Artificial Intelligence in Medicine 87 (2018) 34–48

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

A two-step approach for mining patient treatment pathways in administrative healthcare databases

Ahmed Najjar[a,*], Daniel Reinharz[b], Catherine Girouard[c], Christian Gagné[a]

[a] Laboratoire de vision et systèmes numériques, Département de génie électrique et de génie informatique, Université Laval, Québec, QC G1V 0A6, Canada
[b] Laboratoire de simulation du dépistage, Département de médecine sociale et préventive, Université Laval, Québec, QC G1V 0A6, Canada
[c] CISSS Chaudière-Appalaches, Secteur Alphonse-Desjardins, Lévis, QC G6V 3Z1, Canada

ARTICLE INFO

Article history:
Received 13 November 2016
Received in revised form 8 March 2018
Accepted 22 March 2018

Keywords:
Process clustering
Process mining
Mixed variables
HMM
k-Prototypes
Healthcare databases
Medical treatment process

ABSTRACT

Clustering electronic medical records allows the discovery of information on healthcare practices. Entries in such medical records are usually composed of a succession of diagnostics or therapeutic steps. The corresponding processes are complex and heterogeneous since they depend on medical knowledge integrating clinical guidelines, the physician's individual experience, and patient data and conditions. To analyze such data, we are first proposing to cluster medical visits, consultations, and hospital stays into homogeneous groups, and then to construct higher-level patient treatment pathways over these different groups. These pathways are then also clustered to distill typical pathways, enabling interpretation of clusters by experts. This approach is evaluated on a real-world administrative database of elderly people in Québec suffering from heart failures.

1. Introduction

Health Informatics is a rapidly growing field that is concerned with applying computer science and information technology to medical and health data [3]. The huge and increasing amount of data in the healthcare databases is a source of information of considerable value for those concerned by how to improve health outcomes and how to better control the rising cost of healthcare. Administrative healthcare databases can unveil the constraints of reality, capturing elements from a great variety of real medical care situations. However, the existing databases contain a huge amount of data. For example, an average of 80–86 million medical services are provided annually to the Quebec population. For 2005–2006, there were nearly 714,000 acute care and more than 465,000 one-day surgeries, for which rather details information has been captured [15]. Yet, there is a lack of effective analysis tools to explore and extract the potential rather detailed information that can be brought by analyzing these databases. Exploring the databases to their full potential represents a challenge as it requires complex preprocessing steps and appropriate developed methods.

Process mining [21,22] is an emerging field in the use of all this information for the benefit of the population. This approach is based on the assumption that each event is an instance of a given process. It consists in developing techniques that allow the production of meaningful clusters of similar types of behavior that are grouped together. In healthcare, it is judicious to distinguish between organizational processes and the medical treatment process. As mentioned by Lenz and Reichert [12], the organizational processes help to coordinate inter-operating healthcare professionals and organizational units. The medical treatment process is linked to the patient. The patient specific medical treatment process depends on case-specific decisions. Decisions are made by interpreting patient specific data according to medical knowledge. This decision process is very complex because it includes medical knowledge, medical guidelines and the individual experience of physicians. Yet, over the last decade, some researchers have proposed techniques and methodologies to cluster such healthcare processes.

Although the medical processes are composed of complex events, previous works found in the literature relied on predetermined events. Moreover, many of the proposals made are not suitable for large-scale datasets. Our contribution is to propose a scalable approach based on a two-step clustering method to handle such data. As the first step, we cluster the medical events that are composing the processes. As the second step, we construct higher-level medical treatment pathways and cluster them to iden-

* Corresponding author.
E-mail addresses: ahmed.najjar.1@ulaval.ca (A. Najjar), daniel.reinharz@fmed.ulaval.ca (D. Reinharz), catherine.girouard83@gmail.com (C. Girouard), christian.gagne@gel.ulaval.ca (C. Gagné)

https://doi.org/10.1016/j.artmed.2018.03.004
0933-3657/

tify typical ones. Furthermore, we use a real-world administrative database of elderly people in Québec suffering from heart failures to demonstrate the applicability and the scalability of the proposed methodology.

This paper is organized as follows. We first define the problem in Section 2, followed by an overview of relevant medical process clustering approaches in Section 3. We then present our methodology and the proposed algorithm in Section 4. We propose our methodology to analyze results in Section 5. A case study on the clustering of an administrative healthcare database validating our methodology and algorithm follows in Section 6. Finally, we conclude the paper in Section 7 with some consideration on the relevance and importance of this work.

2. Problem definition

2.1. Case study

The health system and social services in the Province of Quebec are mainly public. The RAMQ (Régie de l'assurance-maladie du Québec) acts as the health insurer for Quebec residents who are covered by a universal public health insurance program (virtually 100% of the people living in the province). The MSSS (Ministère de la Santé et des Services sociaux du Québec) is responsible for the administration of health and social services in the province. Almost all physicians (98%) are participants of the public health system and are exclusively paid by the RAMQ. The health system in Quebec is composed of institutions, community organizations, clinics and private cabinets, all publicly funded. The RAMQ therefore records information on the vast majority of medical, and social services provided to the population.

Heart failure disease is a significant cause of the use of health care resources [13,14]. This disease is a clinical syndrome that normally requires health care to be provided by both specialist and family physicians. Practitioners have guidelines to help them to diagnose and treat the disease. Yet, guidelines that most often apply to an average patient, might not always be relevant for individual cases. This is not surprising given that every patient differs from the "average patient" and has his own peculiar needs. Every health care provider functions with his own biases. The management of heart failure is therefore complex and reflects the integration of decisions made by many actors from different disciplines, but also by the patients [14]. Indeed, in real-life, decisions made on care to be provided often differ from evidence-based recommendations [10].

2.2. Aim

The main goal of our work is to propose a methodology allowing the construction of patient treatment pathways from administrative relational databases, to cluster them in homogenous clusters and to analyze and describe these clusters. Thereby, our approach allows the extraction of latent patterns from patient treatment pathways. As a proof of concept, we are studying medical services given to patients over 65 years old, who live in the province of Quebec (Canada) and who suffer from a heart failure disease.

2.3. Datasets

For this purpose, we have been granted access to administrative health care databases of the RAMQ and MSSS. These databases record all medical acts from health care professionals that are covered by the RAMQ and all hospital stays taking place in the Province of Quebec. Our intent is to exploit these data to reconstruct and cluster patient treatment pathways for elderly people suffering from this disease. We have two databases. The first one contains data for all hospital stays that occurred in Quebec. These hospitals provide general and specialized care. These data, compiled by the hospitals, relate to acute care (physical and mental) and one-day surgery. The data are organized into 5 tables: hospital stays, diagnostics, services, intensive care, and interventions. The second database contains information on physician fees for medical services according to the health insurance plan administered by the RAMQ. It contains the table of medical services and the table of patient information. For our experiments, we selected individuals in these databases with at least one diagnosis of heart failure (i.e., ICD-10 diagnosis codes 428.0, 428.1, or 428.9) made between January 1, 2000 and December 31, 2005. We rejected individuals who were not 65 years or older at the earliest consultation date or earliest departure date from hospital stays. We obtained 180,027 individuals. We were interested in building and mining the treatment pathways between January 1, 2000 and December 31, 2009 for these patients.

3. Related work

In recent years, researchers have become interested in discovering process models from unlabeled event logs. Some proposals have been made for the clustering of medical processes. Ferreira et al. [6] proposed sequence clustering as an approach to deal with processes. They considered that this approach is a good candidate for process clustering. Indeed, sequence clustering is a collection of techniques with the goal of partitioning a number of sequences into meaningful groups. To reach this aim, they used a clustering algorithm based on first-order Markov chains. Afterwards, many studies based on the use of this approach have been proposed to cluster medical processes. For instance, Rebuge et al. [19] proposed a methodology based on a first-order Markov chain to cluster processes composed of the care events occurring in the emergency department of a hospital. They were interested in the radiology workflow of emergency patients, which is an organizational healthcare process. The events used in their studies include 12 different tasks: the exam request and the 11 possible states of the exam. Furthermore, Elghazel et al. [5] assumed that the clinical pathway is a sequence of hospital stays and each hospital stay is characterized by two qualitative items. Therefore, they used a similarity based approach to compute the clinical pathway dissimilarities and a method based on graph coloring to cluster pathways. The behavior of each cluster is then governed by a finite-state Markov chain model.

Huang et al. [8,9] have applied latent Dirichlet allocation (LDA) to discover latent patterns as a probabilistic combination of clinical activities. They assumed that a patient clinical pathway is represented by a mixture of treatment patterns. They applied LDA to two specific care flow logs concerning intracranial hemorrhage and cerebral infarction, extracted from a hospital information system. The model gives the clinical activity density estimation for each pattern, from which the probabilistic association between an activity and a pattern can be obtained.

All of these studies rely on processes composed of relatively simple and well-defined events. Moreover, each of them was concerned with a specific aspect of a patient's pathway but not with providing an overall view of the care provided. Rebuge et al. [19] worked on organizational processes but not on the patient treatment pathways. Huang et al. [8,9] and Elghazel et al. [5] worked on clustering patient treatment pathways. Huang et al. were interested in the patient treatment pathway in hospital stays which gives a local and micro view in a specific site. Elghazel et al. were interested in a succession of hospital stays. This work provides a more macro and global view but does not involve the complete patient treatment pathways since there are services other than hospital

stays. In addition, the method used by Elghazel et al. does not scale well as it uses pairwise dissimilarity. On the other hand, the LDA method can give a model not optimally informative because of low activity probability values that do not differentiate in a clear way one pattern from another.

Business processes clustering studied similar questions in a different application context. Bose and van der Aalst [1] sought to determine multiple feature sets from the traces. An agglomerative hierarchical clustering technique was applied using the Euclidean distance over these feature sets, based on the minimum variance criteria. Two main drawbacks stem from this proposal: (1) the curse of dimensionality over the feature sets, which grows quickly according to the number of activities; and (2) the loss of the order of events execution, also reflected by the lack of context over the process execution. The same authors also proposed another approach to derive the operations costs for a generic edit distance based on 3-grams sets over the event log [2]. However, even if this method solves the lack of context information issue, it still suffers from the curse of dimensionality induced by the use of 3-g sets. De Weerdt et al. [4] proposed an iterative clustering algorithm that mines process models and computes a metric for each trace and model at every step. Contrary to our proposal, the methods proposed in these three papers [1,2,4] are not suitable for large datasets.

Ha et al. [7] proposed to extract a set of features based on graph representation, over which they made a cluster using the *k*-means algorithm. For their part, Nguyen et al. [17] used heterogeneous information networks (HIN) as a representation for process traces, where nodes have some types which are resources, events and traces. They introduced a notion of meta-path to define similarity measures, combining trace-typed nodes in the HIN and the edit distance. These two papers can handle large datasets since they use *k*-means. However, they face the curse of dimensionality and the dependence to the defined similarity problems.

Thaler et al. [20] conducted a survey of clustering techniques used for clustering business processes. They reported that most approaches use an abstract trace representation and most often rely on a Euclidean distance. We noticed that all papers covered by this survey are applied on data with predefined and non-complex events, mainly consisting of business processes, with no large-scale datasets. Also, the only model-based approach relies on a mixture of first-order Markov models using the expectation-maximization (EM) algorithm, also proposed by Rebuge et al. [19].

Hidden Markov models (HMMs) are a common stochastic method for modeling sequential data and have been proven to be efficient and successful in many tasks such as speech recognition [18]. We thus propose to use this method to model the behavior of each cluster as a pattern governed by a HMM. To our knowledge, the work reported here is the first application to healthcare of HMM clustering to handle patient treatment pathways.

4. Proposed methodology for clustering process of complex objects

Our project stems from three main observations. First, patient treatment pathways consist of a succession of diagnostic or therapeutic steps linked to a patient. Second, administrative healthcare databases contain observational data, collected for purposes other than data analysis. They have seldom been used to analyze patient treatment pathways, although they are a rich source of information on health services. Third, process clustering has received increased attention over the years since it allows clustering of the execution traces contained in an event log generated by many latent processes.

Our proposed approach has been designed given the interest of generating clusters of patient treatment pathways composed of complex events. These events differ according to their categories and their variables. We can deal with complex events described by a mix of variable types. In fact, the treatment patient process is an ordered sequence of sub-events. Hence, these sequences have components which are themselves complex objects and may also be clustered. Therefore, we propose a bottom-up approach which can handle this type of data. Our methodology is a two-level clustering method. At the first level, we cluster event datasets and label them according to their categories and clusters. Then, these event labels are used to build process abstraction values. At the second level, we propose the use of HMMs to cluster process values. Fig. 1 shows steps and levels of this approach. In what follows we explain in more detail these two levels.

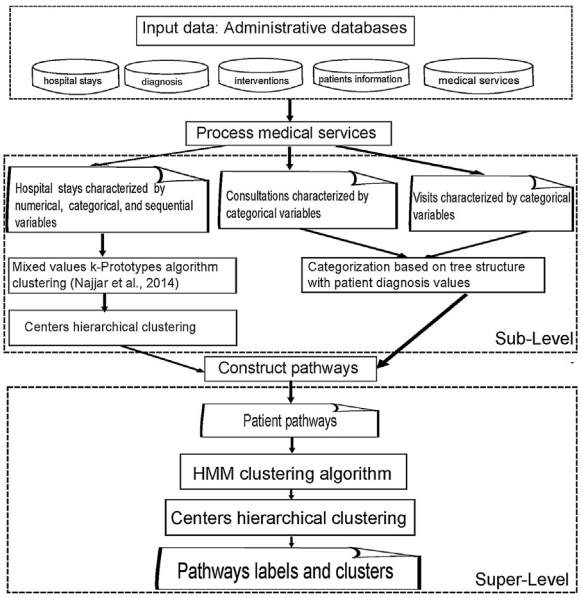


Fig. 1. Description of our approach.

4.1. Sub-level methodology

The aim of this level is to propose a methodology to build complex events from relational databases, to cluster them and to reconstruct a patient's treatment pathway over a specific period of time. First, the event types that composed the process must be identified. For treatment processes, the medical services given to a patient in the treatment process can be categorized into three categories. These categories are:

- **Visits**: defined as a service provided by physicians in ambulatory care without a reference by another physician;
- **Consultations**: defined as a service provided by physicians in ambulatory care following a reference by another physician;
- **Hospital stays**: defined as a service given in the context of a hospitalization of at least one night.

Each service presents an event in the treatment process. It is characterized by its categories and its associated variables. The consultation and visit entities are characterized only by categorical variables, whereas hospital stay entities are more complex. Each hospital stay is described by mixed variable types, with numerical, categorical, and multivalued categorical variables. Databases are used to create the sets of complex entities.

The next step is to cluster these complex objects using the appropriate algorithm according to their variable types. This step aims to

reduce the complexity of the process logs by replacing complex objects by their categories and labels.

Consultations and visits are characterized only by categorical variables. Thereby, we adopted a categorization based on a tree structure and on patient diagnosis codes. We derived subsets by splitting the input objects set according to the diagnostic code of the International Classification of Diseases ICD-10. This classification is a hierarchy composed of chapters that contain three-digit category blocks. The first three-digit diagnosis code allows the determination of its chapter, its blocks and its category. If the subset size exceeds a fixed threshold of a number of objects, we continue the subset partition by dividing the subset into object blocks. Afterwards, if the obtained subset size exceeds the fixed threshold, we divide it into object categories. For example, the I10.0 disease code belongs to diseases of the circulatory system chapter. This chapter covers codes with the first three-digits between I00 and I99. It is divided into homogeneous blocks of three-character categories. An example of a block is hypertensive diseases that contains codes with the first three-digits between I10 and I15. An example of a category in this block is essential hypertension with I10 as the first of the three-digit medical conditions.

Hospital stay entities are more complex. Each hospital stay is described by a mixture of variable types, with numerical, categorical, and multivalued categorical variables. An example of this type of variable is the set of diagnostic codes corresponding to one hospital stay, which can have values such as {O48001,Z370,O62101}. To cluster the hospital stays set we first apply the mixed values k-prototypes algorithm we previously proposed. Thereafter we apply hierarchical clustering for cluster centers given by the k-prototypes algorithm to obtain final clustering.

The k-prototypes algorithm takes as input hospital stay entities. It determines the nearest center of each object according to a given dissimilarity measure and allocates it to the nearest center. To compute this dissimilarity, a representation on Bag-of-Words (BoW) in a defined projection space is used for each multivalued variable. It updates the previous and current centers of the object. The algorithm repeats the processing of all objects until no object is reallocated or another stopping criterion is reached.

After completing this first clustering, we cluster the centers $c_1, \ldots, c_K$ by hierarchical clustering with average linkage criterion to obtain the final clustering. Therefore, each center c_k is presented as $(c_{k,1}, \ldots, c_{k,r}, \ldots, c_{k,q}, \ldots, c_{k,m})$, where the first r elements are numerical values, the next $(q-r)$ elements are categorical values, and the remaining elements are BoW representations of multivalued categorical values. First, the pairwise center distances are calculated by the following dissimilarity measure:

$$d(c_i, c_k) = \frac{r}{m}\frac{\sum_{l=1}^{r}(c_{i,l}-c_{k,l})^2}{\sum_{j=1}^{K}\sum_{l=1}^{r}(c_{i,l}-c_{j,l})^2} + \frac{q-r}{m}\frac{\sum_{l=r+1}^{q} I(c_{i,l}, c_{k,l})}{\sum_{j=1}^{K}\sum_{l=r+1}^{q} I(c_{i,l}, c_{j,l})} + \frac{m-q}{m}\frac{\sum_{l=q+1}^{m} 1-(\cos(c_{i,l}, c_{k,l}))}{\sum_{j=1}^{K}\sum_{l=q+1}^{m} 1-(\cos(c_{i,l}, c_{j,l}))},$$

with $I(x, y) = 1$ if $x = y$ and $I(x, y) = 0$ otherwise.

From this pairwise distance matrix, we apply an agglomerative hierarchical clustering algorithm with average linkage to cluster these centers and assign objects to the new clusters. In the beginning, each center represents one cluster. This algorithm merges cluster pairs as it moves up the hierarchy. To obtain the new clusters, we cut the dendrogram at a chosen height which gives the new partition. The Scipy clustering package[1] was used.

From entity clustering results, we build the process as a succession of complex object labels. The set of processes presents the process log which serves as input for the super-level. An example of the process obtained after the application of this methodology in this level is given by $\{s51 \rightarrow c17C1M2 \rightarrow v2C1M5 \rightarrow v2C1M2 \rightarrow v2C1M2 \rightarrow s50\}$. This pathway means that the patient has a hospital stay labeled s51 followed by a consultation labeled c17C1M2 which in turn is followed by a visit labeled v2C1M5, and so on.

4.2. Super-level methodology

At this level, the aim is to cluster the set of processes obtained by the sub-level. The HMMs using the method proposed by Knab et al. [11] followed by hierarchical clustering where average linkage criterion are used to reach this goal. The HMM based clustering algorithm is described by Algorithm 1. The GHMM library[2] was used as an implementation of the HMM algorithms.

Algorithm 1. HMM for clustering treatment processes.

$$\mathcal{L}(\Lambda^{(0)}) = \sum_{k=1}^{K}\sum_{i=1}^{n_k} \log[P(\mathbf{S}_i|\lambda_k^{(0)})]$$

input $\mathcal{S} = \{\mathbf{S}_1, \ldots, \mathbf{S}_N\}$: a set of sequences, T: maximum number of iterations
output b: sequences labels

Initialize parameters of K HMMs $\Lambda^{(0)} = \{\lambda_0^{(0)}, \ldots, \lambda_K^{(0)}\}$
Compute the log-likelihood $\mathcal{L}(\Lambda^{(0)})$ given by

$$\mathcal{L}(\Lambda^{(0)}) = \sum_{k=1}^{K}\sum_{i=1}^{n_k} \log[P(\mathbf{S}_i|\lambda_k^{(0)})]$$

while $\left(\frac{|\mathcal{L}(\Lambda^{(t)}) - \mathcal{L}(\Lambda^{(t-1)})|}{\mathcal{L}(\Lambda^{(t-1)})} \geq \epsilon\right) \wedge (t \leq T)$ **do**
Generate a partition of the processes by assigning each process to the HMM giving the maximum emission probability $\mathcal{C} = \{C_1, \ldots, C_K\}$ where $C_k = \{\mathbf{S}_i | k = \text{argmax}_j P(\mathbf{S}_i|\lambda_j^{(t-1)})\}$ is the sequences assignment to the kth HMM, $k = 1, \ldots, K$
Compute new parameters Λ^t using the Baum-Welch algorithm with initial parameters Λ^{t-1} and sequences assignment to the HMMs
Compute emission probabilities $P(\mathbf{S}_i|\lambda_k^{(t)})$ for $k = 1, \ldots, K$; $i = 1, \ldots, n_k$ using the forward algorithm
Compute the log-likelihood $\mathcal{L}(\Lambda^{(t)})$ given by

$$\mathcal{L}(\Lambda^{(t)}) = \sum_{k=1}^{K}\sum_{i=1}^{n_k} \log[P(\mathbf{S}_i|\lambda_k^{(t)})]$$

$t \longleftarrow t + 1$
end while
Compute labels $b_{i,k}^{(t)}$ for $i = 1, \ldots, n$ and $k = 1, \ldots, K$ given by

$$b_{i,k}^{(t)} = \begin{cases} 1 & \text{if } k = \text{argmax}_j P(\mathbf{S}_i|\lambda_j^{(t)}) \\ 0 & \text{otherwise} \end{cases}$$

The purpose of Algorithm 1 is to fit a HMM for each pathway cluster. Since the convergence of the algorithm depends on its initialization, we run the algorithm many times. Our initialization policy is based on a Dirichlet distribution and depends on the α parameter of the distribution. In the initialization step, the α parameter (same for all dimensions) is fixed. We generate initial parameters of all HMM using the Dirichlet distribution with this α. The initial sequences assignment is given by the k-means algorithm with a cosine distance and a Bag-Of-Words (BoW) representation. We thus use the BoW representation to project the pathways in the events space and we apply the k-means algorithm to assign each

[1] http://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html#module-scipy.cluster.hierarchy.
[2] http://ghmm.org.

pathway to clusters. We train each HMM in each cluster by the Baum-Welch algorithm to obtain model parameters $\Lambda^{(0)}$.

After the initialization step, the algorithm proceeds in two steps. The first step consists in assigning each patient pathway to the HMM with the highest emission probability. In order to do that, it computes for each pathway value the emission probability $P(\mathbf{S}_i|\lambda_k^t)$ given HMM parameters λ_k^t. This probability is computed by the so-called forward algorithm [18]. As a result we generate a partition of the processes by assigning each process to the HMM giving the maximum emission probability. In the second step, the HMMs are updated according to the log sequences assigned to their clusters. In this step, the algorithm calculates the model parameters in each cluster by training each HMM in each cluster by the well-known Baum-Welch algorithm to obtain model parameters $\lambda_k^{(t)}$ using the previous pathway assignments and previous HMM parameters $\lambda_k^{(t-1)}$. These two steps are repeated until convergence or resource exhaustion.

We run the algorithm for the α parameter (same for all dimensions) between 0.1 and 1 by 0.1 increments many times and we keep the best result obtained according to the log-likelihood indicator.

After the HMM clustering, we obtain a pathways partition into clusters. The last step in this level is to build a hierarchy of these clusters and seek a new clustering in a much smaller number of clusters. In order to achieve this goal, we represent each pathway S_i as a vector of event weights $(w_{i,1}, \ldots, w_{i,L})$, where L is the number of different events in all pathways. Weights $w_{i,j}$ are computed by

$$w_{i,j} = \frac{n_{i,j}}{|S_i|}$$

where $n_{i,j}$ is the number of occurrences of event e_j in the pathway and $|S_i|$ is the pathway length. We represent the pathways center for cluster C_k as $c_k = (c_{k,1}, \ldots, c_{k,L})$. Each $c_{k,j}$ value is computed by

$$c_{k,j} = \frac{\sum_{S_i \in C_k} w_{i,j}}{n_k}$$

Then, we compute the pairwise center distances by the cosine distance and we apply the hierarchical clustering algorithm with an average linkage to cluster these centers and assign pathways to the new clusters.

5. Proposed approach for results analysis

Administrative healthcare databases contain observational data, collected for administrative purposes rather than data analysis. They are a rich source of information on health services and associated processes provided to patients. Being able to analyze and visualize the data extracted from the administrative databases for the health services provided would provide an insight into latent patterns.

As a result of the process clustering, we obtain a partition of processes. This representation is not very informative in its current state. Thus, since the resulting group is a super-level of processes and the sub-events are themselves complex objects, we need an approach to better understand the results. So, we propose an analysis and visualization approach to discover more information contained in these results.

5.1. Frequency analysis

In addition to the process partition obtained by the clustering algorithm, we need to understand this result and analyze it. As one tool for analysis, we define the following measures. The first measure is the service support in each cluster. It is calculated as the ratio of the number of processes which contain the service by the number for all processes in this cluster, as given by Eq. (1):

$$\tau(s) = \frac{\sum_{P_i \in C_k} u_i(s)}{N_k}, \tag{1}$$

where $u_i(s) = 1$ if $s \in P_i$ and 0 otherwise.

The second measure is the service density in each cluster. It is the ratio of the number of service occurrences by the total length of all processes in the cluster, as given by Eq. (2):

$$\rho(s) = \frac{\sum_{P_i \in C_k} o_i(s)}{\sum_{i=1}^{N_k} |P_i|}, \tag{2}$$

where $o_i(s)$ is the number of occurrences of s in pathway P_i, and $|P_i|$ is the pathway length.

The last measure is the transition between services in each cluster. The transition frequency from service s to service s' is defined as the number of occurrences of the pair (s, s') by the total length of all processes in cluster, as given by Eq. (3):

$$\mathcal{A}_k(s, s') = \frac{\sum_{i=1}^{N_k} \tau_i(s, s')}{\sum_{i=1}^{N_k} |P_i|}, \tag{3}$$

where $\tau_i(s, s')$ is the number of transitions from state s to state s' in the process P_i and $|P_i|$ is the length of process P_i.

5.2. Model discovery

To derive a useful model from clustering results, we propose a visualization approach to discover the process model in each cluster. This approach is based on abstraction and pruning of undesired details. Regarding abstraction, for consultation and visits, we remain at the level of chapter for their clustering labels. As for pruning, we are only interested in the events that exceed a fixed threshold of density in the pathways cluster. Then, we remove the undesired events and we use a transitive relation to update the transition between the existing events. For example, if we have A, B, C and D as events for a pathway described by $\{A \rightarrow B \rightarrow C \rightarrow D\}$ and if B and C events are removed, the pathway used for model discovery becomes $\{A \rightarrow D\}$. In model visualization each different event label is represented as a node and the transition as an edge. We compute the number of transitions between two different events in all pathways in the cluster and this number is considered as the weight for the edge between those two nodes.

6. Case study: heart failure in the elderly in Quebec

As proof of concept, we evaluated the proposed approach on the clustering treatment processes of elderly patients over 65 years old who live in the province of Québec (Canada) and are suffering from heart failure, through the access to subsets of the administrative healthcare databases from the RAMQ (universal health insurer for Québec residents) and the MSSS (Ministry of Health). Databases were first preprocessed by gathering various medical services according to three categories: visits, consultations, and hospital stays.

To obtain hospital stay entities, we associated the patient information together with all hospital stay information. Each hospital stay is considered as a first category of complex objects described by a set of numerical, categorical, and multivalued categorical variables. The categorical variables are type of care, origin, and destination. The numerical variable is hospitalization length while the multivalued categorical variables are the sequence of diagnosis, and sequence of interventions that occurred at each hospital stay.

Consultations, which are medical services recommended by a referral doctor who oriented a patient to another doctor, gener-

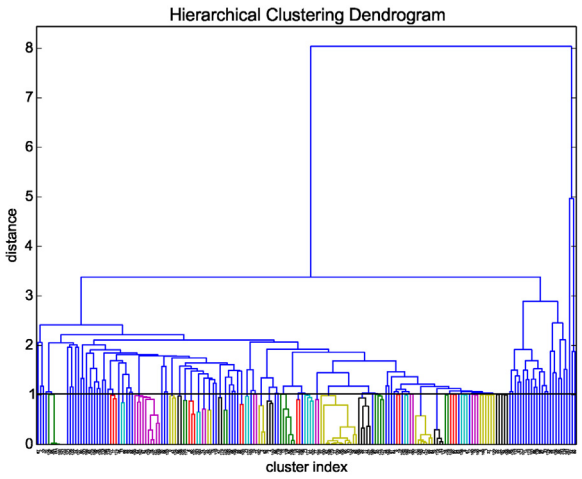


Fig. 2. Hierarchical clustering dendrogram of hospital stay entities clusters.

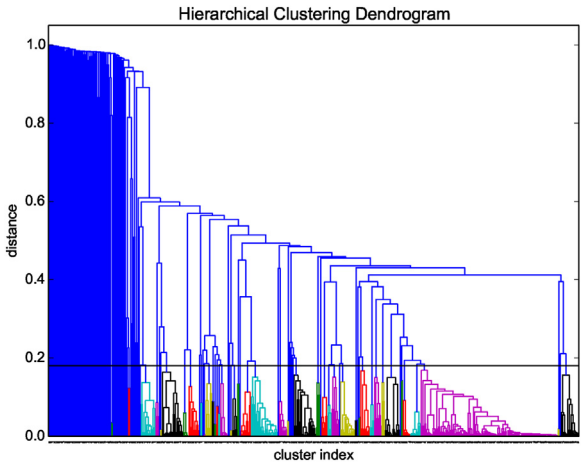


Fig. 3. Hierarchical clustering dendrogram of pathways clusters.

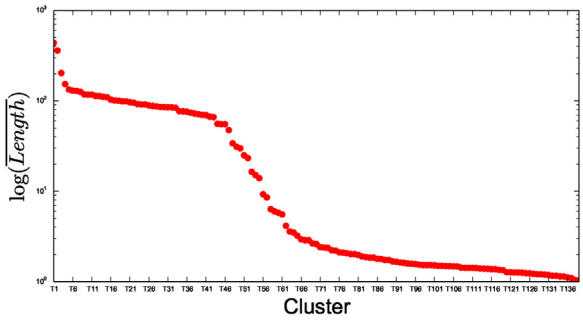


Fig. 4. Log of the average length of the services according to the clusters.

ally a specialist, represent complex objects described by patient information connected with medical act information. They are characterized by categorical variables which are physician specialty, diagnosis code, type of site of contact, and specialty of referent practitioner. Visits are characterized by the same variables minus specialty of referent practitioner.

A cohort of 180,027 individuals allowed us to extract 684,906 hospital stays, 2,594,341 consultations, and 12,510,117 visits that occurred between January 1, 2000, and December 31, 2009. As mentioned in Section 4, to build patient treatment processes, we started by categorizing consultation and visit sets using a tree categorization method based on the diagnosis value. Because we were interested in services given between January 1, 2000 and December 31, 2009, the diagnoses are coded according to two medical classifications: ICD-9 and ICD-10 (the classification changed during these years). To do the categorization, we used twenty-two chapters as described in Table 1. The Codes in the ICD-9 and ICD-10 columns of Table 1 represent the first and the last category codes contained in each chapter. The chapter 0 is dedicated to consultations or visits with indefinite diagnostics coded by white or V999 like radiology consultation.

We fixed the threshold equal to 50,000 objects for consultation and 100,000 for visits. When the number of entities in a chapter exceeds the threshold, we continued categorizing by assigning each entity to one block category. After that, if the number of entities belonging to a block category remained significant, we extended categorization by assigning entities to disease category contained in each block category. We obtained 105 clusters for consultations and 220 for visits. The label for consultations or visits is composed of the letter c for consultation or v for visit followed by the disease chapter number. If we went down to a block category level, we added the letter C and the id number for the block category. If, in addition, we went down to a category level, we added the letter M and the id number of category.

We used the *k*-prototypes algorithm [16] to cluster hospital stay entities into 200 clusters and then we applied hierarchical clustering with average linkage criterion. This gives us the 200 clusters hierarchy described by the dendrogram depicted in Fig. 2. The choice of cut-off value was selected so as to make a trade off between not merging a large number of clusters and having a relatively small number of clusters compared to the initial one. So, the cut-off value was fixed at 1.02. By using this value, 106 clusters were provided, i.e., 106 hospital stay labels. The set of all medical treatment activities included in the pathways were then summarized in 431 service labels. By using the obtained service labels, we built 180,027 patient treatment processes represented as a succession of medical services labels.

To cluster the pathways, we first ran the HMM clustering algorithm with HMMs having 10 hidden states to cluster pathways into 500 clusters. Two repetitions were conducted, where the best value according to log-likelihood was kept. Then, in order to determine the number of clusters to use, we applied hierarchical clustering with an average linkage criterion and we chose 0.18 as the cut-off. We obtained 139 clusters. Fig. 3 gives a hierarchy of 500 pathway clusters and the cut-off level.

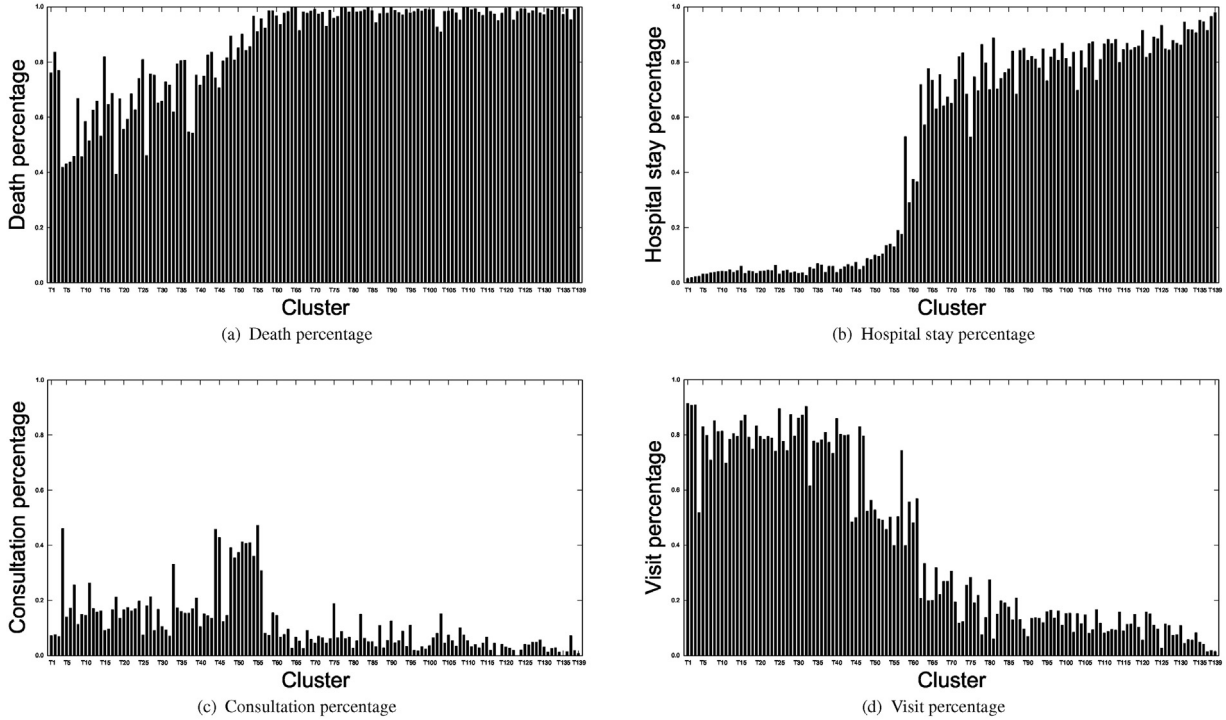
The computation time required to process a database depends on the size of the cohort and the number of groups. It varies from a few minutes to a few days on a modern workstation. More specifically, the time required to cluster our cohort into 200 groups was about one day on one workstation. The time required for clustering the pathways with the HMM algorithm into 500 groups was about 5 days on the same workstation.

6.1. Results and analyses

To see the relevance and benefit of this clustering, statistics on the rate of hospital stays, rate of consultations, rate of visits, mortality rate and average number of medical services in the pathways were first examined. The rates of mortality of the population, the hospital stays, the consultations and the visits were 67.05%, 4.34%, 16.43%, and 79.23% respectively. The average number of medical services in the pathways is 87.71. In order to facilitate the analysis, we then reorganized the pathways clusters in descending order according to the average number of services consumed. Fig. 4 gives the distribution of a common logarithm of this number according to the clusters.

Table 1
Diagnoses chapters.

Chapter	Codes ICD-9	Codes ICD-10	Titles
Chapter 1	000-139	A00-B99	Certain infectious and parasitic diseases
Chapter 2	140-239	C00-D48	Neoplasms
Chapter 3	240-279	E00-E99	Endocrine, nutritional and metabolic diseases
Chapter 4	280-289	D50-D99	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
Chapter 5	290-319	F00-F99	Mental and behavioral disorders
Chapter 6	320-359	G00-G99	Diseases of the nervous system
Chapter 7	360-379	H00-H59	Diseases of the eye and adnexa
Chapter 8	380-389	H60-H99	Diseases of the ear and mastoid process
Chapter 9	390-459	I00-I99	Diseases of the circulatory system
Chapter 10	460-519	J00-J99	Diseases of the respiratory system
Chapter 11	520-579	K00-K99	Diseases of the digestive system
Chapter 12	580-629	N00-N99	Diseases of the genitourinary system
Chapter 13	630-679	O00-O99	Pregnancy, childbirth and the puerperium
Chapter 14	680-709	L00-L99	Diseases of the skin and subcutaneous tissue
Chapter 15	710-739	M00-M99	Diseases of the musculoskeletal system and connective tissue
Chapter 16	740-779	Q00-Q99	Congenital malformations, deformations and chromosomal abnormalities
Chapter 17	780-799	R00-R99	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
Chapter 18	–	V01-Y99	External causes of morbidity and mortality
Chapter 19	800-999	S00-T99	Injury, poisoning and certain other consequences of external causes
Chapter 20	–	P00-P99	Certain conditions originating in the perinatal period
Chapter 21	–	Z00-Z99	Factors influencing health status and contact with health services
Chapter 22	–	U00-U99	Codes for special purposes



(a) Death percentage (b) Hospital stay percentage

(c) Consultation percentage (d) Visit percentage

Fig. 5. Death and service types percentage according to clusters.

To better understand the pathway variants, we analyzed the results from several viewpoints. These analyses allowed the detection of some clusters which were categorically differentiated from others. By examining the variation in the average number of medical services in the pathways in each cluster, we found that there are groups with short pathways, less than 5 services over ten years, which have a high mortality rate greater than 91%. Fig. 5(a) gives the death percentage distribution for each group. This group starts mainly from the patient treatment pathway T62.

To understand what distinguishes these groups from others, we extended the analyses by considering the repartition for each type of service in the clusters. We noted that they are characterized by a percentage of hospital stays that exceeds 53% of total patient services. However, this percentage is lower than 19% for most of the other clusters. Fig. 5(b) shows hospital stays percentage according to clusters. Regarding consultation and visit percentages, we observed some variability between clusters but this variability was not very large except for the clusters that have a low percentage of

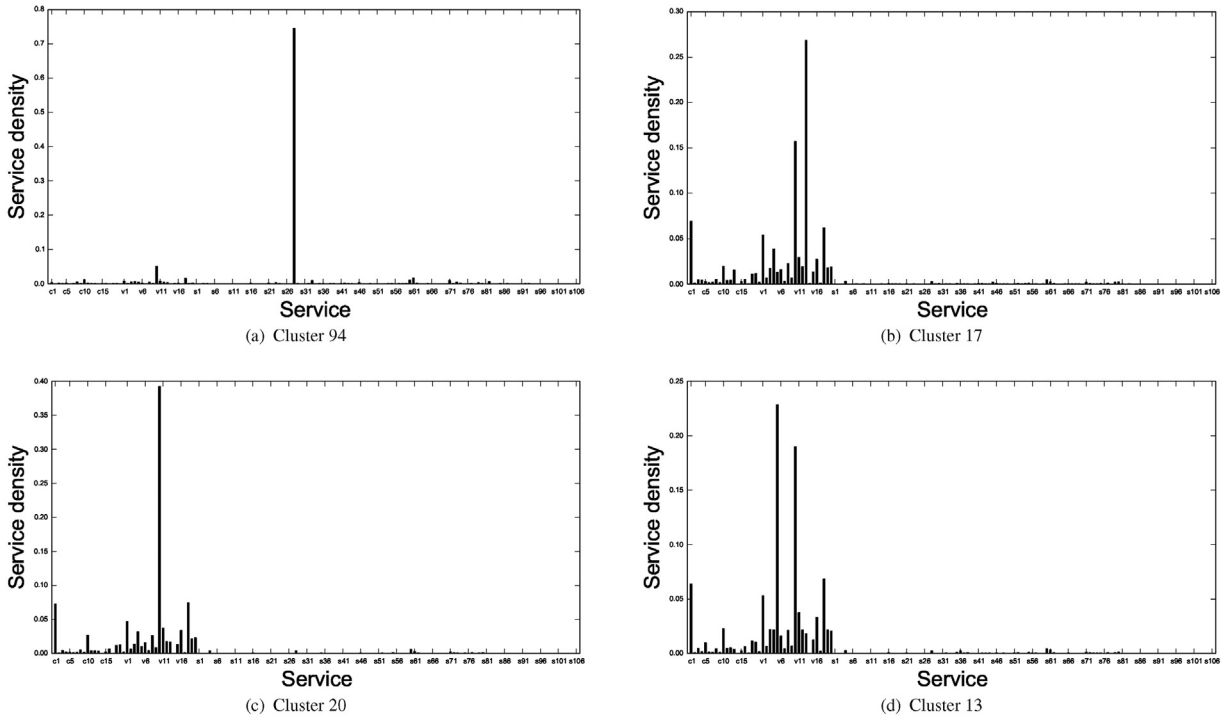


Fig. 6. Densities of all services in some clusters.

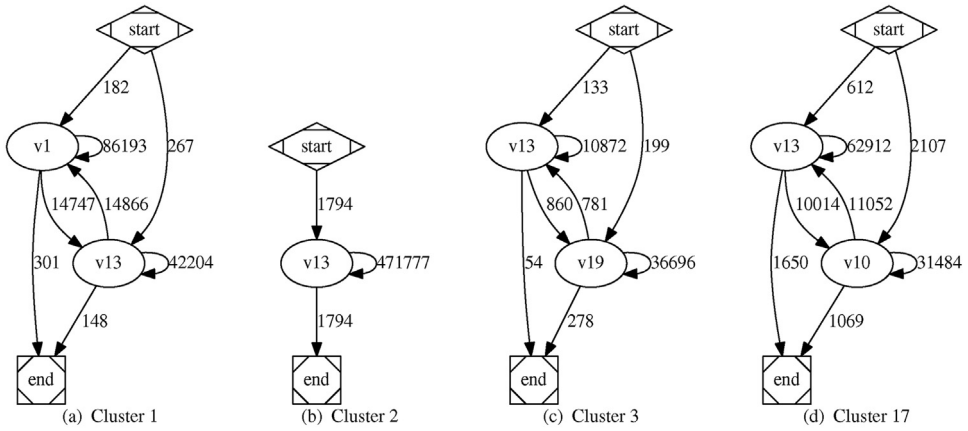


Fig. 7. Discovered models for the kidney problem pattern.

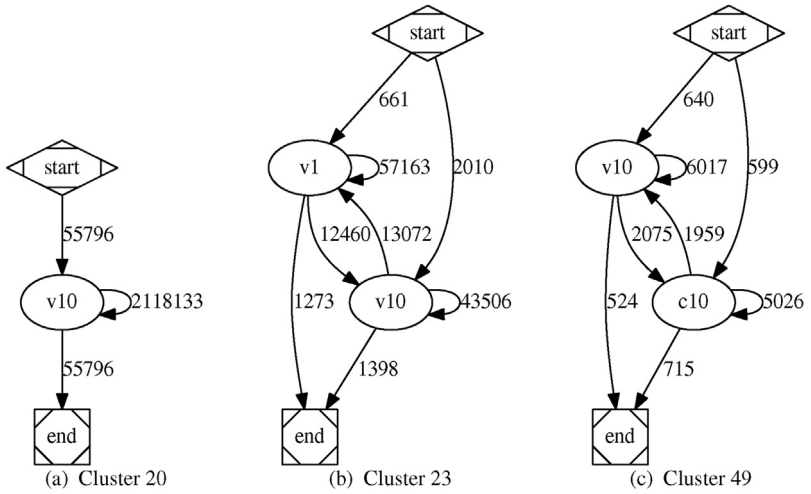


Fig. 8. Discovered models for the heart problem pattern.

Table 2
Description of some clusters with high presence of hospital stays in the pathways.

Clst	Description	Demographics	Most freq service	Cares	Practitioners
C91	**#:** 753 **Avg length:** 1.66 **Death:** 98.67% **Hsp stay:** 82.07% **Consult:** 4.56% **Visit:** 13.37%	**Most freq age:** 85 and over (63.35%) **2nd freq age:** 80–84 (17.93%) **Most freq sex:** F (52.86%)	**Label:** s55 **Support:** 100% **Density:** 65.25% **Type:** Hsp stay	**Origin:** Home (91.64%) **Destination:** Home (66.92%) **Most freq diag:** Pneumonia, organism unspecified **2nd most freq diag:** Pneumonia, unspecified **Most freq interv:** No intervention **2nd most freq interv:** Bronchial endoscopy	**Service:** Medicine (56.13%) **Specialist:** General practitioner (68.84%)
C94	**#:** 2292 **Avg length:** 1.59 **Death:** 99% **Hsp stay:** 84.74% **Consult:** 3.35% **Visit:** 11.90%	**Most freq age:** 85 and over (33.16%) **2nd freq age:** 80–84 (26.48%) **Most freq sex:** M (51.79%)	**Label:** s28 **Support:** 100% **Density:** 74.49% **Type:** Hsp stay	**Origin:** Home (74.49%) **Destination:** Home (54.08%) **Most freq diag:** Acute myocardial infarction **2nd most freq diag:** Coronary atherosclerosis **Most freq interv:** Coronary arteriography using two catheters **2nd most freq interv:** Other coronary arteriography	**Service:** Cardiology (65.06%) **Specialist:** Cardiology (54.86%)
C111	**#:** 1305 **Avg length:** 1.42 **Death:** 98.93% **Hsp stay:** 88.20% **Consult:** 3.18% **Visit:** 8.62%	**Most freq age:** 85 and over (27.13%) **2nd freq age:** 75–79 (25.29%) **Most freq sex:** M (53.41%)	**Label:** s52 **Support:** 100% **Density:** 74.89% **Type:** Hsp stay	**Origin:** Home (92.31%) **Destination:** Home (50.35%) **Most freq diag:** Malignant neoplasm, upper lobe, bronchus/lung **2nd most freq diag:** Malignant neoplasm, bronchus/lung **Most freq interv:** Transfusion, blood cells agglomerated **2nd most freq interv:** Blood transfusion	**Service:** General surgery (24.11%) **Specialist:** General practitioner (35.74%)
C112	**#:** 1334 **Avg length:** 1.42 **Death:** 99.33% **Hsp stay:** 86.61% **Consult:** 3.97% **Visit:** 9.42%	**Most freq age:** 85 and over (38.23%) **2nd freq age:** 80–84 (24.14%) **Most freq sex:** F (55.25%)	**Label:** s61 **Support:** 100% **Density:** 76.72% **Type:** Hsp stay	**Origin:** Home (87.60%) **Destination:** Home (59.59%) **Most freq diag:** Left heart failure **2nd most freq diag:** Congestive heart failure **Most freq interv:** Pacemaker Implantation **2nd most freq interv:** Blood transfusion	**Service:** Cardiology (49.94%) **Specialist:** Cardiology (40.56%)
C119	**#:** 674 **Avg length:** 1.34 **Death:** 97.63% **Hsp stay:** 85.70% **Consult:** 4.10% **Visit:** 10.20%	**Most freq age:** 85 and over (74.48%) **2nd freq age:** 80–84 (14.84%) **Most freq sex:** F (72.26%)	**Label:** s16 **Support:** 100% **Density:** 76.94% **Type:** Hsp stay	**Origin:** Home (83.36%) **Destination:** General and specialized hospital care or hospital center of psychiatric care (30.15%) **Most freq diag:** Transtrochanteric femur neck fracture, simple **2nd most freq diag:** Transcervical fracture femur neck, simple **Most freq interv:** Open fracture reduction + internal fixation, femur **2nd most freq interv:** Blood transfusion	**Service:** Orthopedics (80.20%) **Specialist:** Orthopedic surgery (87.66%)

Table 3
Discovered models for some clusters with a high presence of hospital stays in the pathways. **D:** Most Frequent Diagnosis; **I:** Most Frequent Intervention; **Sp:** Most Frequent Specialist; **Se:** Most Frequent Service.

Cluster	#	Model	Description of services present in model
C91	753 (0.42%)	start → s55 → end	**(s55) D:** Pneumonia, organism unspecified; **I:** No intervention; **Sp:** General practitioner; **Se:** Medicine
C94	2292 (1.27%)	start → s28 → end	**(s28) D:** Acute myocardial infarction; **I:** Coronary arteriography using two catheters; **Sp:** Cardiology; **Se:** Cardiology
C111	1305 (0.72%)	start → s52 → end	**(s52) D:** Malignant neoplasm, upper lobe, bronchus/lung; **I:** Transfusion, blood cells agglomerated; **Sp:** General practitioner; **Se:** General surgery
C112	1334 (0.74%)	start → s61 → end	**(s61) D:** Left heart failure; **I:** Pacemaker Implantation; **Sp:** Cardiology; **Se:** Cardiology
C119	674 (0.37%)	start → s16 → end	**(s16) D:** Transtrochanteric femur neck fracture, simple; **I:** Open fracture reduction + internal fixation, femur; **Sp:** Orthopedic surgery; **Se:** Orthopedics

Table 4
Description of clusters for the kidney problems pattern. **A:** Most Frequent Age; **2nd A:** 2nd Frequent Age; **S:** Most Frequent Sex; **D:** Most Frequent Diagnosis; **2nd D:** 2nd Frequent Diagnosis; **Sp:** Most Frequent Specialist; **Se:** Most Frequent Service; **Es:** Most Frequent Establishment.

Cluster	Description	Demographics	Services (density ≥ 10%)	Model	Description of services present in models
C1	**#:** 454 (0.25%) **Avg length:** 438.35 **Death:** 75.99% **Hsp stay:** 1.47% **Consult:** 7.12% **Visit:** 91.41%	**A:** 70–74 (28.41%) **2nd A:** 75–79 (28.19%) **S:** M (55.51%)	**v1:** 96.92–50.87% **v13:** 93.39–28.75%	Fig. 7(a)	**(v1) D:** not mentioned (100%); **Sp:** general practitioner (66.59%); **Es:** hospitals: outpatient (43.45%) **(v10) D:** Heart failure (16.31%), **2nd D:** Chronic ischemic heart disease (12.33%); **Sp:** general practitioner (67.95%); **Es:** hospitals: outpatient (51.97%) **(v13) D:** Renal failure (38.21%), **2nd D:** Chronic renal failure (31.61%), **Sp:** Nephrology (66.59%); **Es:** hospitals: outpatient (72.71%) **(v19) D:** General routine medical examination (33.69%), **2nd D:** Vaccination against influenza (11.09%); **Sp:** general practitioner (59.56%); **Es:** Private cabinet with the municipality number (46.41%)
C2	**#:** 1804 (1%) **Avg length:** 361.10 **Death:** 83.59% **Hsp stay:** 1.72% **Consult:** 7.56% **Visit:** 90.72%	**A:** 75–79 (26.94%) **2nd A:** 70–74 (26.77%) **S:** M (56.49%)	**v13:** 99.45–72.70%	Fig. 7(b)	
C3	**#:** 337 (0.19%) **Avg length:** 203.32 **Death:** 76.85% **Hsp stay:** 2.22% **Consult:** 6.93% **Visit:** 90.85%	**A:** 85 and over (36.5%) **2nd A:** 75–79 (20.77%) **S:** F (51.34%)	**v19:** 93.18–55.10% **v13:** 64.09–17.20%	Fig. 7(c)	
C17	**#:** 2747 (1.53%) **Avg length:** 101.13 **Death:** 68.58% **Hsp stay:** 4.25% **Consult:** 16.63% **Visit:** 68.58%	**A:** 85 and over (29.38%) **2nd A:** 80–84(23.7%) **S:** M (54.24%)	**v13:** 93.85–26.84% **v10:** 91.88–15.70%	Fig. 7(d)	

Table 5
Description of clusters for the heart problems pattern. **A:** Most Frequent Age; **2nd A:** 2nd Frequent Age; **S:** Most Frequent Sex; **D:** Most Frequent Diagnosis; **2nd D:** 2nd Frequent Diagnosis; **Sp:** Most Frequent Specialist; **Se:** Most Frequent Service; **Es:** Most Frequent Establishment.

Cluster	Description	Demographics	Services (density ≥ 10%)	Model	Description of services present in models
C20	**#:** 56141 (31.18%) **Avg length:** 98.71 **Death:** 55.57% **Hsp stay:** 4.05% **Consult:** 16.58% **Visit:** 79.37%	**A:** 85 and over (28.85%) **2nd A:** 80–84 (22.11%) **S:** F (54.21%)	**v10:** 99.39–39.23%	Fig. 8(a)	**(c10) D:** Heart failure (16.02%), **2nd D:** Chronic ischemic heart disease (12.96%); **Sp:** Cardiology (51.59%); **Es:** Emergency (38.21%) **(v1) D:** not mentioned (100%); **Sp:** general practitioner (66.59%); **Es:** hospitals: outpatient (43.45%) **(v10) D:** Heart failure (16.31%), **2nd D:** Chronic ischemic heart disease (12.33%); **Sp:** general practitioner (67.95%); **Es:** hospitals: outpatient (51.97%)
C23	**#:** 2685 (1.49%) **Avg length:** 92.26 **Death:** 62.61% **Hsp stay:** 4.32% **Consult:** 16.90% **Visit:** 78.78%	**A:** 85 and over (28.34%) **2nd A:** 80–84 (23.46%) **S:** M (52.33%)	**v10:** 96.95–23.40% **v1:** 93.82–28.62%	Fig. 8(b)	
C49	**#:** 1262 (0.70%) **Avg length:** 31.17 **Death:** 80.74% **Hsp stay:** 8.29% **Consult:** 35.43% **Visit:** 56.28%	**A:** 85 and over (41.36%) **2nd A:** 80–84 (24.64%) **S:** F (57.84%)	**v10:** 78.21–21.90% **c10:** 92.23–19.58%	Fig. 8(c)	

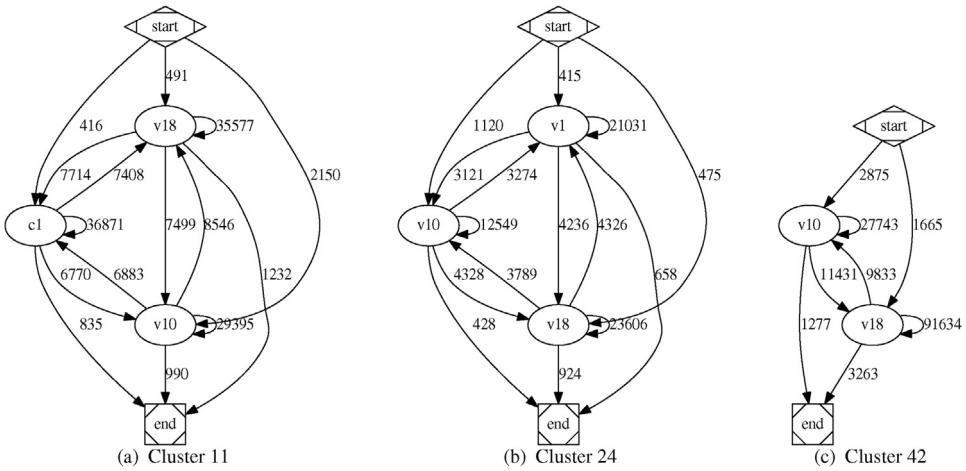


Fig. 9. Discovered models for the decompensated heart problem pattern.

Table 6
Description of clusters for the decompensated heart problem pattern. **A:** Most Frequent Age; **2nd A:** 2nd Frequent Age; **S:** Most Frequent Sex; **D:** Most Frequent Diagnosis; **2nd D:** 2nd Frequent Diagnosis; **Sp:** Most Frequent Specialist; **Se:** Most Frequent Service; **Es:** Most Frequent Establishment.

Cluster	Description	Demographics	Services (density ≥ 10%)	Model	Description of services present in models
C11	**#:** 3066 (1.70%) **Avg length:** 117.28 **Death:** 51.37% **Hsp stay:** 4% **Consult:** 26.30% **Visit:** 69.70%	**A:** 75–79 (23.84%) **2nd A:** 85 and over (23.29%) **S:** F (60.14%)	**v18:** 94.62–14.47% **c1:** 92.14–14.43% **v10:** 94.94–12.74%	Fig. 9(a)	**(c1) D:** not mentioned (100%); **Sp:** diagnostic Radiology (63.48%); **Es:** diagnostic radiology laboratories: general medical laboratory managed by a radiologist (47.46%) **(v1) D:** not mentioned (100%); **Sp:** general practitioner (66.59%); **Es:** hospitals: outpatient (43.45%) **(v10) D:** Heart failure (16.31%), **2nd D:** Chronic ischemic heart disease (12.33%); **Sp:** general practitioner (67.95%); **Es:** hospitals: outpatient (51.97%) **(v18) D:** Dyspnea and respiratory abnormalities (24.80%), **2nd D:** Chest pain (15.02%); **Sp:** general practitioner (80.02%); **Es:** emergency (46.76%)
C24	**#:** 2022 (1.12%) **Avg length:** 91.56 **Death:** 73.94% **Hsp stay:** 6.29% **Consult:** 19.68% **Visit:** 74.04%	**A:** 85 and over (33.78%) **2nd A:** 80–84 (23.99%) **S:** F (57.62%)	**v18:** 93.92–17.63% **v10:** 85.76–11.12% **v1:** 87.88–15.69%	Fig. 9(b)	
C42	**#:** 4585 (2.55%) **Avg length:** 66.78 **Death:** 82.46% **Hsp stay:** 5.70% **Consult:** 14.48% **Visit:** 79.82%	**A:** 85 and over (46.32%) **2nd A:** 80–84 (22.7%) **S:** F (57.3%)	**v18:** 96.25–34.20% **v10:** 81.44–13.21%	Fig. 9(c)	

visits and consultations compared to other groups (see Fig. 5(c) and (d)). These clusters are generally characterized by the consumption of one or two hospital stays related to a disease that is not caused by heart failure.

In further analysis, we examined the micro-level by using frequency analyses, computing service densities and discovering a model in each cluster. Through this analysis, we can dissect the clusters and obtain more details and knowledge that might be of interest to practitioners. As stated above, the first distinct pattern of medical treatment pathways is characterized by a high presence of hospital stays in the pathways. For example, cluster 94 is characterized by a hospital stay due to acute myocardial infarction, possibly severe, since it requires an intervention. In Fig. 6(a) we identified that cluster 94 has one frequent service given to patients in this cluster: hospital stay s28 (hospital stay for heart diseases (see Table 2)). Fig. 11(a) and Table 3 show the discovered model for cluster 94 which is based only on services that have a density higher than 10%. The patient that follows this pattern is simply involved in hospital stay s28. This hospital stay is present in 100% of patient treatment pathways and represents 74.49% of total medical services given in this cluster. These results suggest that there is a distinct pattern of patient treatment pathways which is characterized by a high presence of hospital stays in the pathways. Most of

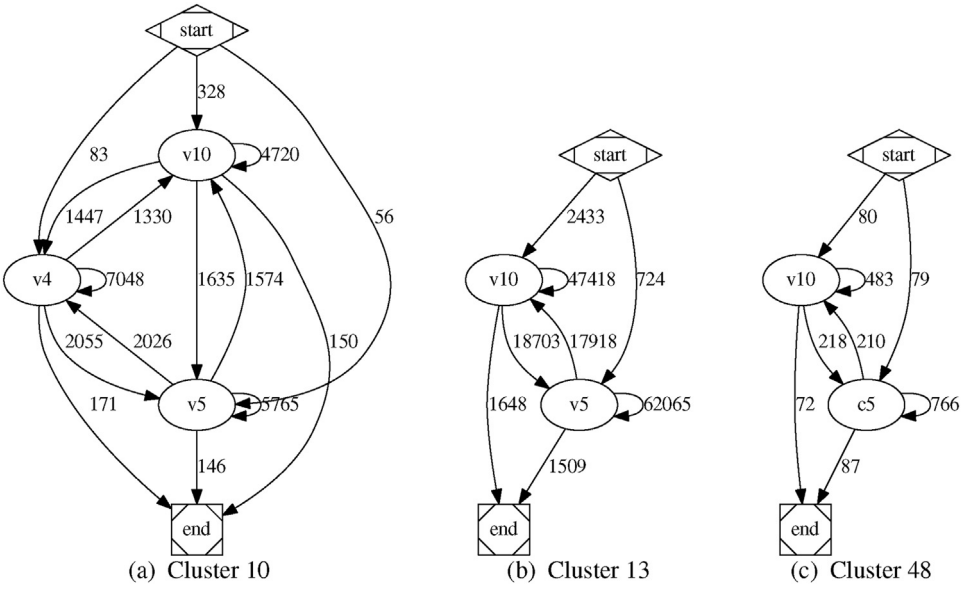


Fig. 10. Discovered models for the anemia problem pattern.

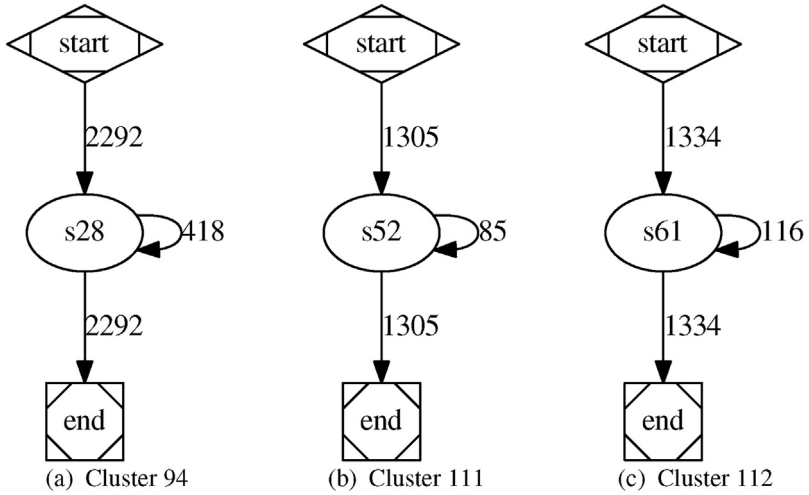
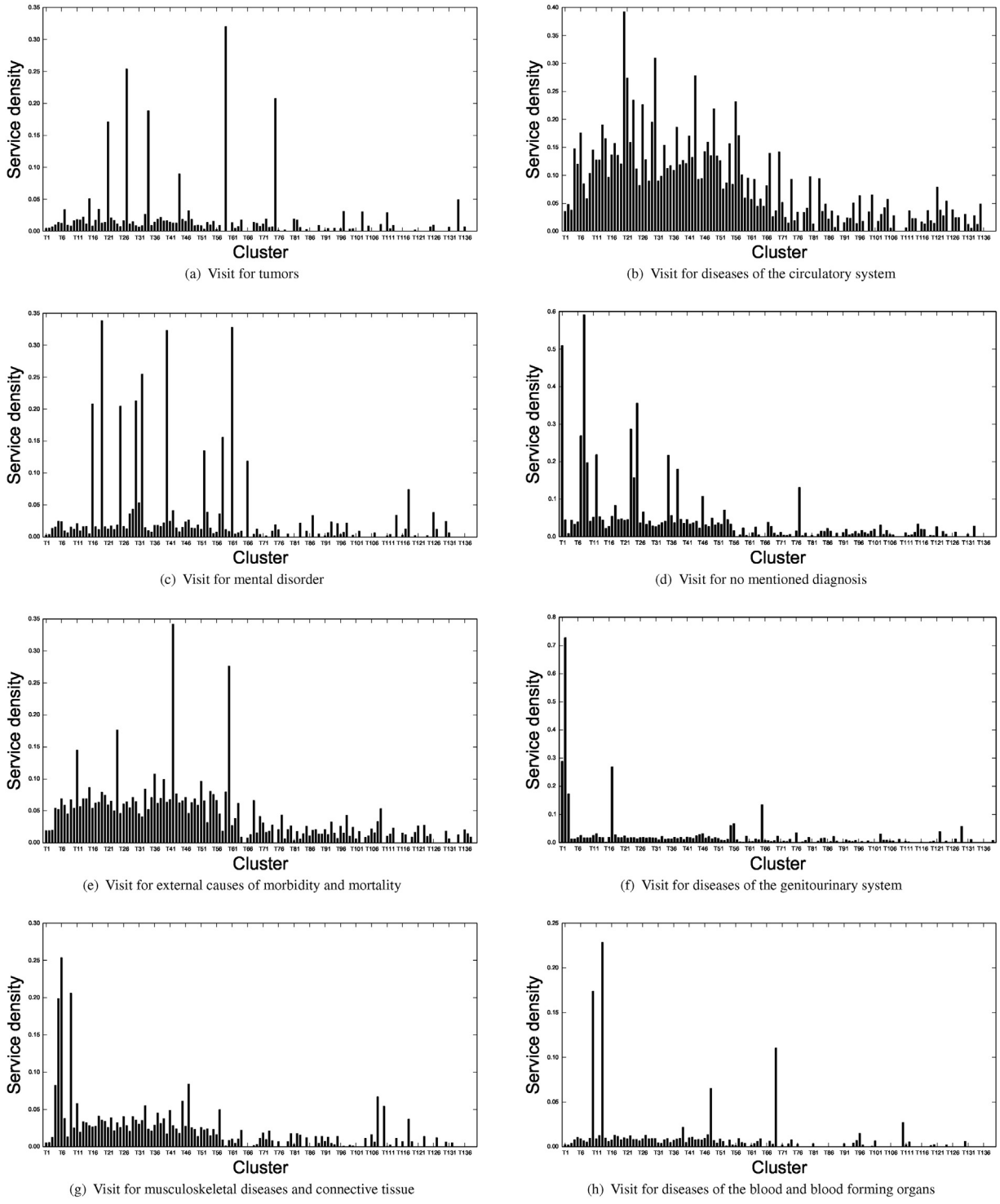


Fig. 11. Discovered models for some clusters with a high presence of hospital stays in the pathways.

these hospital stays are not caused by heart failure but by another disease. Tables 2 and 3 provide some examples for this pattern.

The next steps consisted in analyzing the contents of the other pathway clusters. In order to discover particular patterns, we examined the models produced and services that have a high density in the clusters. The results show that the proposed approach allows the identification and differentiation of patterns. Each of these patterns corresponds to a specific model and to specific treatment services that occur more frequently than other services in other clusters. These treatment services and models allow the description of patterns. One of these patterns was related to diseases of the genitourinary system associated with heart failure. This is presented in Fig. 12(f) that depicts the variability of the density of visits related to diseases of the genitourinary system according to the clusters. In cluster 1, we found a high density of visit v13, characterized by a renal failure as the most frequent diagnosis. With respect to hospitals: outpatient is noted as the most frequent site of contact and the nephrologist as the most frequently seen physician. We also noted a high density of visit v1 that contains no mentioned diagnosis as the most frequent diagnosis. Visit v13 is present in 93.92% of the pathways in the cluster and represents 28.75% of all services made in the pathways cluster. Visit v1 is present in 96.92% of the pathways in the cluster and represents 50.87% of all services made in the pathways cluster. This type of pattern describes regular monitoring of renal problems with a variety of other problems, probably of a less serious nature. Cluster 2 represents this pattern that is characterized by serious kidney problems and the requiring of several medical visits. Cluster 3 is similar to cluster 1 with a lower number of visits related to kidney problems. It is also different in that the v19 visit takes the place of v1. Visit v19 is a general routine medical examination. The last type of this pattern is described by cluster 17. This type is characterized by moderately high comorbidity between kidney failure and heart failure unlike the other types in this pattern. Fig. 7 and Table 4 provide an illustration of this pattern and its types. We can point out that cluster 17 has a relatively low mortality rate despite the fact that this type has 53.08% of the patients who start their pathways at an age higher than 80 years and 42.54% of the services which correspond to kidney or heart diseases. Fig. 6(b) shows the variability of service densities for this type of pattern.

In addition, we can observe the presence of a specific pattern for heart failure without significant comorbidity. We can see three trends in this pattern. The first one is described by cluster 20. It rep-



(a) Visit for tumors

(b) Visit for diseases of the circulatory system

(c) Visit for mental disorder

(d) Visit for no mentioned diagnosis

(e) Visit for external causes of morbidity and mortality

(f) Visit for diseases of the genitourinary system

(g) Visit for musculoskeletal diseases and connective tissue

(h) Visit for diseases of the blood and blood forming organs

Fig. 12. Service densities according to clusters.

resents the pathways for heart failure with regular monitoring and is characterized by a relatively low mortality rate (55.57%) while 50.96% of the patients in this group are over 80 years old. The second one is given by cluster 23. This trend is also characterized by regular monitoring but differing by visit v1. This visit is a visit for radiology or ultrasonography. In this trend, there is an important presence of this visit and a significant correlation with the visit for heart problems v10. This group has almost the same age characteristics as the previous group (51.80% over 80 years). However, it has a slightly higher rate of mortality despite the fact that the density of v10 is less than that of cluster 20 (23.40% vs. 39.23%). The last type is represented by cluster 49. This group has a high mortality rate (80.74%) that can be explained in part by the age (66% over 80 years) and can also be explained by more complicated medical cases since there is a correlation between consultations due to heart failure and visits due to the same problems. Fig. 8 and Table 5 depict a view of this pattern and its types.

Furthermore, we can note that there is another pattern that has emerged from the analyses. This pattern represents patients with decompensated heart failure causing respiratory problems.

Table 7
Description of clusters for the anemia problem pattern. **A:** Most Frequent Age; **2nd A:** 2nd Frequent Age; **S:** Most Frequent Sex; **D:** Most Frequent Diagnosis; **2nd D:** 2nd Frequent Diagnosis; **Sp:** Most Frequent Specialist; **Se:** Most Frequent Service; **Es:** Most Frequent Establishment.

Cluster	Description	Demographics	Services (density ≥ 10%)	Model	Description of services present in models
C10	**#:** 467 (0.26%) **Avg length:** 117.35 **Death:** 58.46% **Hsp stay:** 4.09% **Consult:** 14.62% **Visit:** 81.29%	**A:** 75–79 (29.34%) **2nd A:** 70–74 (22.48%) **S:** F (56.32%)	**v4:** 92.93–19.35% **v5:** 95.50–17.36% **v10:** 96.15–14.51%	Fig. 10(a)	**(c5) D:** Anemia (70.18%), **2nd D:** coagulation abnormalities (7.96%); **Sp:** Gastroenterology (32.35%); **Es:** hospitals: outpatient (48.89%) **(v4) D:** Diabetes mellitus uncomplicated (74.04%), **2nd D:** Hypothyroidism (5%); **Sp:** general practitioner (82.56%); **Es:** Private cabinet with the municipality number (65.05%) **(v5) D:** Anemia (59.03%), **2nd D:** coagulation abnormalities (13.94%); **Sp:** general practitioner (60.39%); **Es:** Private cabinet with the municipality number (43.75%) **(v10) D:** Heart failure (16.31%), **2nd D:** Chronic ischemic heart disease (12.33%); **Sp:** general practitioner (67.95%); **Es:** hospitals: outpatient (51.97%)
C13	**#:** 3161 (1.76%) **Avg length:** 112.92 **Death:** 65.74% **Hsp stay:** 3.74% **Consult:** 15.79% **Visit:** 80.47%	**A:** 85 and over (31.38%) **2nd A:** 80–84 (23.76%) **S:** F (53.08%)	**v5:** 96.24–22.83% **v10:** 95.63–18.99%	Fig. 10(b)	
C48	**#:** 169 (0.09%) **Avg length:** 33.86 **Death:** 89.35% **Hsp stay:** 8.69% **Consult:** 39.02% **Visit:** 52.29%	**A:** 85 and over (44.97%) **2nd A:** 80–84 (23.08%) **S:** M (53.25%)	**c5:** 85.80–18.58% **v10:** 71.01–13.51%	Fig. 10(c)	

This pattern is composed of three trends: cluster 11, cluster 24 and cluster 42. Cluster 11 represents a typical pattern. Clusters 24 and 42 are trends with higher comorbidity with respiratory problems. They are characterized by an acute respiratory problem captured by visit v18. Fig. 9 and Table 6 provide the details on these trends and their differences.

Results also show that cluster 10, 13 and 48 form a specific pattern for patients who have as a comorbidity anemia. This pattern is significant because these patients could require special monitoring and specification of this pattern. Fig. 10 and Table 7 provide information about this pattern and its types.

Clusters 16, 19, 25, 30, 32, 40, and 52 are a specific pattern with a high presence of visits related to anxiety. Fig. 12(c) depicts the variability of the density of these visits according to clusters. In these groups more than two thirds of the patients are older than 80 years of age. Furthermore, we can identify a pattern for patients who have a tumor associated with heart failure represented by clusters 21, 27 and 34. (see Fig. 12(a)).

6.2. Discussion

The relevance of our work can be highlighted by several aspects. First, our work allows the use of administrative databases that contain information embedded in a large volume of data for the building and extracting of latent patterns of patient pathways in the health system. This capacity to deal with many services and pathways is a strong feature of our approach compared to the current literature on trace clustering [20]. It can be seen as a first layer that can assist healthcare system administrators to extract the knowledge included in the common health care databases. To the best of our knowledge, this is the first study using the HMM for pathway clustering and using administrative databases with a large variety of services and a broad range of time. Our contribution for healthcare administrators can be seen in the detection of special patterns representing low mortality rates. It might lead research to undertake further studies of practices in these groups. This might help to improve the practice guidelines.

Another facet of our work is the identification of patients who require monitoring after a number of visits or other selected criteria. This can be done by detecting patients who follow a specific pattern and placing alert mechanisms in the information system. Given our focus on administrative databases that contain raw data and complex events, the proposal made in this paper allows this type of database to be explored to provide a better understanding of the medical processes at work, and eventually improve the healthcare system.

Our method was applied directly over a large and complex healthcare database, for the extraction of a relatively large number of clusters. It should be stressed that heart disease in elderly people is a complex case, with a large set of patients having many comorbidities. We chose this context because it contains significant differences and a lot of variety such that a large number of clusters is required to detect this heterogeneity. In practice, a more dynamic methodology can be considered, by first making a coarse processing where a few dozen pathway clusters are extracted, then analyzing the clusters and subsequently re-applying the method on some data subsets to extract more specific elements. That would make the analysis much easier for the healthcare specialists, with a reduced number of pathway clusters to handle, while allowing the extraction of more detailed knowledge on elements that are specifically of interest in the current context.

7. Conclusion

In this paper we have presented an approach for designing, building, and clustering treatment patient pathways extracted from administrative databases. In general, these processes are complex. However our proposed approach allows the differentiation between some patterns present in these databases even if patients have the same chronic disease. Our work will allow specialists to have a summary of the information contained in these databases and identify specific patterns for further studies.

Acknowledgements

The project was supported by funding from CIHR-Canda and NSERC-Canada. Computation resources have been provided by Calcul Québec/Compute Canada. We are grateful to Annette Schwerdtfeger for conducting an extensive proofreading of the paper.

References

[1] Bose RJC, van der Aalst WM. Trace clustering based on conserved patterns: towards achieving better process models. In: Business process management workshops, vol. 43. 2009. p. 170–81.
[2] Bose RJC, Van der Aalst WM. Context aware trace clustering: towards improving process mining results. In: Proceedings of the 2009 SIAM international conference on data mining. 2009. p. 401–12.
[3] Coiera E. Guide to health informatics. CRC Press; 2015.
[4] De Weerdt J, vanden Broucke S, Vanthienen J, Baesens B. Active trace clustering for improved process discovery. IEEE Trans Knowl Data Eng 2013;25(12):2708–20.
[5] Elghazel H, Deslandres V, Kallel K, Dussauchoy A. Clinical pathway analysis using graph-based approach and Markov models. 2nd international conference on digital information management, 2007 (ICDIM'07), vol. 1 2007:279–84.
[6] Ferreira D, Zacarias M, Malheiros M, Ferreira P. Approaching process mining with sequence clustering: experiments and findings. In: Business process management. Springer; 2007. p. 360–74.
[7] Ha Q-T, Bui H-N, Nguyen T-T. A trace clustering solution based on using the distance graph model. In: International conference on computational collective intelligence. 2016. p. 313–22.
[8] Huang Z, Dong W, Ji L, Gan C, Lu X, Duan H. Discovery of clinical pathway patterns from event logs using probabilistic topic models. J Biomed Inform 2014;47:39–57.
[9] Huang Z, Lu X, Duan H. Latent treatment pattern discovery for clinical processes. J Med Syst 2013;37(2):1–10.
[10] Kent D, Kitsios G. Against pragmatism: on efficacy, effectiveness and the real world. Trials 2009;10(1):48.
[11] Knab B, Schliep A, Steckemetz B, Wichern B. Model-based clustering with hidden Markov models and its application to financial time-series data. In: Between data science and applied data analysis. 2003. p. 561–9.
[12] Lenz R, Reichert M. It support for healthcare processes-premises, challenges, perspectives. Data Knowl Eng 2007;61(1):39–58.
[13] Lloyd-Jones D, Adams RJ, Brown TM, Carnethon M, Dai S, De Simone G, et al. Heart disease and stroke statistics – 2010 update. Circulation 2010;121(12):948–54. On Behalf of the American Heart Association Statistics Committee and Stroke Statistics Subcommittee.
[14] Man JP, Jugdutt BI. Systolic heart failure in the elderly: optimizing medical management. Heart Fail Rev 2012;17(4–5):563–71.
[15] Ministère de la Santé et des Services sociaux du Québec. Statistiques; 2013 http://wpp01.msss.gouv.qc.ca/appl/g74web/statistiques.asp.
[16] Najjar A, Gagné C, Reinharz D. A novel mixed values *k*-prototypes algorithm with application to health care databases mining. In: IEEE symposium series on computational intelligence (IEEE-SSCI). 2014. p. 159–66.
[17] Nguyen P, Slominski A, Muthusamy V, Ishakian V, Nahrstedt K. Process trace clustering: a heterogeneous information network approach. In: Proceedings of the 2016 SIAM international conference on data mining. 2016. p. 279–87.
[18] Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 1989;77(2):257–86.
[19] Rebuge Á, Ferreira DR. Business process analysis in healthcare environments: a methodology based on process mining. Inf Syst 2012;37(2):99–116.
[20] Thaler T, Ternis SF, Fettke P, Loos P. A comparative analysis of process instance cluster techniques. Wirtschaftsinformatik (WI 2015) 2015:423–37.
[21] Van Der Aalst WM, Ter Hofstede AH, Weske M. Business process management: a survey. In: Business process management. 2003. p. 1–12.
[22] Weske M, van der Aalst WM, Verbeek H. Advances in business process management. Data Knowl Eng 2004;50(1):1–8.