tasks in the field of data analysis for healthcare.

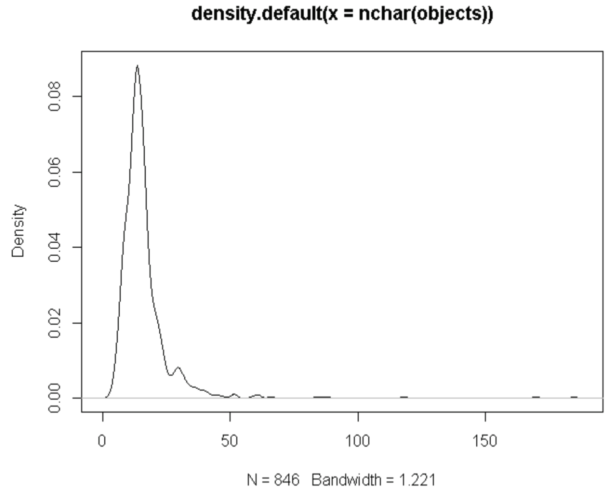


Figure 1. Event log path length distribution

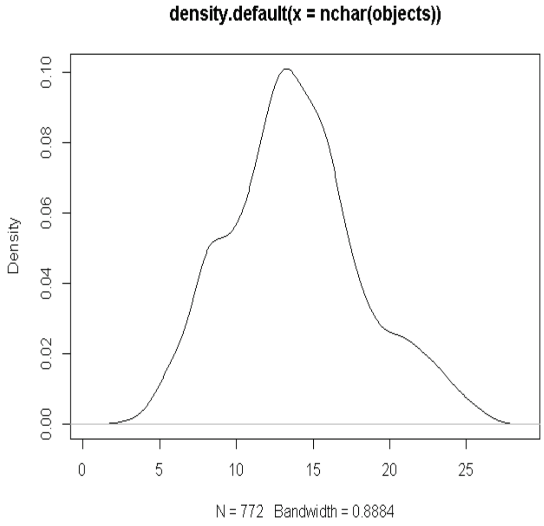


Figure 2. Event log pathways length distribution after constrain

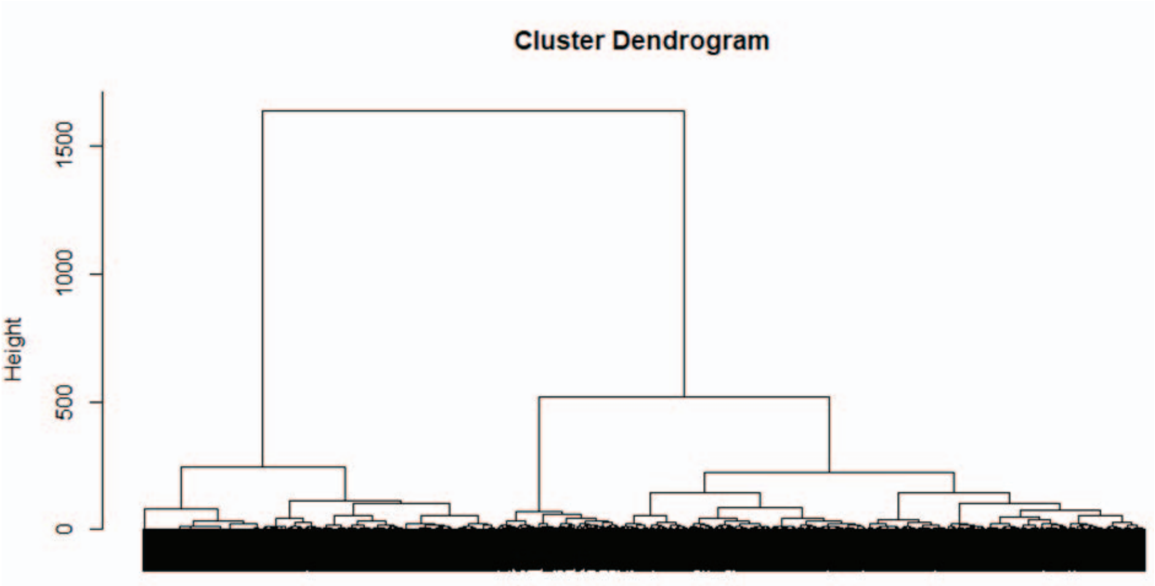


Figure 3. Clinical pathways cluster deprogram

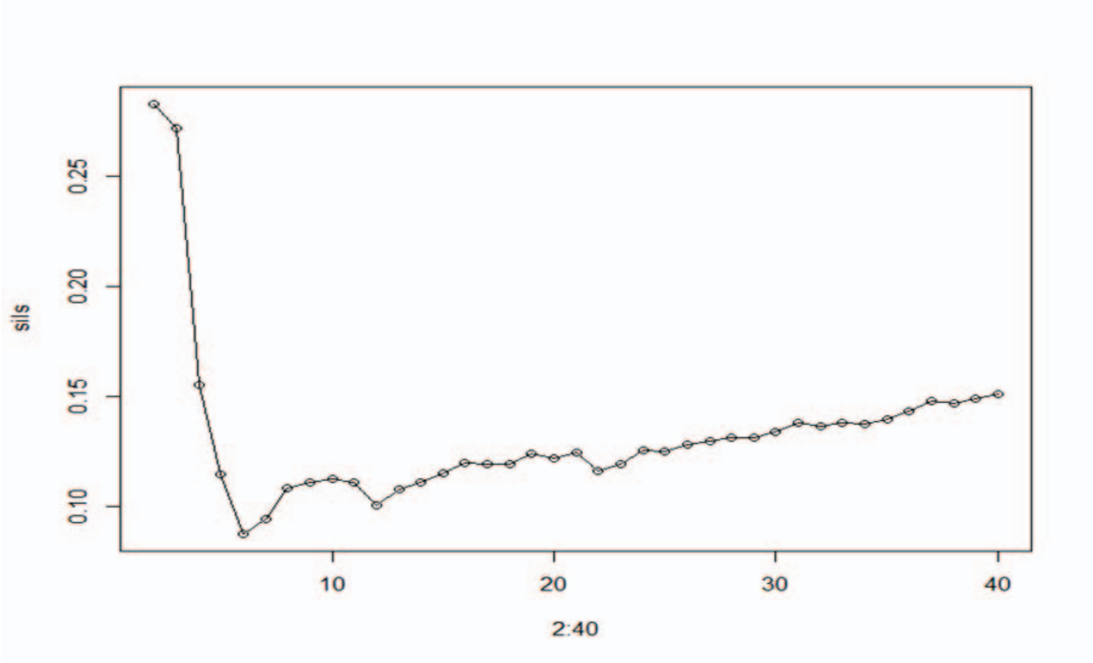


Figure 4. Silhouette coefficient depending on the number of clusters

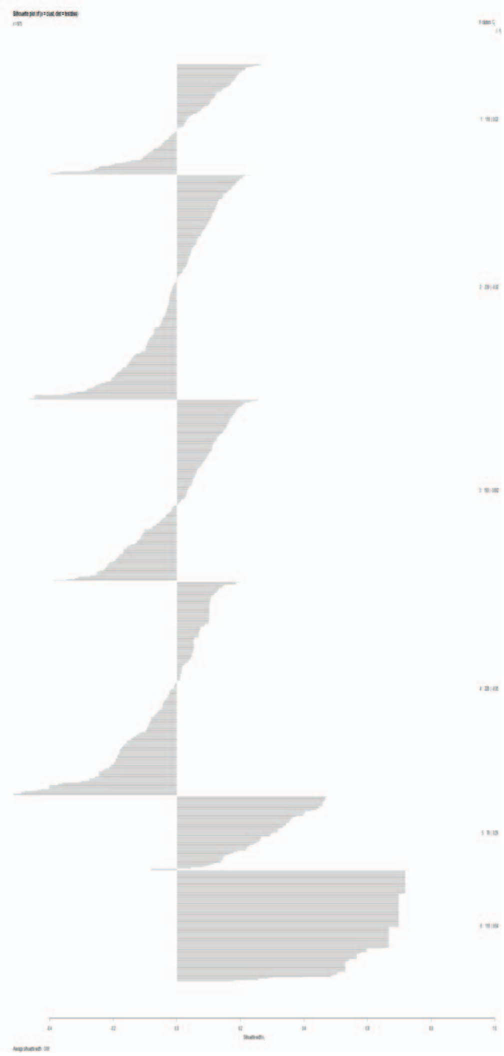


Figure 5. Silhouette coefficient for 6 clusters

As mentioned above, for hard clustering approach all patient traces were encoded by replacing activities with letters of the English alphabet in order. Then, the distribution of pathways' length was analyzed. As it can be seen from the line graph (Fig.1), there are some abnormally long pathways, which were rejected to improve the quality of clustering. The upper limit of the length of the path under consideration was set to 42 events, or Q50 + 3 * (Q75-Q50), where Q50 is the median, Q75 corresponds to 75% of the quantile (Fig.2).

In R programming language (package 'stringdist'), the distance matrix was constructed by method *osa,* (Optimal string alignment,), or restricted Damerau-Levenshtein distance [21]. The method is similar to the Levenshtein distance, however, it allows transposition of neighboring characters [21].

The hierarchical cluster analysis was based on the Ward method. Ward's method is aimed at discovering compact clusters. Next, the silhouette coefficient from 2 to 40 clusters was evaluated to find the optimal number of clusters [19] after the previous hierarchical clustering of the clinical pathways.

In accordance with the results of the analysis of the silhouette coefficient, 6 clusters were selected (Fig.5).

As a result (Fig. 5), it can be seen that the resulting clusters 5 and 6 have a large silhouette coefficient, most of the objects are positive, but the previous ones have the opposite tendency. In this regard, the clusters from 1 to 4 were excluded from the analysis (Fig. 6).
The patterns of clinical pathways for cluster 5 and cluster 6 were visualized in R programming language (Fig.7, Fig.8). The visual representation of obtained clusters is convenient for the further interpretation by medical specialists.

After the first clustering approach, the Additive Regularization of Topic Models (ARTM) was applied to the same dataset. First, the event data was converted to Vowpal Wabbit format, which accepts input data in a specific format: label |A feature1:value1 |B feature2:value2. This format is adapted to divide the features into categories, or modalities, which may be taken into account for model training.

BigARTM open source library for topic modelling was used to proceed this dataset in Python. This library is valuable for working with large collections of text documents and transitional data in healthcare facilities. The converted dataset in Vowal Wabbit format served as input data for the model. This dataset then is converted BigARTM internal format (called *batches*). After creating a class BatchVectorizer the next step is to create and to initialize a *Dictionary*, which is a useful data structure to represent activities of the event log with their frequencies. Finally, the model was created and trained by initial number of topics T= 300.
According to perplexity score (Fig.), sparsity Θ and sparsity Φ, the number of topics was selected. The number of 9 clusters is characterized by perplexity = 63,96, sparsity Θ = 0,44 and sparsity Φ= 0,42.

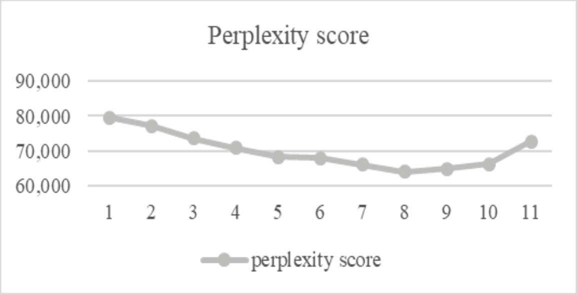
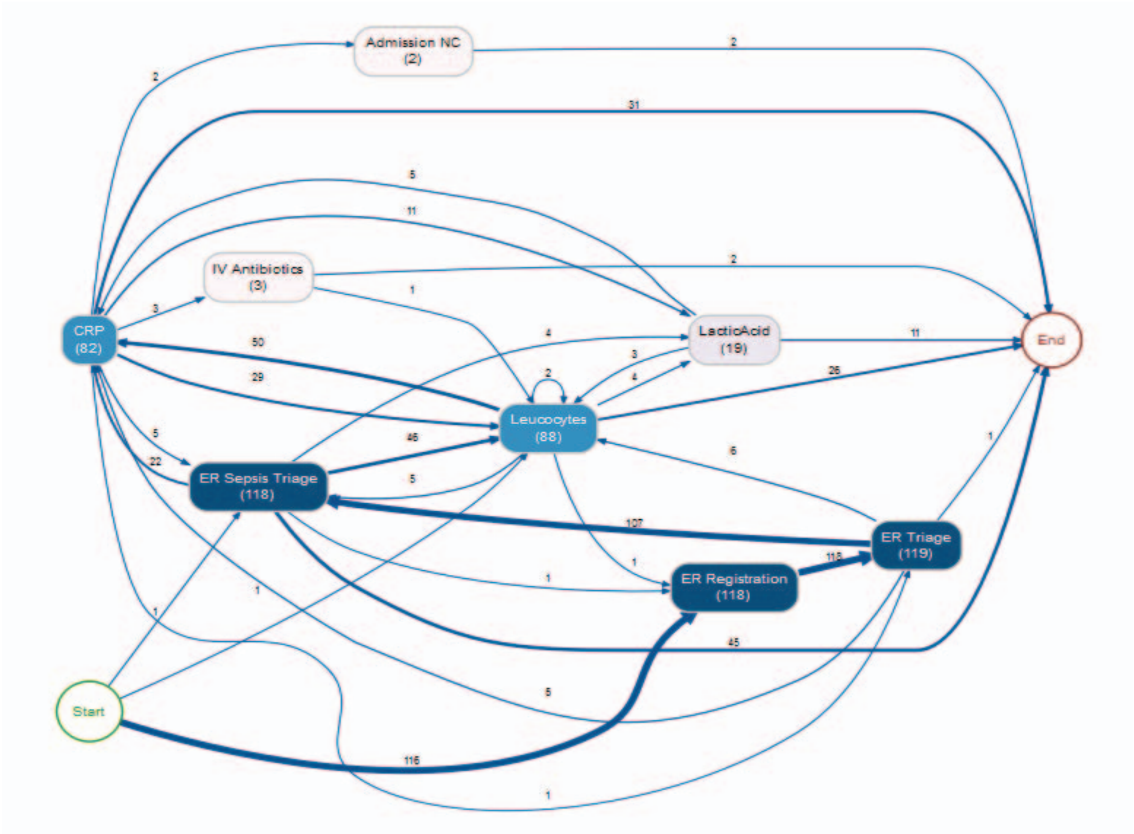


Figure 6. Perplexity score depending on number of clusters

The soft clustering approach defined 9 relevant clusters of patterns of patients' behavior observed in the event log. Each unique patient was assigned a probabilistic assessment of belonging to a particular cluster. This method allows to add a hierarchical representation of routes by introducing a modality. In study context, the resource feature will be used as a modality for the further research. Dedicated clusters will serve as a reference point for improved prediction and potential recommendation service for patients with a certain diagnosis in the medical facility.


Admission NC
(2)
2
2
31
5
11
IV Antibiotics
(3)
2
1
3
CRP
(82)
50
29
4
LacticAcid
(19)
11
End
3
2
4
26
Leucocytes
(88)
5
22
46
6
ER Sepsis Triage
(118)
5
1
107
ER Triage
(119)
1
1
118
ER Registration
(118)
1
45
Start
5
116
1

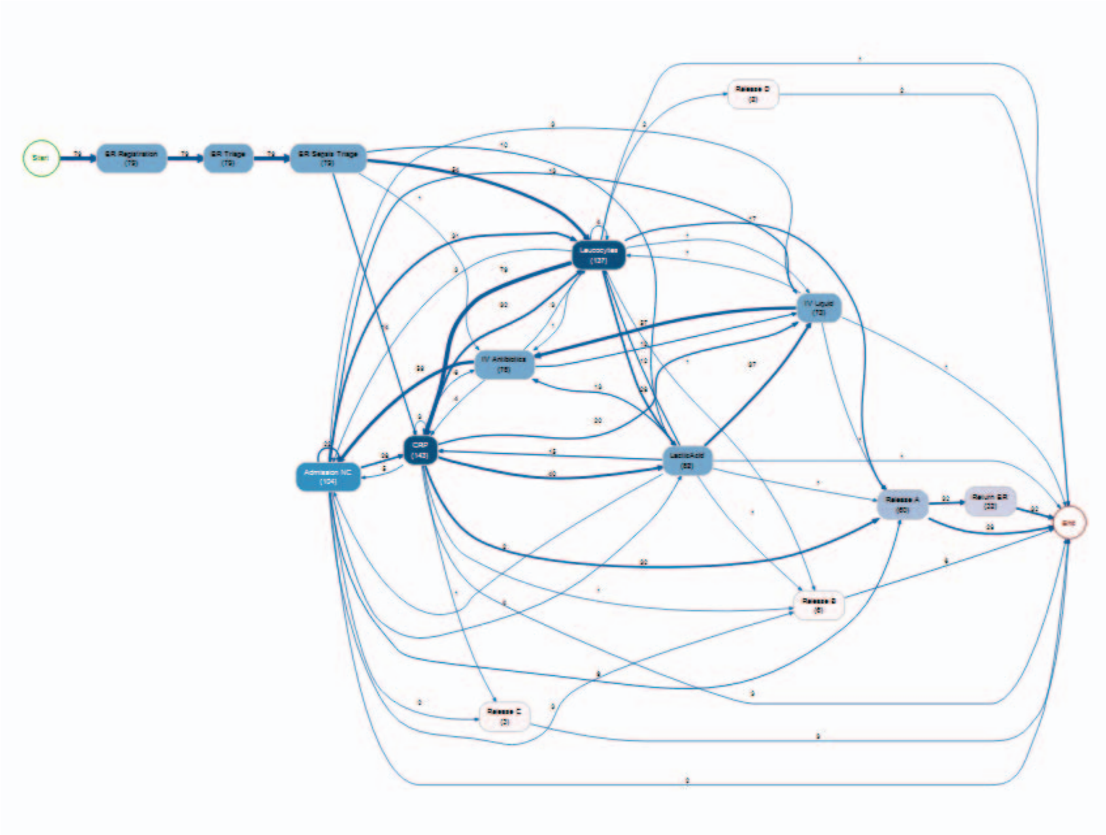

Start
ER Registration
(79)
ER Triage
(79)
ER Sepsis Triage
(79)
Release D
(3)
Leucocytes
(137)
IV Liquid
(72)
IV Antibiotics
(75)
CRP
(143)
Admission NC
(104)
LacticAcid
(82)
Release A
(60)
Return ER
(22)
End
Release B
(6)
Release C
(3)

TABLE I. Part of the resulting ARTM clustering of clinical pathways

Patient	Probability of clinical pathways clusters
A	0^0.20472845 2^0.25094473 3^0.056172576 4^0.34621504 5^0.13793103
B	2^0.31591785 3^0.0804124 4^0.59794307
C	2^0.43433434 3^0.4081694 6^0.1539437
D	0^0.269751 2^0.62204623 3^0.10738771
E	0^0.014719851 4^0.9807415
F	2^0.048205554 3^0.7285043 4^0.2220294
G	0^0.017998157 2^0.059349068 4^0.5676379 6^0.35303143
H	0^0.16274616 1^0.109899595 2^0.08268521 3^0.016755203 6^0.6182206
I	^0.7271091 2^0.040800642 4^0.16320838 6^0.06646766
J	0^0.26885462 2^0.4888821 3^0.16578344 4^0.07595149
K	0^0.025530832 2^0.23489715 3^0.16165423 4^0.23987517 6^0.33226854
L	2^0.5500032 3^0.41118115 6^0.03374122
M	0^0.14285715 1^0.14285715 2^0.14285715 3^0.14285715 4^0.14285715 5^0.14285715 6^0.14285715

V. Conclusion

In this study, the hard and soft clustering techniques of clinical pathways were presented, based on available large data volumes in medical information systems, to improve the management of heterogeneous patient flows in urban medical institutions. For the experiment of the proposed approach the real event journal of the hospital, hosted by the University of Eindhoven University of Technology in 2016, was considered.

For the subsequent hierarchical cluster analysis, the Ward method was chosen, for which initially only one object is included in each cluster. Then, the silhouette coefficient from 2 to 40 clusters was evaluated to find the optimal number of clusters In accordance with the results of the analysis of the silhouette coefficient, 6 clusters were selected and then analyzed for further visualization of trajectories.

After the first clustering approach, the Additive Regularization of Topic Models (ARTM) was applied to the same dataset. The soft clustering approach defined 9 relevant clusters of patterns of patients' behavior observed in the event log by perplexity score analysis. All in all, soft clustering method is more flexible for the clinical pathways segmentation, and for the further directions of research the modalities will be added to dataset.

The presented approaches serve as the basement for further development of simulation model of urban medical facility and recommendation service for patients. Besides, these groups of trajectories will be useful for medical experts to analyze the behavior and to compare it to the standard of the certain disease treatment processes.

Modern data analysis methods have enormous potential, which makes it possible to significantly improve healthcare services. Medical institutions that are the first to introduce these technologies will certainly have a competitive advantage, enabling to gain access to complete information to make more informed decisions.

References

[1] Martin Prodel. Process discovery, analysis and simulation of clinical pathways using health-care data. Other. Université de Lyon, 2017. English.

[2] Kinsman L, Rotter T, James E, Snow P, Willis J. What is a clinical pathway? Development of a definition to inform the debate. BMC Med. 2010;8:31. Published 2010 May 27. doi:10.1186/1741-7015-8-31.

[3] Panella, M., S. Marchisio, and F. Stanislao. 2003. "Reducing Clinical Variations with Clinical Pathways: Do Pathways Work?" International Journal for Quality in Health Care 15:509-521.

[4] Huang, Wei Dong, Lei Ji, Chenxi Gan, Xudong Lu, Huilong Duan, Discovery of clinical pathway patterns from event logs using probabilistic topic models, Journal of Biomedical Informatics, Volume 47, Pages 39-57, ISSN 1532-0464.

[5] F. Mannhardt, M. de Leoni, H.A. Reijers, W.M.P. van der Aalst, Data-Driven Process Discovery - Revealing Conditional Infrequent Behavior from Event Logs, in: 29th Int. Conf. CAiSE 2017, 2017: pp. 545–560. doi:10.1007/978-3-319-59536-8_34.

[6] Wakamiya S, Yamauchi K. What are the standard functions of electronic clinical pathways? Int J Med Inform. 2009 Aug;78(8):543–550. doi: 10.1016/j.ijmedinf.2009.03.003.

[7] Veselý Arnost, Zvárová Jana, Peleska J, Buchtela David, Anger Zdenek. Medical guidelines presentation and comparing with electronic health record. Int J Med Inform. 2006;75(3-4):240–245. doi: 10.1016/j.ijmedinf.2005.07.016.

[8] Kovalchuk, Sergey V. et al. "Simulation of Patients Flow in Multiple Healthcare Units using Process and Data Mining Techniques for Model Identification." Journal of biomedical informatics 82 (2018): 128-142.

[9] E. Rojas, J. Munoz-Gama, M. Sepúlveda, D. Capurro, Process mining in healthcare: A literature review, J. Biomed. Inform. 61 (2016) 224–236. doi:10.1016/j.jbi.2016.04.007.

[10] Huang Z, Lu X, Duan H. On mining clinical pathway patterns from medical behaviors. Artif Intell Med. 2012 Sep;56(1):35–50. doi: 10.1016/j.artmed.2012.06.002.

[11] G. Rakocevic, T. Djukic, N. Filipovic, V. Milutinović, eds., Computational Medicine in Data Mining and Modeling, Springer New York, New York, NY, 2013. doi:10.1007/978-1-4614-8785-2

[12] M. A. Ahmad, A. Teredesai and C. Eckert, "Interpretable Machine Learning in Healthcare," 2018 IEEE International Conference on Healthcare Informatics (ICHI), New York, NY, 2018, pp. 447-447. doi: 10.1109/ICHI.2018.00095

[13] Thomas Rotter, Leigh Kinsman, Erica L. James, Andreas Machotta, Holger Gothe, Jon Willis, Pamela Snow, and Joachim Kugler. Clinical pathways: effects on professional practice, patient outcomes, length of stay and hospital costs. Cochrane Database of Systematic Reviews, 2010. ISSN 1465-1858. doi: 10.1002/14651858.CD006632.pub2. EPOC.

[14] Huang, Wei Dong, Lei Ji, Chenxi Gan, Xudong Lu, Huilong Duan, Discovery of clinical pathway patterns from event logs using probabilistic topic models, Journal of Biomedical Informatics, Volume 47, Pages 39-57, ISSN 1532-0464.

[15] A. I. Orlov, “Grafy pri modelirovanii processov upravleniya promyshlennymi predpriyatiyami”, UBS, 30.1 (2010), 62–75

[16] Goto, Yoshikazu & Maeda, Tetsuo & Goto, Yumiko. (2013). Decision-tree model for predicting outcomes after out-of-hospital cardiac arrest in the emergency department. Critical care (London, England). 17. R133. 10.1186/cc12812.

[17] Wil M.P. van der Aalst. Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer Publishing Company, Incorporated, 1st edition, 2011. ISBN 3642193447, 9783642193446.

[18] Rebuge, A., and Ferreira, D.R., Business process analysis in healthcare environments: A methodology based on process mining. Inf. Syst. 37(2):99–116, 2012.

[19] Murtagh, Fionn & Legendre, Pierre. (2014). Ward’s Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward’s Criterion?. Journal of Classification. 31. 274-295. 10.1007/s00357-014-9161-z.

[20] Rousseeuw P. J. Silhouettes (1987) A graphical aid to the interpretation and validation of cluster analysis //Journal of computational and applied mathematics, vol. 20, pp. 53-65.

[21] Van der Loo M (2014). “The stringdist package for approximate string matching.” The R Journal, 6, pp. 111-122. https://CRAN.R-project.org/package=stringdist.

[22] Ward, J. H. (1963). Hierarchical grouping to optimize an objective function, Journal of Amer. Statist. Assoc. 58: 236-244.

[23] Huang Z, Lu X, Duan H. Latent treatment topic discovery for clinical pathways. J Med Syst 2013;37(2):1–10.

[24] D.M. Blei, A.Y. Ng, M.I. Jordan. Latent Dirichlet allocation. J Mach Learn Res, 3 (2003).

[25] Blei, D.M. (2010). Introduction to Probabilistic Topic Models.

[26] Voroncov K. V. Obzor veroyatnostnyh tematicheskih modelej. 2019.

[27] Vorontsov K. V., Potapenko A. A. Additive regularization of topi models // Mahine Learning, Speial Issue on Data Analysis and Intel ligent Optimization, 2014.