DOI: 10.17323/2587-814X.2020.1.19.31

Clinical pathways analysis of patients in medical institutions based on hard and fuzzy clustering methods

Elizaveta S. Prokofyeva[a]
E-mail: prokofyeva.liza@gmail.com

Roman D. Zaytsev[b]
E-mail: Roman.Zaitsev@fors.ru

[a] National Research University Higher School of Economics
Address: 20, Myasnitskaya Street, Moscow 101000, Russia

[b] FORS Group
Address: 3, Trifonovskiy Tupik Street, Moscow 129272, Russia

Abstract

Modeling the processes in a healthcare system plays a large role in understanding its activities and serves as the basis for increasing the efficiency of medical institutions. The tasks of analyzing and modeling large amounts of urban healthcare data using machine learning methods are of particular importance and relevance for the development of industry solutions in the framework of digitalization of the economy, where data is the key factor in production. The problem of automatic analysis and determination of clinical pathways groups of patients based on clustering methods is considered in this research. Existing projects in this area reflect a great interest on the part of the scientific community in such studies; however, there is a need to develop a number of methodological approaches for their further practical application in urban outpatient institutions, taking into account the specifics of the organization being analyzed. The aim of the study is to improve the quality of management and segmentation of patient input flow in urban medical institutions based on cluster analysis methods for the further development of recommendation services. One approach to achieving this goal is the development and implementation of clinical pathways, or patient trajectories. In general, the clinical pathway of a patient might be interpreted as the trajectory when receiving medical services in respective institutions. The approach of developing groups of patient routes by the hierarchical agglomerative algorithm with the Ward method and Additive Regularization of Topic Models (ARTM) is presented in this article. A computational experiment based on public data on the routes of patients with a

diagnosis of sepsis is described. One feature of the proposed approach is not just the automation of the determination of similar groups of patient trajectories, but also the consideration of clinical pathways patterns to form recommendations for organizing the resource allocation of a medical institution. The proposed approach to segmenting the input heterogeneous flow of patients in urban medical institutions on the basis of clustering consists of the following steps: 1) preparing the data of the medical institution in the format of an event log; 2) encoding patient routes; 3) determination of the upper limit of the clinical pathway length; 4) hierarchical agglomerative clustering; 5) additive regularization of topic models (ARTM); 6) identifying popular patient route patterns. The resulting clusters of routes serve as the foundation for the further development of a simulation model of a medical institution and provide recommendations to patients. In addition, these groups may underlie the development of the robotic process automation system (RPA), which simulates human actions and allows you to automate the interpretation of data to manage the resources of the institution.

Key words: cluster analysis; data; hierarchical clustering; topic modeling; silhouette coefficient; healthcare; clinical pathways; process mining.

Citation: Prokofyeva E.S., Zaytsev R.D. (2020) Clinical pathways analysis of patients in medical institutions based on hard and fuzzy clustering methods. *Business Informatics*, vol. 14, no 1, pp. 19–31.
DOI: 10.17323/2587-814X.2020.1.19.31

Introduction

The rapidly evolving data analysis technologies play a huge role in healthcare. The current level of automation of medical care allows you to process large amounts of information and use the accumulated data to solve optimization problems. Medical institutions have data on receptions, but the traditional approach to documenting visits does not allow a complete representation of the main trajectories of patients and its automatic analysis.

An important area of data processing technologies in healthcare is work optimization of medical institutions: an effective schedule for medical personnel, forecasting the patient flow, planning and distribution of resources, reducing queues and other tasks. Modern analytical technologies facilitate the development of decision-making tools based on empirical data. For example, aggregated data on the actual movements of patients between medical institutions and specialists within these institutions allow one to plan the load of resources, ensure a high level of service availability and optimize the organization's work based on the real demand for these services. Based on such data, patient flow management information systems are actively developing [1, 2].

The development and implementation of clinical pathways, or patient paths, is an important tool in healthcare management. In general, the clinical pathway of a patient is a trajectory when receiving services in respective institutions. According to the source [3], clinical pathways have been introduced internationally since the 1980s. This methodology was presented in medical institutions in Sweden in the mid-1990s; and in the United States, according to source [3], approximately 80% of hospitals used clinical pathways to improve the quality of care.

In the study [4], the authors described the clinical pathway as a plan which reflects the goals for patients and determines the sequence and time of actions necessary to achieve these goals with optimal efficiency.

According to sources [5,6], the developed clinical pathways were integrated into the electronic document management of medical institutions. However, the rapid growth

of available data and digital images revealed the need for automatic determination of the patient's clinical pathway based on these data. The solution for the automatic identification of a personal clinical pathway was the technology of process mining [7, 8], data mining [9], machine learning algorithms [10, 11] and others.

In research papers, there are various definitions of a clinical pathway. For example, according to [12], the clinical pathway can be defined as a structured care plan with the indicated main stages and terms of treatment of patients. It is important to note that each trajectory of the patient is unique and corresponds to his or her medical history [12]. The clinical pathways may include a chain of events corresponding to the profile of the medical institution: initial appointment with the therapist, laboratory tests, obtaining advice from a specific specialist, and others. In [13], the authors pointed out that the identification of clinical pathway patterns can potentially complement the information about the intentions and behavior of the patient and can serve as a basis for further analysis of patient movements.

In this article, hard and fuzzy clustering methods are considered to solve the problem of automatic analysis and segmentation of clinical pathways in order to improve standardization of management and further solution of optimization problems as well as development of relevant services for patients.

The article has the following structure. Section 1 presents the existing approaches to modeling the patient's clinical pathways. Section 2 contains formal definitions generally accepted for modeling clinical pathways, and a methodology for their cluster analysis. A computational experiment using the example of open hospital data and its results are described in Section 3. The Conclusion lists the main trends and directions for further work.

1. Existing approaches to modeling clinical pathways of patients

There are many papers [14–18] devoted to the research and analysis of clinical pathways based on initial data of a medical institution. Clinical pathways modeling might be developed on the methodological base of probability theory, mathematical statistics, data mining, graph theory, semantic technologies, process mining, etc.

In a study [14], the authors noted the importance of developing an adaptive approach to modeling clinical pathways due to the high variability of patient trajectories and their individual characteristics. Based on proposed graphs of sequences and data mining methods, patterns or patterns of the clinical pathways of stroke patients are distinguished to predict the trajectories of new patients.

The semi-Markov model of individual patient experience in a family practice clinic is presented in [15]. The scheme of the general patient flow in this clinic is represented by an oriented graph, the vertices of which are the rooms and departments of the clinic, and edges correspond to the direction of movement of the analyzed flow. This model allows one to predict the duration of patient care, but such parameters as the waiting time in the queue and the length of the queue are not available for analysis.

Modeling patient trajectories by Markov chains in [16] made it possible to identify typical clinical pathways during disease progression and visualize it. Due to the ability of Markov chains to take into account nested models, in [16] the clinical pathway is presented in the form of four levels of aggregation. According to the authors of [12], the nested design allows one to simplify the clinical pathway model and highlight the most important regularities of the process. However, as the main drawback of the Markov chain for modeling the clinical pathway, the authors of the study highlight a limited number of states to handle.

The probabilistic topic modeling method, in particular, Latent Dirichlet Allocation (LDA), was adapted to model the clinical pathways of patients in study [13], where the authors suggested that LDA would allow the hidden treatment patterns of patients to be presented as probabilistic combinations of initial events from the event log (*Figure 1*). The latent Dirichlet distribution is a generative hierarchical probabilistic model described in 2003 in a study [17] and originally developed to characterize text documents. The parameters of this model are generated from the a priori Dirichlet distribution, and the Bayesian approach methods are used to train the model [17]. The document in the LDA model is represented by a distribution of hidden (latent) topics, each of which is characterized by a distribution of words. In the framework of this study, topics of documents correspond to the patterns of clinical pathways obtained.

A special place among the methods for modeling the clinical pathways of patients is taken by the application of process mining. The development of a process model is based on initial data on the real behavior of patients, their routes and the main characteristics that affect the choice of a particular trajectory. The purpose of process mining is to extract new information about processes from event logs. Thus, process mining as a discipline combines machine learning, data mining, and process modeling techniques. The main ideas of this discipline are described in [19–22].

A number of studies are devoted to the development of process analysis algorithms: for example, the eMotivia algorithm for analyzing the movements of nine patients over 25 weeks [23]. A detailed list of algorithms for process mining in healthcare is given in [7], based on a review of 74 studies in this field. It is important to note that the results of the application of process mining allows one to objectively evaluate the past and current movement of patient flow, however, for a detailed study of the system's behavior and the experimental part for its improvement, it is necessary to develop a simulation model [12, 20, 24].

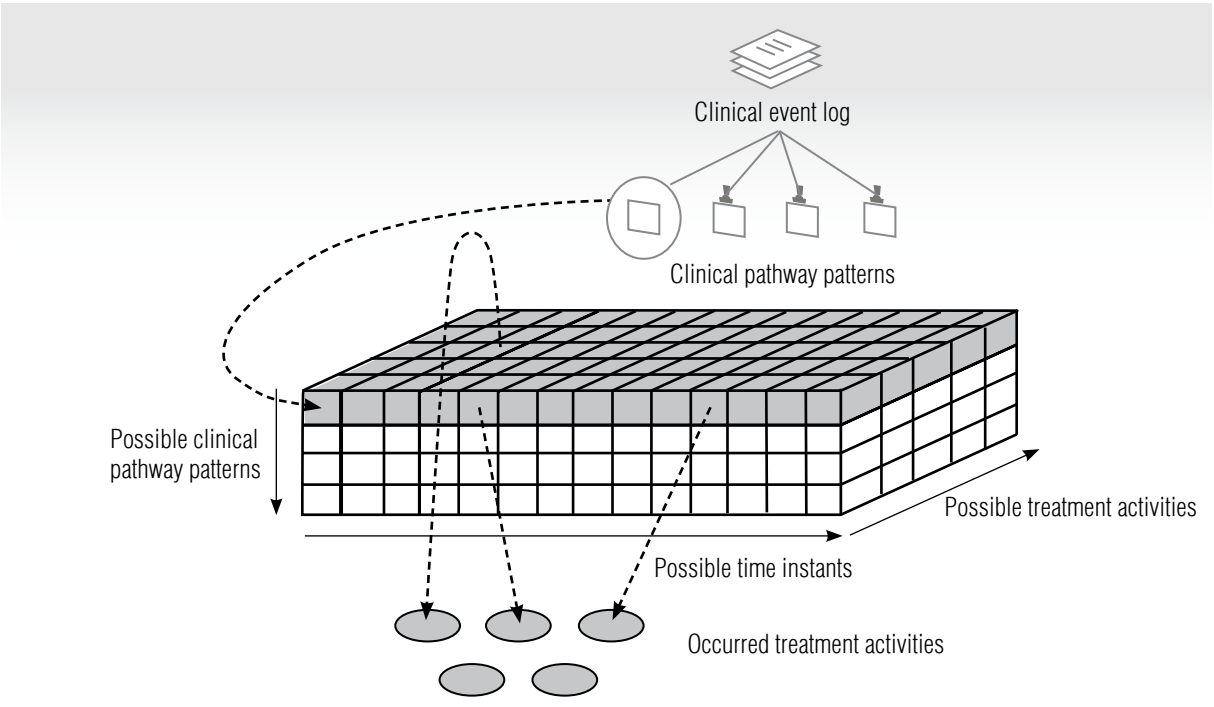

Fig. 1. Study of clinical pathway patterns based on probabilistic topic modeling [13]

2. Methods of cluster analysis of clinical pathways of patients

2.1. Terms and definitions

The specific structure of clinical pathways allows us to use cluster analysis of time series. This is due to the fact that the series by which clinical paths are expressed can have both categorical and numerical nature. One of the ways to significantly increase the accuracy of the model so developed is to segment the initial sample into subgroups of objects similar to each other and to construct separate "personalized" models for each of the selected groups in the future. This study aims to identify such groups, or clusters, of clinical pathways based on event logs of medical institutions.

Some formal definitions generally accepted for modeling clinical pathways [12–14]. Let E be the set of all real events in the study area that occurred during the medical care process: $E \subseteq A \times T$, where A is the finite set of event identifiers, T is the set of time attributes. Then the event is a pair $e = (a, t)$, where $a \in A$ and $t \in T$. The type of activity and time label of the clinical event is denoted as $e \cdot a$ and $e \cdot t$. It is important to note that when modeling clinical pathways, each event is uniquely determined by a combination of its attributes. Trace σ is the event chain of the patient, a non-empty sequence of events of the clinical path: $\sigma = \langle e_1, e_2, ..., e_n \rangle$, where $ei \in E$ $(1 \le i \le n)$, $n \in N$ is the length of the patient's route. The set of all routes over E is denoted by E *. The event log L is a non-empty set of patient routes over E*: $L = \{\sigma_1, ..., \sigma_m\}$, where $\sigma_i \in E^*$ $(1 \le i \le m)$, $m \in N$

In the framework of process mining terminology adapted for modeling the clinical pathways of patients, the following examples of definitions' correspondence can be given:

✦ trace σ_i – patient attached to the analyzed clinic;

✦ event e_i – visit to the therapist for an initial consultation;

✦ attribute a_i – patient characteristic (gender, age, diagnosis, etc.);

✦ event log L – the source database of the medical institution.

The structure of the event log L assumes the presence of the following attributes [22]:

✦ identifier (patient_id): stores objects for which sequences of events are built;

✦ activity (activity_name): stores actions performed as part of journal events;

✦ timestamp: stores the date and time of recording events of the journal, for example, the time of the visit to the therapist;

✦ resource: stores the main actors of the events of the journal (those who perform actions within the framework of the events of the journal). In the context of a study, a resource may be provided by a medical professional or research equipment;

✦ other data: other data that is potentially useful for modeling the processes of a medical institution.

2.2. Hard clustering

There are different patient route encoding systems. For example, in a study [25], the authors use alphanumeric characters to indicate clinical pathway activities (which may include diagnoses, procedures, analyzes, and treatment regimens) according to the Unicode standard. Thus, the authors indicate the ability to encode 65,536 activities of the clinical pathway.

The choice of coding system depends on the maximum number of activities of the event log of a medical institution. In this work, at the initial stage, all patient routes are coded by replacing events with letters of the English alphabet in order, since the initial data set for the experimental part contains 16 types of activities: ER Registration, ER Triage, IV Liquid, IV Antibiotics, CRP, Admission IC, ER Sepsis Triage, Leucocytes, Lactic Acid, Admission NC,

Release A, Release B, Release C, Release D, Release E, Return ER.

In the context of this study, the upper limit of the length of the pathway was assumed to be 26 events: $Q50 + 3 \cdot Q(Q75 - Q50)$, where $Q50$ is the median and $Q75$ corresponds to 75% quantile. Therefore, pathways containing more than 26 events are considered abnormally long. After analyzing the distribution of the clinical pathways' lengths of the event log, abnormally long paths were excluded from the analysis to improve the quality of clustering.

Two clustering methods were compared: *k*-medoids [26] and the Ward hierarchical agglomerative algorithm [27]. Ward's method is based on analysis of variance methods to estimate distances between groups of objects and, along with the complete linkage method, leads to the formation of small compact clusters. The method is applicable for tasks of more fractional classification of objects with closely spaced clusters [29]. Each sample object in the Ward method is initially considered as a separate cluster [30]. At the next step of the iteration of the algorithm, the closest clusters are combined, the distance between them being measured by the following formula:

$$\Delta(\mathrm{A},\mathrm{B}) = \sum_{i\in \mathrm{A}\cup \mathrm{B}} \left\|\vec{x}_i - \vec{m}_{\mathrm{A}\cup \mathrm{B}}\right\|^2 - \sum_{i\in \mathrm{A}} \left\|\vec{x}_i - \vec{m}_{\mathrm{A}}\right\|^2 - \\ - \sum_{i\in \mathrm{B}} \left\|\vec{x}_i - \vec{m}_{\mathrm{B}}\right\|^2 = \frac{n_{\mathrm{A}} n_{\mathrm{B}}}{n_{\mathrm{A}} + n_{\mathrm{B}}} \left\|\vec{m}_{\mathrm{A}} - \vec{m}_{\mathrm{B}}\right\|^2, \quad (1)$$

where A, B – are joined clusters;

$\vec{x}_i$ – the cluster object;

$\vec{m}_{\mathrm{A}\cup \mathrm{B}}$ – the center of joined cluster AB;

$\vec{m}_{\mathrm{A}}$ – the center of cluster A;

$\vec{m}_{\mathrm{B}}$ – the center of cluster B;

n_{A} – number of objects in cluster A;

n_{B} – – number of objects in cluster B.

Based on the results of solving the problem of identifying the most specific clusters described later in this section, the Ward algorithm was chosen.

At the next step, a matrix of distances between clinical pathways was constructed on the basis of the limited Damerau–Levenshtein distance [31] – a measure of the difference of two lines of characters defined as the minimum number of insertion, deletion, replacement and permutation operations of neighboring characters needed to transfer one line to another.

To estimate the Damerau–Levenshtein distance between two lines *a* and *b,* the function is determined [31], see *formula 2* below, where $1_{(a_i \neq b_j)}$ – is the indicator function equal to 1 when $a_i \neq b_j$ and equal to 0 otherwise. Besides, each recursive call matches one of the cases covered by this distance:

$d_{a,b}(i-1, j)+1$ corresponds to a deletion (from *a* to *b*),

$d_{a,b}(i, j-1)+1$ corresponds to an insertion (from *a* to *b*),

$$d_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min\begin{cases} d_{a,b}(i-1,j)+1 \\ d_{a,b}(i,j-1)+1 \\ d_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \\ d_{a,b}(i-2,j-2)+1 \end{cases} & \text{if } i,j>1,\ a_i = b_{j-1} \text{ and } a_{i-1} = b_j, \\ \min\begin{cases} d_{a,b}(i-1,j)+1 \\ d_{a,b}(i,j-1)+1 \\ d_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & \text{otherwise}, \end{cases} \quad (2)$$

$d_{a,b}(i-1, j-1)+1_{(a_i \neq b_j)}$ corresponds to *a* match or mismatch,

$d_{a,b}(i-2, j-2)+1$ corresponds to *a* transposition between two successive symbols.

The optimal number of clusters was defined within a silhouette coefficient [32]. This coefficient is based on the idea of determining the proximity of each investigated object to its cluster. Suppose that the distance *d* on the set to be clustered is given, and using a certain method a clusterization model is obtained. Let for each object of the sample *i* belonging to the cluster C_i the quantity $a(i)$ be equal to the average distance from *i* to each of the objects *j* of the same cluster:

$$a(i)=\frac{1}{|c_i|}\sum_{i\neq j} d(i,j), j \in C_i. \quad (3)$$

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:

$$C'=\arg\min_{C_k \in C \setminus C_i}\left(\frac{1}{|C_k|}\sum_{j\in C_k} d(i,j)\right). \quad (4)$$

The average distance from point *i* to a neighboring cluster defined as $b(i)$:

$$b(i)=\frac{1}{|C'|}\sum_{j\in C'} d(i,j). \quad (5)$$

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

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i);b(i)\}}. \quad (6)$$

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

$$s(C_k)=\frac{1}{|C_k|}\sum_{j\in C_k} s(j). \quad (7)$$

and the overall silhouette of the clustering model

$$s(C)=\frac{1}{|C|}\sum_{C_k\in C} s(C_k). \quad (8)$$

The largest 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. Since the results of cluster analysis should be well interpreted, it is necessary to choose a model containing clusters with the greatest silhouette. Let $K_i=\{C_1^i, C_2^i, ...C_{N_i}^i\}$ be the clustering model dividing the sample into *Ni* clusters, and $C_{\max}^i$ the cluster with the largest silhouette among all clusters of the *i*-th clustering model:

$$C_{\max}^i=\operatorname{argmax}_{C_j^i\in K_i}\left(s\left(C_j^i\right)\right), j\in 1...N_i.. \quad (9)$$

Then there is a need to find model, which contains the cluster with the largest silhouette among all *K* models:

$$K_{opt}=\operatorname{argmax}_{K_i\in K}\left(C_{\max}^i\right). \quad (10)$$

2.3. Soft clustering

Probabilistic fuzzy, or overlapping, clustering of patient routes by groups of clinical patterns allows one to develop a more flexible approach when describing the total flow of patients, where each sample object belongs to a cluster with a certain weight or probability. The application of this approach is based on topic modeling, originally developed to determine the topics of a collection of text documents. In terms of topic modeling, the events of the patient during medical care are correlated with the words of the model. The patient's route is represented by a sequence of such events similar to a document with words. Thus, hidden topics discovered by the algorithm are interpreted as patterns of the patient's clinical pathways [13].

According to study [13], the application of the LDA method allows us to choose a set of clinical pathway patterns for each patient with different emphasis on the significance of these patterns. Thus, we model a mixture of route patterns as a polynomial probability distribution along a clinical pathway to pattern z. Similarly, the importance of each clinical action a with each template is modeled as a polynomial probability distribution $P(a|\sigma)$ according to the patient's activities. These two distributions allow us to calculate the probability of a separate clinical activity in a patient:

$$P(a|\sigma) = \sum_{z=1}^{K} P(a|z)\,P(z|\sigma). \qquad (11)$$

In probabilistic generating models (for example, LDA), the available data are considered as the result of the generating process, including hidden variables [33]. In this paper, it is also noted that the generating process determines 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 observed or posterior probabilities. The choice of a probabilistic approach to modeling is due to the complexity of medical processes and the high variability of patient behavior during treatment.

Application of the LDA method allows each patient to choose a set of clinical pathways patterns with a different emphasis on the significance of these patterns. However, LDA chooses one of the possible solutions, without giving the researcher the opportunity to compare and choose the best solution for a specific task. In connection with this limitation, an alternative approach of Additive Regularization of Topic Models (ARTM) was developed, leading to the modularity of topic modeling technology [34]. In this work, the BigARTM library was used to determine the groups of clinical pathways, based on additive regularization. In this study, the following regularities are applied:

✦ decorrelation of the distribution of terms in topics in order to increase the diversity of these topics;

✦ smoothing the distribution of topics in documents;

✦ smoothing the distribution of terms in topics;

✦ sparse out the distribution of terms in topics;

✦ sparse out the distribution of topics in documents.

To assess the quality of modeling and determine the optimal number of topics, perplexity is used – one of the metrics implemented in the BigARTM library. In the context of this study, perplexity determines the number of basic patient patterns in the log of a medical institution [13]:

$$P = \left[\exp -\frac{\sum_{\sigma\in L}\log P(e_\sigma|M)}{\sum_{\sigma\in L}|\sigma|}\right], \qquad (12)$$

where M – model;

e_σ – set of hidden events in the patient's trace σ.

3. Computational experiment

For the experiment of the proposed approach, the public event log of a Dutch hospital is considered. The event log contains 1,143 patient routes and 150,291 events. The choice of source is due to the fact that the databases contain complete and open information necessary for research tasks in the field of healthcare.

After removing anomalously long pathways, the upper limit of the pathways' length was assumed to be 26 events: $Q50 + 3 \cdot (Q75 - Q50)$, where $Q50$ is the median and $Q75$ corresponds to 75% of the quantile. The programming language R was chosen for data analysis. This language is well suited to research tasks, because it contains a rich library of packages for various scenarios [28]. The use of the R language also helps us to visualize data for understanding the general picture of the studied subject area

[28]. Using the Stringdist package, the distance matrix was constructed using the Osa method (a Damerau–Levenshtein distance measure).

The model with maximum silhouette coefficient in a cluster was discovered by the compared Ward method and *k*-medoids (*Figure 2*), where the clusters are located along the X axis, and the silhouette values along the Y axis.

In accordance with the analysis of the coefficient values, the Ward method was chosen and the trends of the groups obtained were identified (*Figure 2*). Clusters with a low value were excluded (Figure 3), thus, as a result of experiments, clusters 5 and 6 were selected with the highest silhouette coefficient (*Table 1*).

A free ProcessmapR[1] package is used to build a map of the processes of the original dataset. However, without preliminary separation of routes into clusters, it is difficult to interpret the resulting clinical pathways.

After determining the optimal number of clusters, process maps for the obtained groups were separately generated. For example, on the process map for cluster 5 (*Figure 4*), the nodes of the graph indicate the main stages of the clinical pathway for patients diagnosed with sepsis: start, registration in the appropriate unit, taking antibiotics, etc. The edges of the graph represent the transitions of patients at these stages of treatment; the numbers on the edges correspond to the number of people making this transition between nodes. More significant patient paths are marked by wider edges. Thus, the process map allows one to quickly assess the busiest routes of the medical institution. In addition, such maps can be interpreted by medical specialists in the context of comparing them with accepted medical standards to identify overloaded resource units and further reorganize the service process.

The next step was fuzzy, or "soft" clustering of the initial data by topic modeling methods in which the patient's path can refer to several patterns (topic clusters) with different probabilities. Latent Dirichlet allocation (LDA) [17] is used in the definition of clinical path clusters [13, 32] and is considered one of the standard

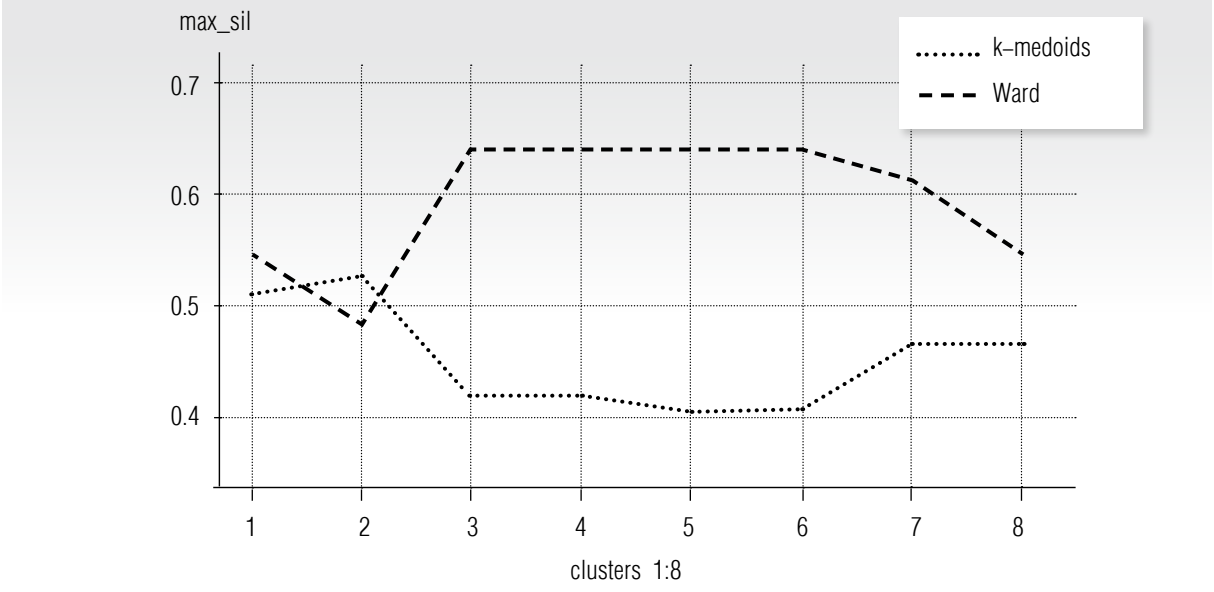

Fig. 2. Dependence of the maximum silhouette coefficient in the model on the number of clusters for k–medoids and Ward methods

[1] https://cran.r-project.org/web/packages/processmapR/processmapR.pdf

Table 1.

Cluster Silhouette Values

Cluster number	Number of objects	Silhouette coefficient value
1	118	0.02
2	239	–0.02
3	193	0.002
4	228	–0.05
5	79	0.29
6	118	0.64

methods of topic modeling. When constructing such a topic model, an infinite number of solutions arise, leading to instability and poor interpretability of topics [34]. In order to solve such problems with the choice of the best solution, additional regularizations or optimality criteria are specified [34]. Thus, it became necessary to develop a new multicriteria approach - additive regularization of topic models (ARTM), proposed in [34].

The application of this more flexible approach to clustering the clinical pathways of patients has not been previously considered in studies. This article used the BigARTM open source library in Python, which is based on additive regularization. The data was converted to Vowpal Wabbit format, which accepts input in a specific structure: label | A feature1: value1 | B feature2: value2. This format is adapted to be categorized or modal when training a model. The model was created and trained on the initial number of topics $T = 300$. Based on the calculated perplexity parameters of 63.97 and sparseness factors $\Theta = 0.44$ and $\Theta = 0.42$, the optimal number of clusters was chosen, equal to nine.

Each unique patient was assigned a probabilistic assessment of belonging to a particular cluster. For example, for one of the patients in the sample, the distribution according to the clinical path patterns is as follows: $P_1 = 0.017998157$, $P_2 = 0.059349068$, $P_4 = 0.5676379$, $P_6 = 0.35303143$. Accordingly, the next step of the patient with a probability of about 57% will correspond to the behavioral pattern 4 of the cluster.

The fuzzy clustering method allows one to add a hierarchical representation of patient routes, displaying the resources of medical institutions. Dedicated clusters will be the starting point for improving the forecast of the flow of the patients, as well as for forming recommendations on the resource equipment of hospitals in the development of services.

Conclusion

Currently, medical institutions have large amounts of data, however, the traditional approach to documenting processes does not allow a complete picture of all patient trajectories and their automatic analysis in real time, taking into account predicted flow estimates.

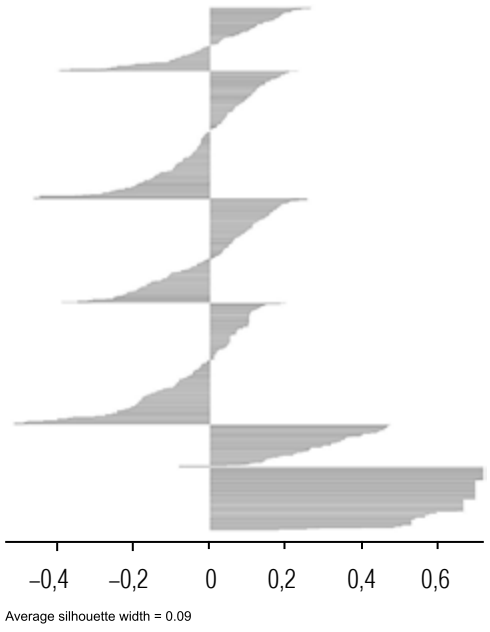

Fig. 3. Silhouette coefficients of the six resulting clusters

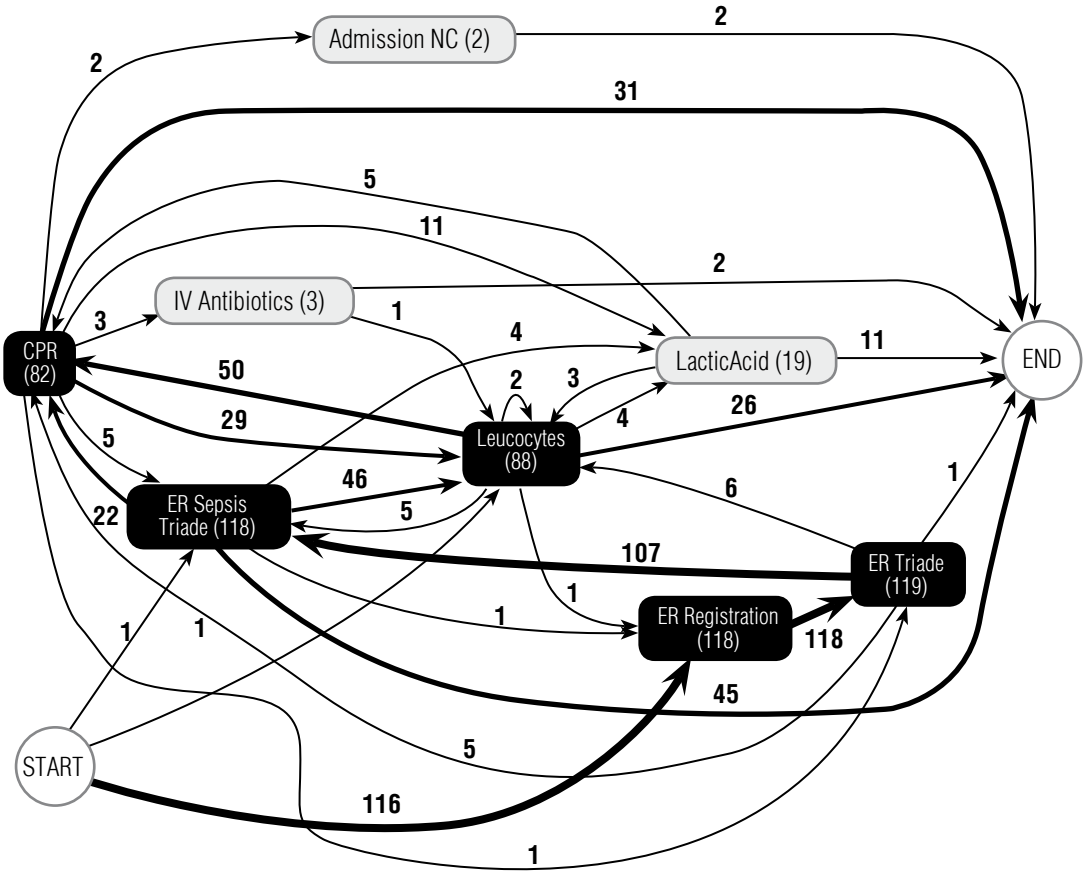

Fig. 4. Process map for cluster 5

In addition, the diverse nature of diseases is reflected in the high variability of routes.

Based on the results of analyzing a number of studies, the main methods for modeling clinical pathways were considered and the limitations of their application were identified. A methodology for the formation of patient route groups by the hierarchical agglomerative algorithm with the Ward connection method is presented. For the first time, Additive Regularization of Topic Models (ARTM) is considered to determine patterns of clinical pathways. A computational experiment based on data on the routes of patients with a diagnosis of sepsis, placed in the public domain.

The results obtained make it possible to conduct a preliminary assessment of the clinical pathways of patients of any event log, to identify bottlenecks in the system and to visualize process maps of the medical institution.

The described approaches to segmentation of the input heterogeneous flow serve as the foundation for the further development of a simulation model of a medical institution and for providing advisory services to patients, for example, chat bots on web pages of a clinic for consulting services.

Medical institutions that are the first to implement these technologies will certainly have a competitive advantage. Consequently, managers and other interested parties will be able to gain access to complete information, which will allow them to make more informed decisions. ■

References

1. Ilyushin G.Ya., Limanskij V.I. (2015) Development of the patient flow management system. *Systems and Approaches of Informatics*, vol. 25, no 1, pp. 186–197 (in Russian).
2. Azanov V.G. (2016) Structural-functional model of patient flow management. *Systems and Approaches of Informatics*, vol. 26, no 1, pp. 13–29 (in Russian).
3. Kinsman L., Rotter T., James E., Snow P., Willis J. (2010) What is a clinical pathway? Development of a definition to inform the debate. *BMC Medicine*, vol. 8, no 31. DOI: 10.1186/1741-7015-8-31.
4. Pearson S.D., Goulart-Fisher D., Lee T.H. (1995) Critical pathways as a strategy for improving care: Problems and potential. *Annals of Internal Medicine*, vol. 123, no 12, pp. 941–948.
5. Wakamiya S., Yamauchi K. (2009) What are the standard functions of electronic clinical pathways? *International Journal of Medical Informatics*, vol. 78, no 8, pp. 543–550. DOI: 10.1016/j.ijmedinf.2009.03.003.
6. Veselý A., Zvárová J., Peleska J., Buchtela D., Anger Z. (2006) Medical guidelines presentation and comparing with electronic health record. *International Journal of Medical Informatics*, vol. 75, no 3–4, pp. 240–245. DOI: 10.1016/j.ijmedinf.2005.07.016.
7. Rojas E., Munoz-Gama J., Sepúlveda M., Capurro D. (2016) Process mining in healthcare: A literature review. *Journal of Biomedical Informatics*, no 61, pp. 224–236. DOI:10.1016/j.jbi.2016.04.007.
8. Huang Z., Lu X., Duan H. (2012) On mining clinical pathway patterns from medical behaviors. *Artificial Intelligence in Medicine*, vol. 56, no 1, pp. 35–50. DOI: 10.1016/j.artmed.2012.06.002.
9. Rakocevic G., Djukic T., Filipovic N., Milutinović V. (2013) *Computational medicine in data mining and modeling.* N.Y.: Springer. DOI: 10.1007/978-1-4614-8785-2.
10. Ahmad M. A., Teredesai A., Eckert C. (2018) Interpretable machine learning in healthcare. Proceedings of the *2018 IEEE International Conference on Healthcare Informatics (ICHI), New York, NY, USA, 4–7 June 2018*, pp. 447–447. DOI: 10.1109/ICHI.2018.00095.
11. Rotter T., Kinsman L., James E.L., Machotta A., Gothe H., Willis J., Snow P., Kugler J. (2010) Clinical pathways: effects on professional practice, patient outcomes, length of stay and hospital costs: Cochrane database of systematic reviews and meta-analysis. *Evaluation & the Health Professions*, vol. 35, no 1, pp. 3–27. DOI: 10.1177/0163278711407313.
12. Prodel M. (2017) *Process discovery, analysis and simulation of clinical pathways using health-care data.* Université de Lyon. Available at: https://tel.archives-ouvertes.fr/tel-01665163/document (accessed 25 November 2019).
13. Huang Z., Dong W., Ji L., Gan C., Lu X., Duan H. (2014) Discovery of clinical pathway patterns from event logs using probabilistic topic models. *Journal of Biomedical Informatics*, no 47, pp. 39–57. DOI: 10.1016/j.jbi.2013.09.003.
14. Lin F., Chou S., Pan S., Chen Y. (2001) Mining time dependency patterns in clinical pathways. *International Journal of Medical Informatics*, vol. 62, no 1, pp. 11–25. DOI: 10.1016/S1386-5056(01)00126-5.
15. Cote M.J., Stein W.E. (2007) A stochastic model for a visit to the doctor's office. *Mathematical and Computer Modelling*, vol. 45, no 3–4, pp. 309–323. DOI: 10.1016/j.mcm.2006.03.022.
16. Zhang Y., Padman R., Patel N. (2015) Paving the cowpath: Learning and visualizing clinical pathways from electronic health record data. *Journal of Biomedical Informatics*, no 58, pp. 186–197.
17. Blei D.M., Ng A.Y. Jordan M.I. (2003) Latent Dirichlet allocation. *The Journal of Machine Learning Research*, no 3, pp. 993–1022.
18. Fernández-Llatas C., Benedi J.-M., García-Gómez J.M., Traver V. (2013) Process mining for individualized behavior modeling using wireless tracking in nursing homes. *Sensors (Basel)*, vol. 13, no 11, pp. 15434–15451. DOI: 10.3390/s131115434.
19. van der Aalst W.M.P. (2011) *Process mining: Discovery, conformance and enhancement of business processes.* Springer. DOI: 10.1007/978-3-642-19345-3.
20. van der Aalst W.M.P. (2018) Process mining and simulation: A match made in heaven! Proceedings of the *50th Computer Simulation Conference (SummerSim 2018). Bordeaux, France, 9–12 July 2018.* DOI: 10.22360/summersim.2018.scsc.005.
21. van der Aalst W.M.P. (2016) *Process mining: Data science in action.* Berlin: Springer-Verlag.

22. van der Aalst W.M.P. (2011) Process mining manifesto. *Business Process Management Workshops*. Springer, pp. 169–194. DOI: 10.1007/978-3-642-28108-2_19.

23. Fernández-Llatas C., Meneu T., Benedí J.M., Traver V. (2010) Activity-based process mining for clinical pathways computer aided design. Proceedings of the *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology. Buenos Aires, Argentina, 31 August – 4 September 2010*, pp. 6178–6181. DOI: 10.1109/IEMBS.2010.5627760.

24. Kovalchuk S.V., Funkner A.A., Metsker O.G., Yakovlev A.N. (2018) Simulation of patient flow in multiple healthcare units using process and data mining techniques for model identification. *Journal of Biomedical Informatics*, no 82, pp. 128–142.

25. Williams R., Buchan I., Prosperi M., Ainsworth J. (2014) Using string metrics to identify patient journeys through care pathways. Proceedings of the *AMIA Annual Symposium, Washington, DC, USA, 15–19 November 2014*, pp. 1208–1217.

26. Kaufmann L., Rousseeuw P. (1987) Clustering by means of medoids. *Data analysis based on the L1-norm and related methods*, pp. 405–416.

27. Ward J.H. (1963) Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, vol. 58, no 301, pp. 236–244.

28. Zaitsev R.D., Britkov V.B. (2015) The use of the R language for multidimensional clustering of time series in order to analyze the dynamics of scientific and technological development. *Transactions of the Second Youth Scientific Conference "Problems of Modern Computer Science", Moscow, 29–30 October 2015*, pp. 92–98.

29. Ferreira L., Hitchcock D. (2009) A comparison of hierarchical methods for clustering functional data. *Communications in Statistics – Simulation and Computation*, no 38, pp. 1925–1949. DOI: 10.1080/03610910903168603.

30. Konnov I.V., Kashina O.A., Gilmanova E.I. (2019) Solving the clustering problem by optimization methods on graphs. *Scientific Letters of the Kazan University, Series Physical and Mathematical Sciences*, vol. 161, pp. 423–437 (in Russian). DOI: 10.26907/2541-7746.2019.3.423-437.

31. Boytsov L. (2011) Indexing methods for approximate dictionary searching. *Journal of Experimental Algorithmics*, vol. 16, no 1, article no 1.1. DOI: 10.1145/1963190.1963191.

32. Rousseeuw P.J. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65. DOI: 10.1016/0377-0427(87)90125-7.

33. Huang Z., Lu X., Duan H., Fan W. (2013) Summarizing clinical pathways from event logs. *Journal of Biomedical Informatics*, vol. 46, no 1, pp. 111–127. DOI: 10.1016/j.jbi.2012.10.001.

34. Vorontsov K.V., Potapenko A.A. (2014) Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization. *AIST'2014, Analysis of Images, Social Networks and Texts. Communications in Computer and Information Science*. Springer, pp. 265–267.

About the authors

Elizaveta S. Prokofyeva

Doctoral Student, Department of Innovation and Business in Information Technologies,
National Research University Higher School of Economics,
20, Myasnitskaya Street, Moscow 101000, Russia;

E-mail: prokofyeva.liza@gmail.com

ORCID: 0000-0003-1322-2932

Roman D. Zaytsev

Senior Expert for Data Analysis, FORS Group,
3, Trifonovskiy Tupik Street, 129272 Moscow, Russia;

E-mail: roman.zaitsev@fors.ru

ORCID: 0000-0002-8313-3727