Journal of Biomedical Informatics 47 (2014) 39–57

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Discovery of clinical pathway patterns from event logs using probabilistic topic models

Zhengxing Huang [a,1], Wei Dong [b,1], Lei Ji [c], Chenxi Gan [a], Xudong Lu [a], Huilong Duan [a,*]

[a] College of Biomedical Engineering and Instrument Science, Zhejiang University, Zhou Yiqing Building 510, Zheda Road 38#, Hangzhou, Zhejiang 310008, China
[b] Department of Cardiology, Chinese PLA General Hospital, China
[c] IT Department, Chinese PLA General Hospital, China

ARTICLE INFO

Article history:
Received 14 February 2013
Accepted 7 September 2013
Available online 25 September 2013

Keywords:
Clinical pathway analysis
Topic models
Latent Dirichlet Allocation
Pattern discovery
Clinical event log

ABSTRACT

Discovery of clinical pathway (CP) patterns has experienced increased attention over the years due to its importance for revealing the structure, semantics and dynamics of CPs, and to its usefulness for providing clinicians with explicit knowledge which can be directly used to guide treatment activities of individual patients. Generally, discovery of CP patterns is a challenging task as treatment behaviors in CPs often have a large variability depending on factors such as time, location and patient individual. Based on the assumption that CP patterns can be derived from clinical event logs which usually record various treatment activities in CP executions, this study proposes a novel approach to CP pattern discovery by modeling CPs using mixtures of an extension to the Latent Dirichlet Allocation family that jointly models various treatment activities and their occurring time stamps in CPs. Clinical case studies are performed to evaluate the proposed approach via real-world data sets recording typical treatment behaviors in patient careflow. The obtained results demonstrate the suitability of the proposed approach for CP pattern discovery, and indicate the promise in research efforts related to CP analysis and optimization.

1. Introduction

In order to increase the quality of care services in an unfavorable economic scenario and under the financial pressure by governments, health-care organizations have to introduce clearly defined clinical pathways (CPs) for patients, and these pathways must be improved continuously [1–5]. A CP is a defined set of therapy and treatment activities that represent the steps required to achieve a specific treatment objective in patient careflow. It has been proven that CPs can break functional boundaries and offer an explicit process-oriented view of health-care where the efficient collaboration and coordination of physicians become the crucial issue [6,2,7–13]. Since actual treatment activities are extremely complex, with numerous variations across various stages in CPs, they often bear no relation to the ideal as envisaged by the designers of CPs. To this end, CP analysis is nonetheless vital for healthcare management due to its usefulness for capturing the actionable knowledge to administrate, automate, and schedule the best practice for individual patients in the executions of CPs [14–19].

Traditionally, health-care practitioners analyze CPs by looking at aggregated data seen from an external perspective, e.g., length of stay (LOS), charge, bed utilization, medical service levels, and so on [20]. A new and promising way of acquiring insights into CPs is to analyze clinical event logs using process mining techniques [21–24], which have been widely studied in the domain of business process management, which attempt to extract non-trivial and useful information from event logs [24,16]. Regarding clinical settings, many hospital information systems usually record various kinds of treatment events in patient careflow, and each event refers to a specific case (i.e., a patient trace following a specific CP), is related to a particular clinical activity type (i.e., a well-defined step in a specific CP), has an associated occurring time stamp, and other properties. Clinical event logs conceal an untapped reservoir of knowledge about the way of specific therapy and treatment activities being performed on particular patients in their careflow. It is, therefore, possible to mine clinical event logs, extract non-trivial knowledge from these logs, and exploit these for further analysis.

In this study, we focus on discovering CP patterns from clinical event logs. We define CP patterns to be temporal regularities of CPs. A CP often involves several underlying patterns of treatment activities from admission to discharge, possibly over different time scales and for varying time intervals. The underlying patterns form the backbone of CPs and should be conserved, and the absence of presence of such patterns may indicate the cause of an anomaly or a malfunction (e.g., medical error) [25]. Therefore, we argue that discovery of CP patterns offers potentially rich information about patient treatment intent and behaviors, and can provide a basis for further CP analysis.

* Corresponding author. Fax: +86 571 87951960.
E-mail address: duanhl@zju.edu.cn (H. Duan).
[1] These authors contributed equally to this work.

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.09.003

However, discovery of CP patterns is in general a challenging task as the diversity and complexity of treatment behaviors in CPs are far higher than that of common business processes [16]. Although it is possible to employ existing process mining techniques in mining CP patterns from clinical event logs, many of them often generate spaghetti-like patterns which are difficult to be comprehended by clinical experts and either not amenable or lack of assistance to efforts of CP analysis and improvement. In our previous work [19], we developed a new process mining algorithm to derive a concise summary from a clinical event log. While the constructed summary is capable of describing the entire structure of a CP, it is challenged by describing underlying treatment patterns and handling the different levels of complexity of these patterns. Note that what makes the discovery of CP patterns complex is that they are typically composed of various and heterogeneous treatment activities and the composition of activities has a large variability depending on factors such as time, location and patient individual.

To this end, we propose to leverage the power of probabilistic topic models (1) to extract latent CP patterns from event logs, and (2) to enable the recognition of patient traces as a composition of such patterns. In our previous work [18], we adapted a probabilistic topic model, Latent Dirichlet Allocation (LDA) to build a discovery model of latent CP patterns. LDA is an unsupervised probabilistic clustering technique used to discover latent topics from bags of words in text by finding co-occurrences of words in documents. Here, treatment activities are extracted and are mapped to words. These are then collated for a specific patient trace and become analogous to a document. The latent topics discovered by LDA in this way are interpreted as CP patterns. The probability distribution derived from LDA surmises the essential features of CP patterns, and CPs can be accurately described by combining different classes of distributions [18].

While interpretable and meaningful CP patterns can be discovered, our previous work has a major shortcoming that it does not take the occurring time stamps of treatment activities into account. Note that a CP pattern captures not only the low-dimensional structure of treatment activities, but also when these activities are performed in patient careflow. Standard LDA does not distinguish differences in occurring time of treatment activities in CPs. While the occurring time stamp is often critical to capturing the meaning of treatment behaviors in CP analysis, neglecting it may result in spurious associations and misleading inference on capturing the meaning of underlying CP patterns from clinical event logs. Although it is possible to take pairs of "treatment activity–time stamp" as words in topic models, it will not be able to discriminate the generation of treatment activities and that of occurrences.

In this work, we exploit to extend our previous work using LDA in [18] to propose an extension of LDA, i.e., clinical pathway model (CPM), show that richer CP patterns can be automatically discovered from clinical event logs. The proposed model finds latent CP patterns that are influenced by both treatment activities and their occurring time stamps in an unsupervised manner, thus disclosing temporal structure of discovered patterns. The ability to infer latent CP patterns is a vital and foundational component for CP analysis. Our investigations enable both richer representation and more accurate extractions of different aspects of treatment behaviors in CPs, and hence the outcomes of out study can be potentially valuable to CP analysis and redesign.

The remainder of the paper is organized as follows. Section 2 presents preliminary knowledge of the proposed approach. Section 3 describes the proposed CPM for discovering underlying CP patterns from clinical event logs. Section 4 carefully presents our experimental results. Section 5 discusses the results obtained. Related work is outlined in Section 6. Finally, some conclusions are given in Section 7.

2. Preliminaries

2.1. Representation of a patient trace

The objective of this study is to discover latent CP patterns from clinical event logs. In particular, the proposed approach assumes that it is possible to record clinical events sequentially such that each event refers to a treatment activity (i.e., a well-defined step in a CP). Furthermore, additional information such that the occurring time stamp of the event is used in this study. To explain the kind of input needed for the proposed approach, we first define the concept of a clinical event.

Definition 1. (Clinical event)[2]. Let A be a finite set of event identifiers (clinical terms describing treatment activities), and T the time domain (set of time stamp primitives). A **clinical event** e is a pair $e = (a,t)$ where $a \in A$ and $t \in T$. Formally, we use $e{\cdot}a$ and $e{\cdot}t$ to denote the activity type, and the occurring time stamp of a clinical event, respectively. We denote by $E \subseteq A \times T$ the set of all valid clinical events of a particular domain.

Note that clinical events could be characterized by various properties, e.g., an event has an occurring time stamp, it corresponds to an activity type, and has associated costs, etc. We do not impose a specific set of properties, however, given the focus of this paper, we assume that the activity type and occurring time stamp of the event are present. Thus, unfolding in time stamp and activity type, a clinical event e is a treatment activity–time stamp pair $\langle$activity,time stamp$\rangle$ that the activity has been occurred at a specific time point. For example, let '(Admission, day 1)' is a clinical event indicating that a patient was in admission on the first day in his or her length of stay; likewise '(PCI, Day 4)' means the patient was performed a percutaneous coronary intervention on the 4th day in his/her length of stay. For simplicity, the time stamps of these event examples are integer values, however it could be presented in a date-format time stamp.

Definition 2. (Patient trace). Let E be the domain of clinical events (i.e., the number of unique clinical events in the collection). A **patient trace** is a non-empty sequence of clinical events performed on a particular patient, i.e., $\sigma = \langle e_1, e_2, \ldots, e_n \rangle$, where $e_i \in E$ $(1 \leqslant i \leqslant n)$ is a particular clinical event.

In general, a patient trace consists of different categories of clinical events. For a particular patient trace, certain temporal constraints exist between the events. Table 1 depicts an example of eight patient traces. The corresponding meanings of treatment activities listed in Table 1 can be found in Table 4. We will continue to use this example in the rest of the paper to illustrate various points.

From the practical point of view, we make the assumption that treatment behaviors in the executions of CPs are correctly recorded in a clinical event log. Information in the log is expected to (1) refer to clinical events of specific patient traces, and (2) the activity type and occurring time stamp of clinical events. The formal definition of a clinical event log is as follows:

Definition 3. (Clinical event log). Let $\mathcal{C}$ be the set of all possible patient traces. A **clinical event log** $\mathcal{L}$ is a set of patient traces, i.e., $\mathcal{L} \subseteq \mathcal{C}$.

[2] Some clinical events might have a duration, i.e., they are conducted not at a specific time stamp, but over a time period. However, such a clinical event can be assumed to consist of a pair of sub clinical activities, i.e., a start event and an end event, which correspond to a start event and an end event, respectively. In this study, we assume that clinical events are time point events, and intervals are represented by starting and ending time point events.

Table 1
An example clinical event log for the unstable angina clinical pathway. The traces in the log are simplified information extraction from patient records of the Chinese PLA General hospital.

σ_1	$\langle(A_0, 1), (A_1, 1), (A_{42}, 1), (A_9, 1), (A_7, 1), (A_5, 1), (A_{18}, 1), (A_{47}, 1), (A_{36}, 1), (A_{37}, 1), (A_5, 2), (A_7, 2), (A_8, 2), (A_9, 2), (A_{38}, 3), (A_{64}, 3), (A_{41}, 3), (A_9, 3), (A_7, 3), (A_7, 3), (A_5, 3), (A_{65}, 3), (A_9, 4), (A_5, 4), (A_7, 4), (A_9, 5), (A_{40}, 5), (A_{11}, 5), (A_5, 5), (A_7, 5), (A_9, 6), (A_5, 6), (A_7, 6), (A_9, 7), (A_5, 7), (A_7, 7), (A_{11}, 7), (A_{42}, 7), (A_9, 8), (A_5, 8), (A_7, 8), (A_9, 9), (A_5, 9), (A_7, 9), (A_2, 9)\rangle$
σ_2	$\langle(A_0, 1), (A_1, 1), (A_{42}, 1), (A_5, 1), (A_{13}, 1), (A_{22}, 1), (A_9, 1), (A_{37}, 1), (A_{47}, 1), (A_{15}, 1), (A_{32}, 1), (A_{46}, 1), (A_{34}, 1), (A_{31}, 1), (A_7, 1), (A_{66}, 1), (A_{18}, 1), (A_8, 1), (A_5, 2), (A_6, 2), (A_7, 2), (A_9, 2), (A_5, 3), (A_6, 3), (A_7, 3), (A_9, 3), (A_{40}, 4), (A_9, 4), (A_{53}, 4), (A_5, 4), (A_7, 4), (A_{13}, 4), (A_6, 4), (A_5, 5), (A_6, 5), (A_9, 5), (A_7, 5), (A_{13}, 5), (A_{18}, 5), (A_9, 6), (A_{53}, 6), (A_5, 6), (A_{13}, 6), (A_7, 6), (A_6, 6), (A_9, 7), (A_{16}, 7), (A_{13}, 7), (A_7, 7), (A_5, 7), (A_6, 7), (A_9, 8), (A_{16}, 8), (A_{13}, 8), (A_7, 8), (A_5, 8), (A_6, 8), (A_9, 9), (A_5, 9), (A_{16}, 9), (A_7, 9), (A_{13}, 9), (A_6, 9), (A_{38}, 9), (A_8, 9), (A_{41}, 9), (A_9, 10), (A_{16}, 10), (A_{13}, 10), (A_7, 10), (A_5, 10), (A_6, 10), (A_{57}, 10), (A_9, 11), (A_{16}, 11), (A_{13}, 11), (A_7, 11), (A_5, 11), (A_6, 11), (A_9, 12), (A_{16}, 12), (A_{13}, 12), (A_7, 12), (A_5, 12), (A_2, 12)\rangle$
σ_3	$\langle(A_0, 1), (A_1, 1), (A_{40}, 1), (A_{10}, 1), (A_7, 1), (A_{36}, 1), (A_{31}, 1), (A_{37}, 1), (A_{32}, 1), (A_{23}, 1), (A_9, 1), (A_{13}, 1), (A_5, 1), (A_{12}, 1), (A_{42}, 1), (A_{10}, 2), (A_7, 2), (A_9, 2), (A_5, 2), (A_{13}, 2), (A_{12}, 2), (A_7, 3), (A_{10}, 3), (A_{13}, 3), (A_9, 3), (A_5, 3), (A_{12}, 3), (A_7, 4), (A_{10}, 4), (A_{13}, 4), (A_9, 4), (A_5, 4), (A_{12}, 4), (A_7, 5), (A_{10}, 5), (A_9, 5), (A_{13}, 5), (A_5, 5), (A_{12}, 5), (A_7, 6), (A_{10}, 6), (A_9, 6), (A_{13}, 6), (A_5, 6), (A_{12}, 6), (A_{10}, 7), (A_7, 7), (A_9, 7), (A_{13}, 7), (A_5, 7), (A_{12}, 7), (A_7, 8), (A_{10}, 8), (A_{13}, 8), (A_9, 8), (A_{11}, 8), (A_5, 8), (A_{12}, 8), (A_{38}, 8), (A_7, 9), (A_{10}, 9), (A_{13}, 9), (A_9, 9), (A_5, 9), (A_{12}, 9), (A_7, 10), (A_{10}, 10), (A_9, 10), (A_{13}, 10), (A_{12}, 10), (A_5, 10), (A_{10}, 11), (A_7, 11), (A_9, 11), (A_{13}, 11), (A_5, 11), (A_{12}, 11), (A_7, 12), (A_{10}, 12), (A_9, 12), (A_{13}, 12), (A_5, 12), (A_{12}, 12), (A_7, 13), (A_{10}, 13), (A_9, 13), (A_{13}, 13), (A_5, 13), (A_{12}, 13), (A_7, 14), (A_{10}, 14), (A_9, 14), (A_{13}, 14), (A_5, 14), (A_{12}, 14), (A_2, 14)\rangle$
σ_4	$\langle(A_0, 1), (A_1, 1), (A_{10}, 1), (A_5, 1), (A_9, 1), (A_{32}, 1), (A_{46}, 1), (A_{22}, 1), (A_{34}, 1), (A_{31}, 1), (A_{37}, 1), (A_{36}, 1), (A_9, 1), (A_7, 1), (A_{18}, 1), (A_{57}, 1), (A_5, 2), (A_9, 2), (A_7, 2), (A_8, 2), (A_{57}, 2), (A_5, 3), (A_{64}, 3), (A_{38}, 3), (A_8, 3), (A_9, 3), (A_7, 3), (A_{57}, 3), (A_9, 4), (A_5, 4), (A_7, 4), (A_{57}, 4), (A_9, 5), (A_5, 5), (A_7, 5), (A_{57}, 5), (A_2, 5)\rangle$
σ_5	$\langle(A_0, 1), (A_1, 1), (A_5, 1), (A_{18}, 1), (A_{13}, 1), (A_9, 1), (A_{10}, 1), (A_7, 1), (A_8, 1), (A_{46}, 2), (A_{32}, 2), (A_{34}, 2), (A_{65}, 2), (A_{37}, 2), (A_{22}, 2), (A_{31}, 2), (A_{36}, 2), (A_{28}, 2), (A_{64}, 2), (A_{38}, 2), (A_9, 2), (A_{11}, 2), (A_5, 2), (A_{10}, 2), (A_7, 2), A_{54}, 3), (A_9, 3), (A_{10}, 3), (A_5, 3), (A_7, 3), (A_{53}, 4), (A_{31}, 4), (A_9, 4), (A_7, 4), (A_{10}, 4), (A_5, 4), (A_9, 5), (A_{57}, 5), (A_1, 5), (A_5, 5), (A_{10}, 5), (A_7, 5), (A_9, 6), (A_{10}, 6), (A_5, 6), (A_7, 6), (A_{31}, 7), (A_9, 7), (A_5, 7), (A_{10}, 7), (A_7, 7), (A_2, 7)\rangle$
σ_6	$\langle(A_0, 1), (A_{41}, 1), (A_{28}, 1), (A_{42}, 1), (A_5, 1), (A_7, 1), (A_8, 1), (A_9, 2), (A_7, 2), (A_5, 2), (A_{67}, 2), A_9, 3), (A_5, 3), (A_{67}, 3), (A_7, 3), (A_8, 3), (A_9, 4), (A_{47}, 4), (A_{37}, 4), (A_{65}, 4), (A_{36}, 4), (A_7, 4), (A_5, 4), (A_9, 5), (A_5, 5), (A_7, 5), (A_1, 5), (A_9, 6), (A_5, 6), (A_7, 6), (A_9, 7), (A_5, 7), (A_7, 7), (A_2, 7)\rangle$
σ_7	$\langle(A_0, 1), (A_1, 1), (A_{42}, 1), (A_9, 1), (A_7, 1), (A_5, 1), (A_{11}, 1), (A_{32}, 1), (A_{46}, 1), (A_{22}, 1), (A_{31}, 1), (A_{34}, 1), (A_{22}, 1), (A_8, 1), (A_{38}, 2), (A_{41}, 2), (A_{13}, 2), (A_5, 2), (A_8, 2), (A_{11}, 2), (A_9, 2), (A_7, 2), A_9, 3), (A_{11}, 3), (A_7, 3), (A_5, 3), (A_9, 4), (A_5, 4), (A_7, 4), (A_9, 5), (A_7, 5), (A_5, 5), (A_9, 6), (A_{11}, 6), (A_5, 6), (A_7, 6), (A_2, 7)\rangle$
σ_8	$\langle(A_0, 1), (A_{28}, 1), (A_5, 1), (A_{42}, 1), (A_9, 1), (A_{46}, 1), (A_7, 1), (A_{31}, 1), (A_{34}, 1), (A_{32}, 1), (A_8, 1), (A_{18}, 1), (A_{13}, 2), (A_9, 2), (A_5, 2), (A_7, 2), (A_{38}, 2), (A_8, 2), (A_{41}, 2), A_8, 3), (A_9, 3), (A_5, 3), (A_9, 4), (A_5, 4), (A_{15}, 4), (A_2, 5)\rangle$

Table 1 shows an example of a clinical event log about the unstable angina CP, which consists of eight patient traces. Note that we assume that clinical events in CPs are regularly recorded in an event log by various kinds of hospital information systems, e.g., electronic medical record system (EMRs), radiology information system, picture archiving and communication system, laboratory information system, etc., which effectively reflects the real executing conditions in CPs.

2.2. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a powerful tool, introduced by [26] and initially developed to characterize text documents, in which each document is modeled as a multinomial distribution of topics, and each topic is modeled as a multinomial distribution of words. LDA has been extended to other collections of discrete data. They are probabilistic generative models that can be used to explain multinomial observations by unsupervised learning.

In our previous work, we employed LDA to discover latent CP patterns from clinical event logs, viewing this task as an inference problem over the latent CP pattern variables [18]. LDA allows to infer the inherent CP patterns from a clinical event log. For a particular patient trace σ, it picks a set of CP patterns with different emphasis. Thus, we model the mixture of CP patterns as multinomial probability distribution $P(z|\sigma)$ over CP pattern z. Similarly, the importance of each clinical activity a for each CP pattern z is also modeled as a multinomial probability distribution $P(a|z)$ over activities a of an activity domain. Given these two distributions, we can compute the probability of a clinical activity a occurring in patient trace σ:

$$P(a|\sigma) = \sum_{z=1}^{K} P(a|z)P(z|\sigma) \tag{1}$$

assuming that there are K CP patterns the log contains. Having many patient traces in the log, we observe a matrix of observed $P(a|\sigma)$ as depicted on the left-hand side of the equation in Fig. 1. This resulting clinical activity-patient trace matrix is assumed to be the product out of two matrices, one of the activity probabilities within CP patterns and another of the CP pattern probabilities within patient traces on the right-hand side, thereby recovering the characteristic clinical activities for each CP pattern and the mixture of CP patterns for each patient trace.

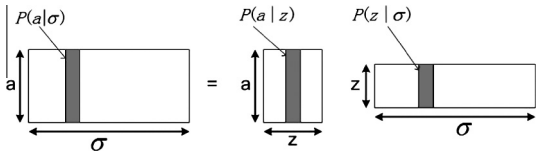


Fig. 1. Matrix representation. By introducing an unobserved, latent CP pattern z, the observed matrix of P $(a|\sigma)$ is decomposed into a CP pattern-treatment activity matrix of P $(a|z)$ and a patient trace-CP pattern matrix of P $(z|\sigma)$.

Using the LDA model shown in Fig. 2A, each patient trace in the log is modeled as a finite mixture over an underlying set of K CP patterns. The CP pattern mixture is drawn from a Dirichlet prior to the entire log. The generative process begins by choosing a distribution over CP patterns $\mathbf{z} = (z_{1:K})$ for a given patient trace σ. Given a distribution of CP patterns for a patient trace, treatment activities are generated by sampling CP patterns from this distribution. The result is a vector of $|\sigma|$ activities $\mathbf{a} = (e_{1:|\sigma|} \cdot a)$ for a patient trace σ $(e_1, e_2, \ldots, e_{|\sigma|} \in \sigma)$. The steps adapted for patient traces are summarized below:

1. Select a multinomial distribution ϕ_z for each CP pattern z from a Dirichlet distribution with parameter β.
2. For each patient trace σ in a clinical event log $\mathcal{L}$, select a multinomial parameter θ_σ over K CP patterns is sampled from a Dirichlet distribution with parameter α.
3. For the activity type a of each event e in σ, select a CP pattern z from θ_σ,
4. Select an activity a from ϕ_z.

In LDA, each patient trace is a mixture of CP patterns represented by θ_σ and each pattern is a distribution over treatment activities represented $\phi_{z,a} = P(a|z)$, which represents CP patterns

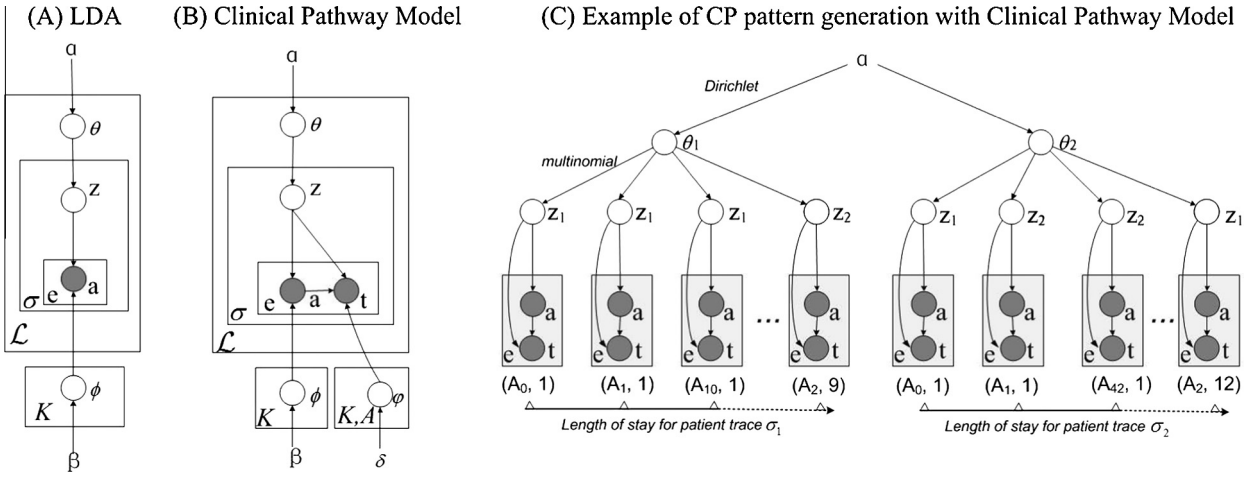


Fig. 2. Graphical representation of two probabilistic models (A) Latent Dirichlet Allocation (LDA), (B) Clinical Pathway Model (CPM), and (C) A possible generative process for CP patterns when modeled as CPM, which can be viewed as the expanded graphical model of the plate representation in (B). In this example, we assume that there are two underlying CP patterns z_1 and z_2. A patient trace containing a set of clinical events, which are spread along the time-line of length of stay, is mixed with both CP patterns.

that we are interested to infer. Exact inference in LDA is known to be intractable. Options include the variational approach, expectation propagation or collapse Gibbs sampling [26,27]. In our previous work [18], we use Gibbs sampling to iteratively draw samples from the conditional distribution for each CP pattern z_i after marginalizing out the parameters:

$$P(z_i^{\sigma} = z|\mathbf{z}_{-i}, e_i.a = a, \mathbf{a}, \alpha, \beta) \propto \frac{m_{\sigma,z}^{-i} + \alpha_z}{n_{\sigma,*}^{-i} + \alpha_*} \times \frac{n_{z,a}^{-i} + \beta_a}{n_{z,*}^{-i} + \beta_*} \quad (2)$$

where $z_i^{\sigma} = z$ represents the assignments of the ith event in a patient trace σ to pattern z, and $z_{-i,\sigma}$ represents all pattern assignments not including the ith event. Furthermore, $m_{z,a}^{-i}$ is the number of times clinical events whose activity types are a are assigned to pattern z, not including the current instance, $n_{\sigma,z}^{-i}$ is the number of times pattern z has occurred in trace σ, not including the current instance, and the dot $*$ denotes the summing operation at the corresponding index, e.g., $m_{z,*} = \sum_{a \in A} m_{z,a}$ or $\beta_* = \sum_{a \in A} \beta_a$, where A is the universe of treatment activities.

The first term of Eq. (2) is proportional to the number of the current CP pattern z within the patient trace σ and the second term is proportional to the count of the current activity a in σ to z. Intuitively the effect of co-occurrence is achieved by assigning higher probability to two activities in the same trace being assigned to the same pattern. For any sample from this Markov chain, being an assignment of each activity to a treatment pattern and using symmetric Dirichlet distribution (i.e., each distribution is parameterized by a single parameter), we can estimate ϕ and θ as follows:

$$\phi_{z,a} = \frac{m_{z,a} + \beta}{m_{z,*} + |A|\beta} \quad (3)$$

$$\theta_{\sigma,z} = \frac{n_{\sigma,z} + \alpha}{n_{\sigma,*} + K\alpha} \quad (4)$$

For more details of LDA-based CP pattern discovery from clinical event logs, please refer to [18].

3. Clinical pathway model

Realizing the needs above for the discovery of meaningful CP patterns, in this paper, we extend our previous work to propose a richer Bayesian topic model, i.e., clinical pathway model (CPM), for the task of hidden CP pattern extraction from clinical event logs, which offers solutions to the previously mentioned problem.

The proposed CPM can discover latent CP patterns that are determined by both treatment activities and their occurring time stamps. The proposed model first generates a treatment activity to be performed, and then generates its occurring time stamp. The generative process is similar with that of standard topic models. For a specific event log $\mathcal{L}$ recording routine treatment behaviors of a particular CP, it has pattern proportions θ that are sampled from a Dirichlet distribution, and each CP pattern is associated with a multinomial distribution ϕ_z over treatment activities and multinomial distribution $\varphi_{z,a}$ over the time stamp of the occurred activity a for pattern z. So, θ, ϕ_z and $\varphi_{z,a}$ have a symmetric Dirichlet prior with hyper parameters α, β and δ, respectively.

In summary, the proposed model assumes the following generative process for a clinical event log:

1. Draw Discrete distribution ϕ_z from a Dirichlet prior β for each CP pattern z;
2. Draw Discrete distribution $\varphi_{z,a}$ from a Dirichlet prior δ for each CP pattern z and each treatment activity a;
3. For each patient trace σ, draw a Discrete distribution θ_σ from a Dirichlet prior α, and then for each clinical event e_i in patient trace σ:

   (a) Draw CP pattern z_i from Discrete θ_σ; and
   (b) Draw treatment activity $e_i{\cdot}a$ from Discrete ϕ_{z_i}
   (c) Draw time stamp $e_i{\cdot}t$ from Discrete $\varphi_{z_i,e_i.a}$.

Fig. 2B shows a graphical model representation of the proposed CPM for clinical event logs, which represents dependencies among variables. Here the shaded and unshaded nodes indicate observed and latent variables, respectively. The proposed model is an extension of the standard topic model. For a particular event e_i, the time stamp $e_i{\cdot}t$ is generated depending on CP pattern z and activity $e_i{\cdot}a$ in the proposed CPM. Fig. 2C shows a possible generative process for CP patterns when modeled as CPM with the patient traces σ_1 and σ_2 in the example log shown in Table 1. It can be viewed as the expanded graphical model of the plate representation in Fig. 2B. A clinical event e has two specific properties, i.e., treatment activity $e{\cdot}a$ and the occurring time stamp $e{\cdot}t$, in which $e{\cdot}a$ depends on the immediate CP pattern variable z, and $e{\cdot}t$ depends on both the CP pattern variable z and the treatment activity $e{\cdot}a$. Thus with CPM, explicit CP patterns are modeled.

As indicated in [26], inference in the LDA model family is intractable, thus we resort to an approximate technique, Gibbs sampling, to derive efficient sampling routines for CPM. Formally, we denote

$z_\sigma = (z_{1,\sigma}, z_{2,\sigma}, \ldots, z_{|\sigma|,\sigma})$ as the vector of CP pattern assignments within the patient trace σ and $\mathbf{z} = \{z_\sigma | \sigma \in \mathcal{L}\}$ as the concatenated pattern assignments for the entire log. This convention is also applied for clinical events $\{e_\sigma, \mathbf{e}\}$, treatment activities $\{a_\sigma, \mathbf{a}\}$, and time stamps $\{t_\sigma, \mathbf{t}\}$.

Our interest is in the distribution $P(\mathbf{z}, \mathbf{e}|\alpha, \beta, \delta)$. As we mentioned above, exact inference is known to be intractable in the LDA family, and thus we resort to approximate inference using Gibbs sampling. This is achieved by integrating out the parameter random variables, and taking advantages of conjugacy to derive a close form for the Gibbs conditional distribution $P(z_{\sigma,i}|z_{\sigma,-i}, \mathbf{e}, \alpha, \beta, \delta)$. Because the parameter variables are integrated out during sampling, this is also known as collapsed Gibbs sampling. We start the joint distribution as follows:

$$P(\mathbf{z}, \mathbf{e}|\alpha, \beta, \delta) = P(\mathbf{z}, \mathbf{a}, \mathbf{t}|\alpha, \beta, \delta) = P(\mathbf{z}|\alpha)P(\mathbf{a}|\mathbf{z}, \beta)P(\mathbf{t}|\mathbf{z}, \mathbf{a}, \delta) \tag{5}$$

For $P(\mathbf{z}|\alpha)$, we have:

$$P(\mathbf{z}|\alpha) \propto \prod_{\sigma=1}^{|\mathcal{L}|} \frac{\prod_{z=1}^{K} \Gamma(m_{z,\sigma} + \alpha_*)}{\Gamma(m_{\sigma,*} + \alpha_*)} \tag{6}$$

where $\Gamma(\cdot)$ is the gamma function, $m_{z,\sigma}$ is the count of observing that patient trace σ is assigned to CP pattern z, and $*$ denotes the summing operation at the corresponding index, e.g., $m_{\sigma,*} = \sum_{z=1}^{K} m_{z,\sigma}$ or $\alpha_* = \sum_{z=1}^{K} \alpha_z$.

For $P(\mathbf{a}|\mathbf{z}, \beta)$, we have:

$$P(\mathbf{a}|\mathbf{z}, \beta) \propto \prod_{z=1}^{K} \frac{\prod_{a=1}^{|A|} \Gamma(\beta_a + n_{z,a})}{\Gamma(\beta_* + n_{z,*})} \tag{7}$$

where $n_{z,a}$ is the count of observing that activity a is assigned to pattern z, and the dot $*$ denotes the summing operation at the corresponding index, e.g., $n_{z,*} = \sum_{a=1}^{|A|} n_{z,a}$ or $\beta_* = \sum_{a=1}^{|A|} \beta_a$.

For $P(\mathbf{t}|\mathbf{z}, \mathbf{a}, \delta)$, we have:

$$P(\mathbf{t}|\mathbf{z}, \mathbf{a}, \delta) \propto \prod_{z=1}^{K} \prod_{a=1}^{|A|} \frac{\prod_{t=1}^{|T|} \Gamma(\delta_t + q_{z,a,t})}{\Gamma(\delta_* + q_{z,a,*})} \tag{8}$$

where $q_{z,a,t}$ is the count of observing that activity a occurring at time stamp t is assigned to pattern z, and $*$ denotes the summing operation at the corresponding index, e.g., $q_{z,a,*} = \sum_{t=1}^{|T|} q_{z,a,t}$ or $\delta_* = \sum_{t=1}^{|T|} \delta_t$.

Our objective is to derive the conditional Gibbs distribution $P(z_{\sigma,i} = z|z_{\sigma,-i}, \mathbf{e}, \alpha, \beta, \delta)$, where $z_{\sigma,-i} = \mathbf{z}/\{z_{\sigma,i}\}$ denotes the set of remaining pattern variables except at position i in the patient trace σ. Substituting Eqs. (6)–(8) into Eq. (5), and using symmetric Dirichlet distribution, we obtain the conditional Gibbs distribution as follows:

$$P(z_{\sigma,i} = z|z_{\sigma,-i}, \mathbf{e}, \alpha, \beta, \delta) \propto \frac{m_{\sigma,z} + \alpha}{m_{\sigma,*} + K\alpha} \cdot \frac{n_{z,a} + \beta}{n_{z,*} + |A|\beta} \cdot \frac{q_{z,a,t} + \delta}{q_{z,a,*} + |T|\delta} \tag{9}$$

Details of the derivation are in Appendix A. In all cases, the current sampling position is always excluded during counting.

Consider Eq. (9), which computes a probability of a certain CP pattern for the present event i in the patient trace σ. In particular, the occurring time information of clinical activities is inferred from Eq. (9). Suppose that we are currently determining the probability that the pattern of the present activity a is z. Eq. (9) determines the probability of the occurring time of a under z.

The pseudo-code for CP pattern extraction with CPM is shown in Algorithm 1, where the posterior estimates in Eqs. (10)–(12) are used to derive the patterns.

Algorithm 1. Gibbs sampling for CPM.

1: **Input**:
2: A clinical event log $\mathcal{L}$, hyper-parameters α, β, and δ
3: **Steps**:
4: **Initialize** $\mathbf{z}_\sigma, \mathbf{e}_\sigma$ for all $\sigma \in \mathcal{L}$ randomly
5: **Iterate** over a large number of iterations (e.g. 1000):
6: **Iterate** over each patient trace σ in the log $\mathcal{L}$
7: **Iterate** over each clinical event e in the trace σ
8: **Sample** a CP pattern z assignment for e according to Eq. (8)
9: **Output**:
10: Estimate the model parameters as follows:

$$\hat{\phi}_{z,a} = \frac{n_{z,a} + \beta}{n_{z,*} + |A|\beta} \tag{10}$$

$$\hat{\theta}_{\sigma,z} = \frac{m_{\sigma,z} + \alpha}{m_{\sigma,*} + K\alpha} \tag{11}$$

$$\hat{\varphi}_{z,a,t} = \frac{q_{z,a,t} + \delta}{q_{z,a,*} + |T|\delta} \tag{12}$$

The complexity of driving a sample from Eq. (9) is $O(K)$, making an over complexity of $O(NK)$ for each Gibbs round, where K is the number of CP patterns in consideration, and $N = \sum_{\sigma \in \mathcal{L}} |\sigma|$ is the number of events in clinical event log $\mathcal{L}$. Thus, the total complexity for S Gibbs samples is $O(SNK)$. In practice, K and S are often fixed in advance, making N the main factor that contributes to the complexity. In other words, the proposed CPM scales linearly with the number of events recorded in the log.

We utilize the proposed CPM as a means of characterizing the underlying CP patterns from a specific clinical event log. Fig. 3 presents a schematic representation of assumptions about learning, which are built into the model. It is assumed that each patient trace contains a set of CP patterns with initial proportions. During patient careflow, treatment behaviors performed on the patient updates the CP pattern proportions. The task for the model is to

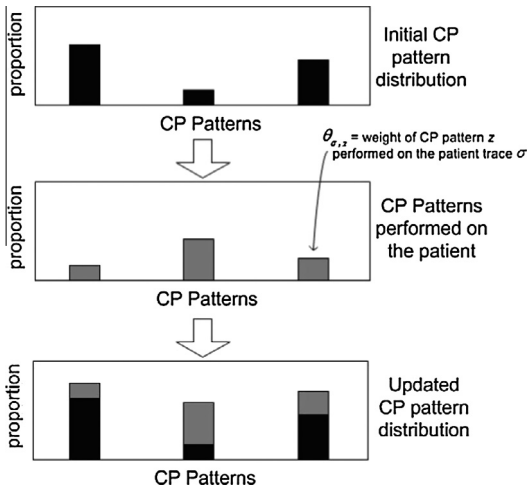


Fig. 3. Notional model of CP pattern discovery based on CPM. A patient trace is a mixture of CP patterns that they are performed on the patient in his or her careflow. It focuses on establishing the predictive power of a CPM fit to pretest measures of initial CP pattern proportions, estimates of pattern based upon treatment behaviors recording in the patient trace, and posttest measures of updated CP pattern proportions.

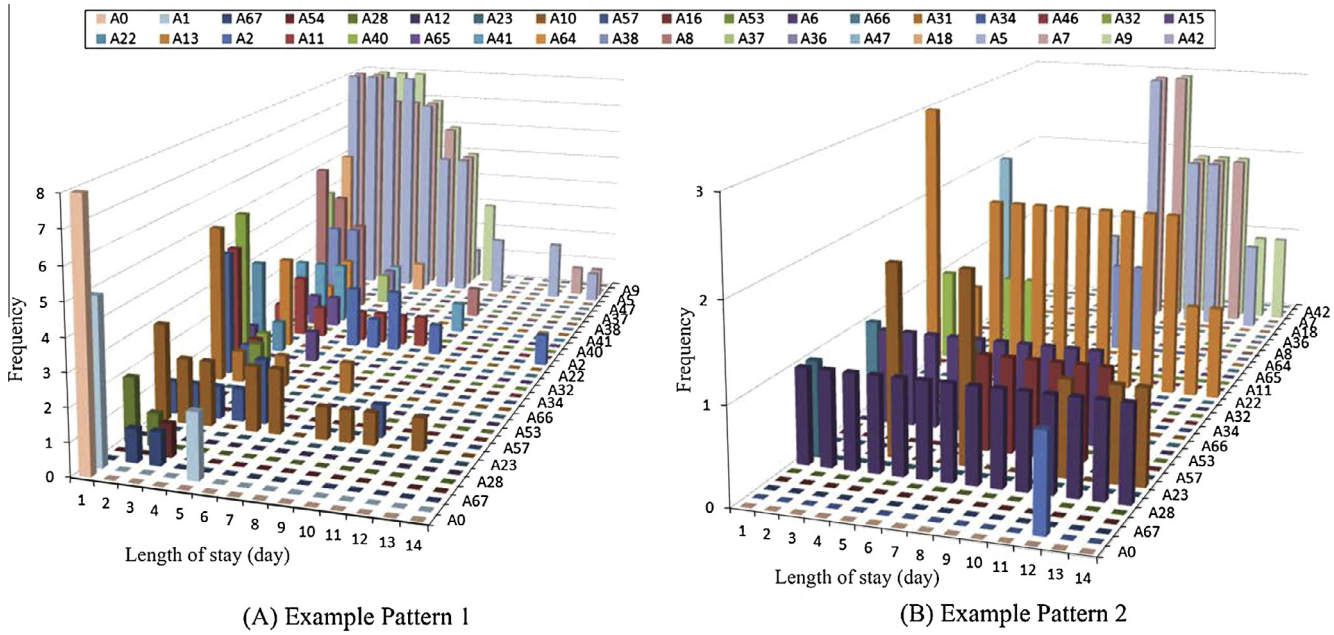


Fig. 4. Patterns discovered from the example log shown in Table 1.

Table 2
Pattern-trace distribution of the example log.

Patient trace no.	Pattern no.	Probability	Pattern no.	Probability
σ_1	1	0.885	2	0.115
σ_2	1	0.494	2	0.506
σ_3	1	0.396	2	0.604
σ_4	1	0.997	2	0.003
σ_5	1	0.883	2	0.117
σ_6	1	0.997	2	0.003
σ_7	1	0.997	2	0.003
σ_8	1	0.996	2	0.004

induce the CP patterns mixed in the patient trace from the treatment behaviors performed on his or her careflow, and to relate them to update the probability distributions of CP patterns over the patient trace.

Taking the log shown in Table 1 as an example, and assuming the number of latent CP patterns K is 2. The typical activity labels and their occurring time stamps for the derived patterns shown in Fig. 4. The CP pattern-patient trace distribution $p(z|\sigma)$ measures the connection (or relatedness) of a specific patient trace with a specific CP pattern (i.e., the conditional probability of a CP pattern in a given patient trace). We used this statistical probability to group patient traces by associating them with patterns. Taking the log shown in Table 1 as an example, the pattern-trace distribution values are listed in Table 2. The traces are grouped into specific clusters based on their pattern-trace distribution values. For example, the traces σ_1, σ_3, σ_4, σ_5, σ_6, σ_7 and σ_8 have the similar probability distribution to belong to pattern 1. Thus, these six traces are grouped into the same cluster. While for patient traces σ_2 and σ_3 in the example log, they have similar probability distributions over the derived patterns such that both traces can be grouped into the other cluster.

As we can see, CPM is now explicitly capturing the associations between treatment activities and their occurring time stamps. By learning the parameters of the model, we obtain the set of CP patterns that appear in a clinical event log and their relevance to different patient traces, which truly models essential treatment activities and also possible occurring time instants of these activities. Thus we are able to reveal the temporal structure of discovered CP patterns.

4. Experiments and results

In order to evaluate both suitability and generality of our approach, in this work we have considered cases of study in two different scenarios where optimal CPs are critical: a typical cardiovascular disease – unstable angina, and oncology, through the cooperations with the Chinese PLA General hospital and Zhejiang Huzhou Central hospital of China, respectively.

In experiments, we validate the proposed CPM on the collected unstable angina event log and oncology log, respectively. The baseline in comparisons is LDA, a probabilistic generative model originally developed for modeling text documents [26] and has already applied in CP pattern mining in our previous work [18]. All experiments were performed on a Lenevo Compatible PC with an Intel Pentium IV CPU 2.8 GHz, 4G byte main memory running on Microsoft Windows 7. The algorithms were implemented using Microsoft C#.

4.1. Model selection

An input required for the proposed CPM is the number of topics, i.e., the number of CP patterns to be discovered. In the case studies, we use a common measure on the ability of a model to generalize to unseen data, i.e., perplexity, for this model selection task.

Perplexity is defined as the reciprocal geometric mean of the likelihood of a test corpus given a model. The perplexity score has been widely used in LDA to determine the number of topics, which is a standard measure to evaluate the prediction power of a probabilistic model suggested in many literature [26,28–30]. It is defined as the reciprocal geometric mean of the likelihood of a clinical event log given a model,

$$\text{Perplexity} = exp\left[-\frac{\sum_{\sigma\in\mathcal{L}} logP(\mathbf{e}_\sigma|\mathcal{M})}{\sum_{\sigma\in\mathcal{L}}|\sigma|}\right] \quad (13)$$

where $\mathcal{M}$ is the model, and $\mathbf{e}_\sigma$ are the set of unseen events in the patient trace σ.

4.2. Unstable angina log

In the first case study, the experimental log is collected from the cardiology department at the Chinese PLA General hospital. The CP of unstable angina is selected in this case study. Unstable angina is a kind of chest discomfort or pain that occurs in a continuous and unpredictable way. The cause of angina is commonly the poor blood flow in coronary vessels caused by atherosclerosis and the lack of oxygen supply to the myocardium. The unstable pain can result from the disruption of an atherosclerotic plaque in narrowed coronary vessels with lessened flexibility, embolization and vasospasm. Unstable angina lay its symptoms between exertional stable angina and acute myocardium infarction and a further sudden death. While the risk of unstable angina is high, the population of unstable angina is huge, especially for aged people and those with associated disease such as hypertension and diabetes [31]. Thus, the discovery of underlying patterns in the unstable angina CP will be of significant value and interest. Discovered patterns can provide the user explicit suggestions of treatment actions to influence medical behaviors in concern for the patient's benefit.

In this case study, 2934 patient traces following the unstable angina CP were selected from the Department of Cardiology to demonstrate the ability of the proposed method to discover latent patterns of the unstable angina CP. These patient traces have 483,349 clinical events within 67 event types. Details of the unstable angina log are shown in Tables 3 and 4.

4.2.1. CP pattern discovery

Regarding latent CP pattern discovery by the proposed CPM, we set Dirichlet prior α, β, and δ of CPM as 0.1, 0.01, and 0.01, respectively, which are common settings in literature. The number of iterations of the Markov chain for Gibbs sampling is set to 1000. Note that Gibbs sampling usually converges before 1000 iterations for the collected log. For LDA, we use the same model parameters as those for CPM. Specifically, $\alpha = 0.1$, $\beta = 0.01$, and we run 1000 iterations of the Gibbs sampling algorithm for LDA.

Fig. 5 shows the perplexity curve with respect to the number of CP patterns, using the proposed CPM. The lower the perplexity, the better the derived model fits with the collected log. In general, the model perplexity decreases with the number of pattern increases. On the other hand, if the number of patterns is larger, the derived model may be over explain the log and it spends more sampling computation and storage as well [32]. Thus, it needs to choose a balance between simplicity of the model and the degree of fitness.

Table 3
The details of the experimental event logs.

Disease type	Number of behavioral instances	Number of events	Number of activity types	Maximum length of stay (day)	Minimum length of stay (day)	Average length of stay (day)
Unstable Angina	2934	483349	67	93	1	9
Oncology	258	11028	266	66	2	25.39

Table 4
The common set of treatment activities contained in the unstable angina log.

Abbreviation	Description	Abbreviation	Description
A_0	Admission	A_{34}	ABO Blood Type
A_1	Second-Grade Nursing	A_{35}	Blood Gas Analysis
A_2	Discharge	A_{36}	Urine Routine
A_3	Transfer	A_{37}	Fecal Routine
A_4	Expired	A_{38}	PCI (Percutaneous Coronary Intervention)
A_5	Unblocking	A_{40}	Oxygen Uptake
A_6	BGR (Blood Glucose Regulation)	A_{41}	ECG Monitoring and Oxygen Saturation Monitoring
A_7	Antiplatelet	A_{42}	Heart Rate Regulation
A_8	Anticoagulation	A_{43}	Sputum Bacterial Culture and Strain Identification
A_9	Blood Lipid Regulation	A_{44}	Tuberculosis Three Items Test
A_{10}	Blood Pressure Regulation	A_{45}	Eliminating Phlegm
A_{11}	Sedative	A_{46}	Serum Four Items Test
A_{12}	Liver Recovery	A_{47}	Occult blood test
A_{13}	Diuretic	A_{48}	ECT
A_{14}	Antibiotics	A_{49}	Blood Bacterial Culture and Strain Identification
A_{15}	ECG Examination	A_{50}	Body Fluid Bacterial Culture and Strain Identification
A_{16}	Blood Glucose Test	A_{51}	Exercise Testing
A_{17}	Heart Ultrasonography	A_{52}	Holter
A_{18}	X-ray	A_{53}	Thrombelastometry
A_{19}	CT	A_{54}	Abdominal ultrasound
A_{20}	MRI	A_{55}	Promoting blood circulation for removing blood stasis
A_{21}	Coronary Angiography	A_{58}	Vascular ultrasound
A_{22}	Blood Biochemistry	A_{59}	Nursing care for seriously ill patient
A_{23}	Biochemistry Full Term	A_{60}	Nursing care for critically ill patient
A_{24}	Liver Function	A_{61}	Fecal Bacterial Culture and Strain Identification
A_{25}	Renal Function	A_{63}	Electrical Defibrillation
A_{28}	First-Grade Nursing	A_{64}	Multifunctional Intensive Monitoring
A_{29}	Immunological General Examination	A_{65}	High-sensitivity C-reactive protein (HS-CRP)
A_{30}	Thyroid Function	A_{66}	Echocardiogram
A_{31}	Blood Routine	A_{67}	TnT
A_{32}	Coagulation Tests	A_{68}	Permanent Pacemaker Implantation
A_{33}	ESR (Erythrocyte Sedimentation Rate)		

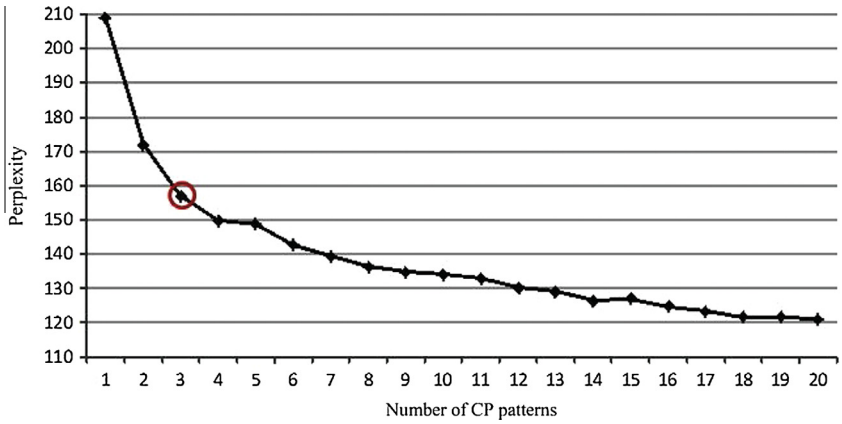


Fig. 5. Choosing number of CP patterns using perplexity for the unstable angina log.

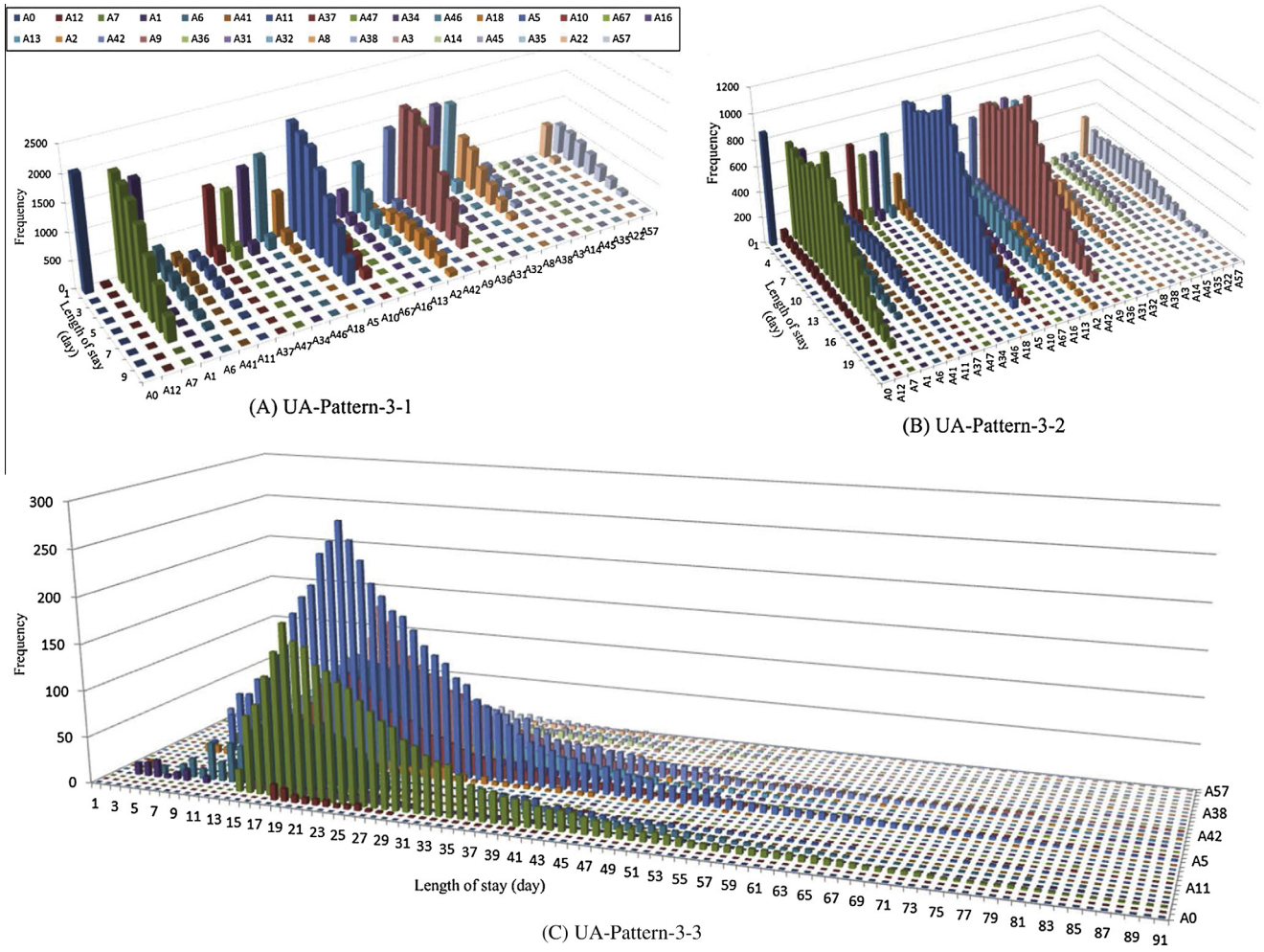


Fig. 6. The discovered CP patterns by CPM with $K = 3$, for the unstable angina log.

To select the appropriate number of CP patterns, we examine the discovered patterns by CPM with different value of K. In particular, we interpret the results by examining the probability $\hat{\phi}_{z,a}$ of a treatment activity a given a CP pattern z, and the probability $\hat{\varphi}_{z,a,t}$ of an occurring time stamp t of the treatment activity a given the pattern z, with a fixed number of CP patterns. In other words, we look at those activities which are assigned to a CP pattern z and their occurring time points, with high probability. In the experiments, we select a set of representative treatment activities $\{a|\forall z \in Z, P(a|z) > 0.01\}$ to represent discovered patterns.

Figs. 6–8 show the results of discovered patterns by the proposed CPM with $K = 3$, $K = 4$ and $K = 5$, respectively, in which each row corresponds to a specific treatment activity, and the height of each bar depicts the frequency of activity accumulated over all collected Gibbs samples. The detected frequency of how often a treatment activity occurring on a particular time instant for that CP

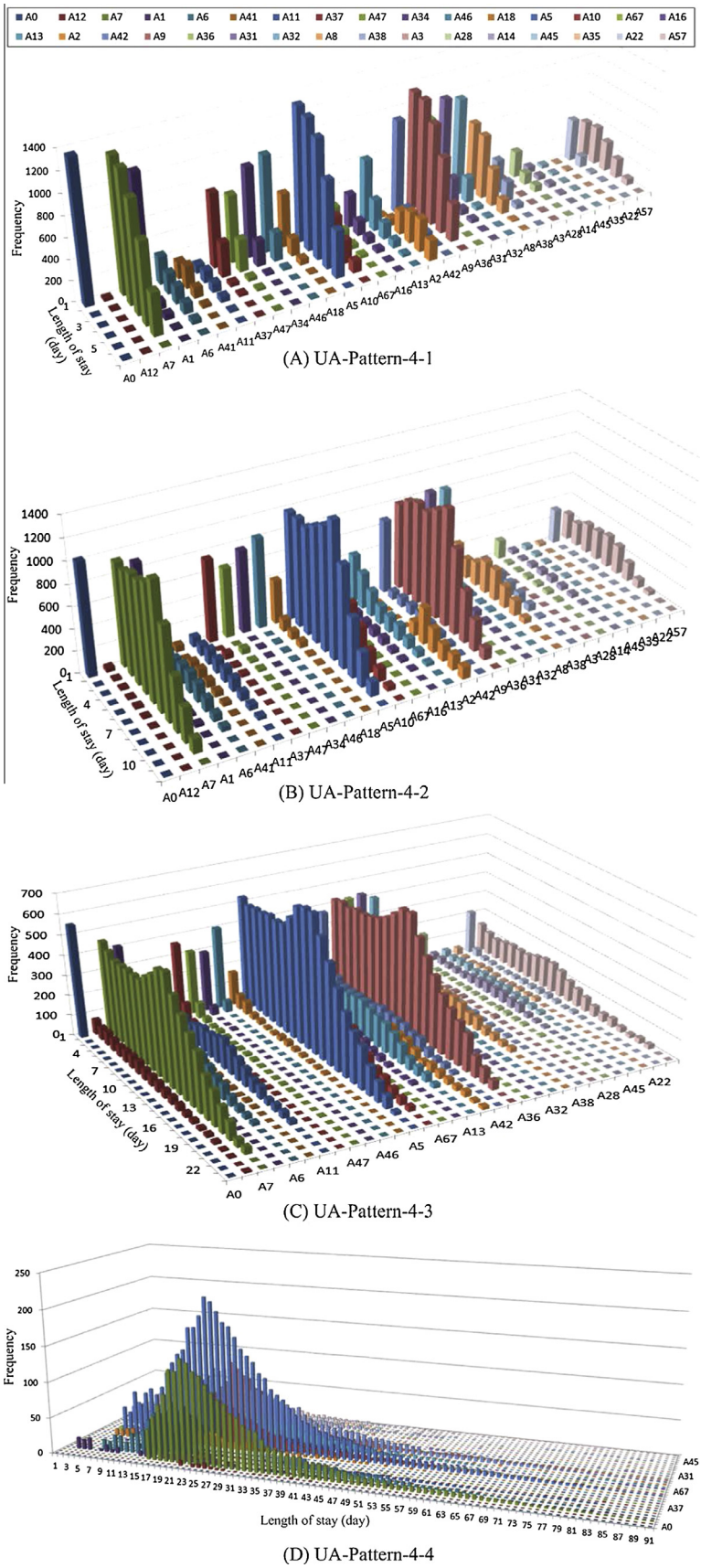


Fig. 7. The discovered patterns by CPM with $K = 4$, for the unstable angina log.

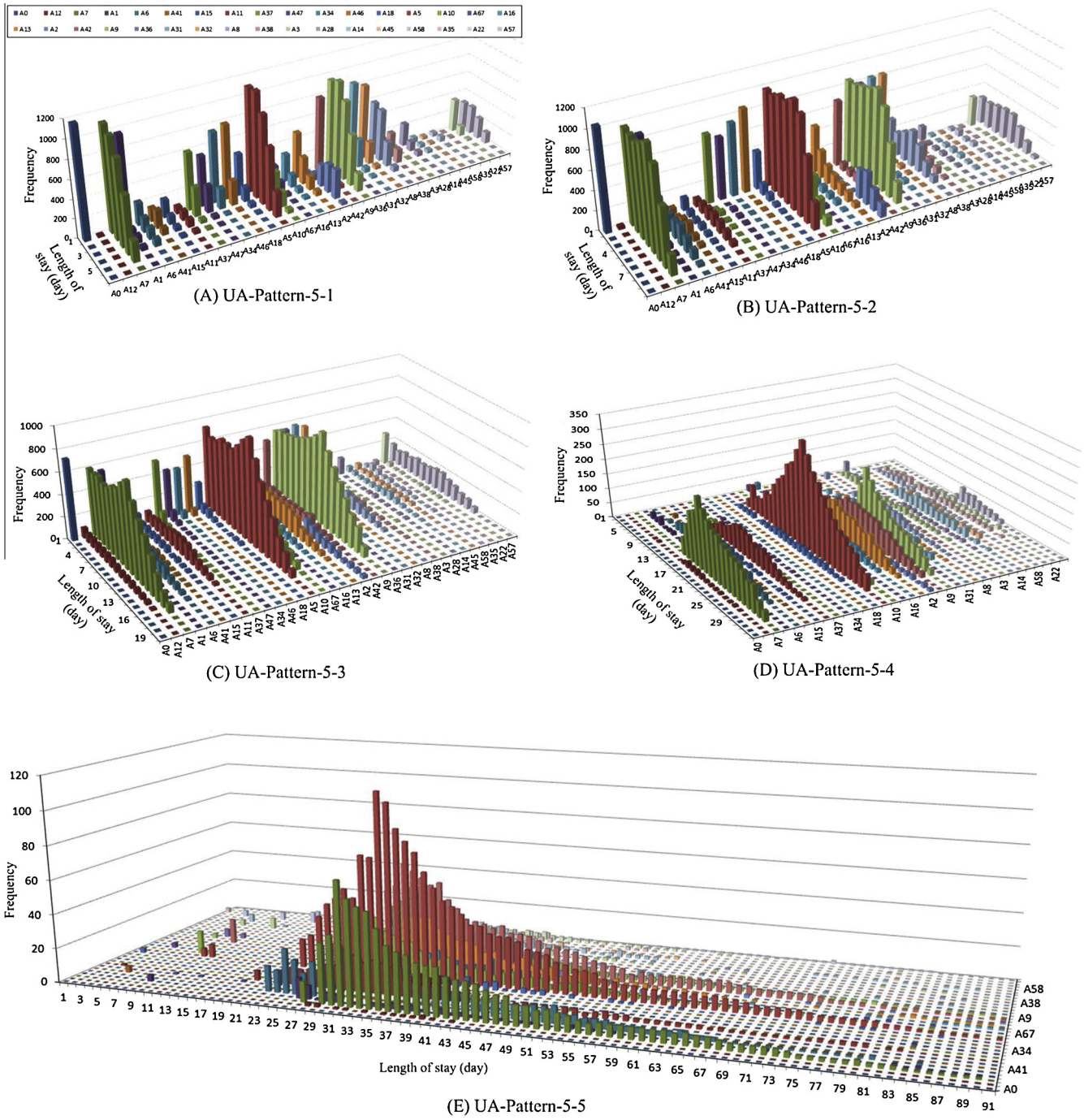


Fig. 8. The discovered patterns by CPM with $K = 5$, for the unstable angina log.

pattern, after being normalized, becomes its probability assigned to the pattern. As shown in Figs. 6–8, the proposed CPM derives quite similar CP patterns with different K (e.g., "UA-Pattern-3-3", and "UA-Pattern-4-4"). In addition, the derived patterns share almost the same representative treatment activities with specific probability distributions. However, a larger value of K may make the learned model over-fitting. For example, as shown in Fig. 7, discovered patterns "UA-Pattern-4-1" and "UA-Pattern-4-2" have similar probability distributions for the representative treatment activities although occurred time instants of these activities are slightly different with each other. As indicated by clinical experts, both patterns can be merged into one (i.e., "UA-Pattern-3-1") to effectively represent typical treatment behaviors of UA patient careflow in which there is little variation occurred. As a result, we empirically choose the number of patterns $K = 3$ for the unstable angina log, where the perplexity also seems to decrease rapidly and appear to settle down as shown in Fig. 5.

In comparison with discovered patterns by using LDA (as shown in Table 5), the results from CPM are rather interesting, which disclose the temporal structure to the pattern explicitly. In particular, from Fig. 6, we can see that related treatment activities are clustered for specific patient groups. For example, the first discovered pattern "UA-Pattern-3-1" indicates typical treatment behaviors of unstable angina patient careflow in which there is little variation occurred and common treatment activities (e.g., 'PCI-surgery') are carried out smoothly. In clinical practice, patients who follow this

Table 5
CP pattern discovery results with LDA for the unstable angina log. Top treatment activities are listed for these patterns, ranked by $p(a|z)$.

Pattern id	Significant activities
1	Unblocking, Blood Lipid Regulation, Antiplatelet, Blood Pressure Regulation, Promoting Blood Circulation, Diuretin, Sedative, Anticoagulation, Blood Routine, Admission, Discharge, Coagulation, Urine Routine, Fecal Examination, Serum Test, Heart Rate Regulation, Occult Blood Test, ABO, X-ray
2	Unblocking, BGR, Antiplatelet, Heart Rate Regulation, Diuretin, Blood Glucose Test, Blood Lipid Regulation, Blood Pressure Regulation, Sedative, Blood Routine, Antibiotics, X-ray, Eliminating Phlegm, Anticoagulation, TnT, Blood Gas Analysis, Coagulation, Transfer to Cardiovascular Surgery Department, Discharge
3	Unblocking, Blood Lipid Regulation, Antiplatelet, Anticoagulation, Diuretin, Blood Routine, Admission, Discharge, Heart Rate Regulation, Coagulation, Urine Routine, Serum Test, Fecal Examination, ABO, Occult Blood Test, Liver Recovery, X-ray, PCI (Percutaneous Coronary Intervention), Blood Biochemical Examination, ECG Examination

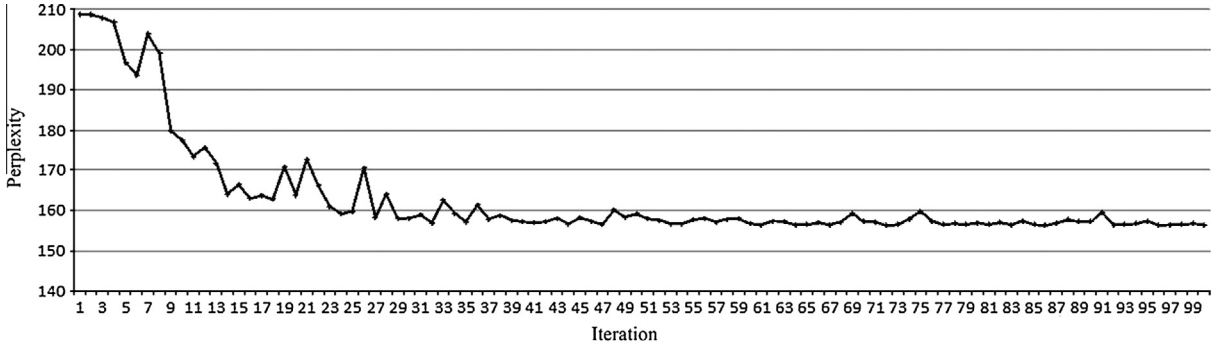


Fig. 9. Perplexity of CPM over iterations for the unstable angina log.

dominating pattern have shorter length of stay (in general 3–9 days) than others, and almost all physical examinations (e.g., 'X-ray', 'Occult Blood Test', 'BGR', etc.) are performed on the first day after admission. "UA-Pattern-3-2" discloses treatment behaviors of unstable angina patients who cannot take 'PCI' surgery due to specific physical problems, e.g., coronary stenosis. In general, LOS of patients who follow "UA-Pattern-3-2" is larger than 9 days (i.e., 'Discharge' occurs on 9–21 days, as shown in Fig. 6B). Moreover, it is interesting to see that "UA-Pattern-3-3", as shown in Fig. 6C, has captured typical treatment behaviors of unstable angina patients who are transferred to the Cardiovascular Surgery Department during their LOS. Note that this variant pattern is quite normal in the unstable angina CP (312 out of 2934 patient traces in the collected log). In addition, we find that most patient traces which follow "UA-Pattern-3-3" also take "UA-Pattern-3-1" or "UA-Pattern-3-2" as well. Typically, they are best represented as a mixture of CP patterns.

LDA has also discovered certain meaningful patterns, but the results are much less interpretable. In particular, patterns discovered by LDA have failed to capture the frequent occurring time stamps of treatment activities. In contrast, CPM grouping gives better and more meaningful interpretation of CP patterns. Note that discovered patterns share a lot of common treatment activities, such as 'Unblocking', 'Blood Lipid Regulation', etc., as shown in the experimental results. However, the occurring time instants of these activities are variable from discovered patterns. In clinical settings, the variability of treatment behaviors is an important aspect that needs to be addressed in CP analysis and optimization. One of the strengths of the proposed CPM is the ability to consider variations of patients' treatment behaviors. Indeed, given the high variability of patient careflow, the same treatment activity can be performed in different time instants. The variations in patient careflow are thus reflected in the CP patterns extracted by the proposed CPM, which provide comprehensive knowledge to medical staff, and allow them to realize/study which treatment behaviors can be performed and during which time periods these behaviors can be performed in CPs.

Fig. 9 shows the perplexities of the proposed model over iterations on inference in the unstable angina log, with a fixed number of CP patterns ($K = 3$). As the number of iteration increases, the perplexity decreases, and eventually converges to a certain point. It confirms our assumption that Gibbs sampling usually converges before 1000 iterations for the unstable angina log.

Next, we study how the proposed CPM performs with the increasing size of a clinical event log. Fig. 10 shows how the presented approach scales up as the number of input-clinical pathway traces is increased, from 300 to 2934. Note that the experiments were performed on the unstable angina log with the number of CP patterns $K = 3$. The running times are normalized with respect to the time for the 300 input-traces. It can be observed that the presented algorithm has a linear scalability in terms of the runtimes against the increasing number of traces.

4.3. Oncology log

We argue that the proposed CPM can assist in getting better insights into clinical pathways, and provide a solid basis for further CPA tasks. In this sub section, we present a case study on an oncology log extracted from the hospital information system of Zhejiang Huzhou Central hospital of China. The collected log contains typical treatment behaviors in CPs of five specific types of cancers, i.e., bronchial lung cancer, colon cancer, rectal cancer, breast cancer, and gastric cancer. We, firstly, derive latent CP patterns from the log, and then present a possible CPA task, i.e., patient trace clustering, to illustrate the advantage of the proposed CPM. Since the oncology log contains general categories of cancers, they can be used as benchmark clusters for evaluating the overall performance of clustering. In detail, there are 258 patient traces, 11,028 clinical events within 266 event types. The average LOS of these traces is 25.39 days while some traces take a very short time, e.g., only

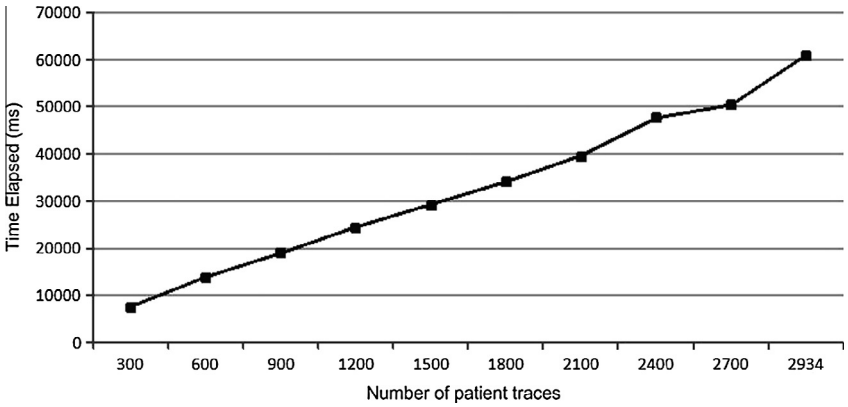


Fig. 10. Scale-up: Number of input traces of the unstable angina log.

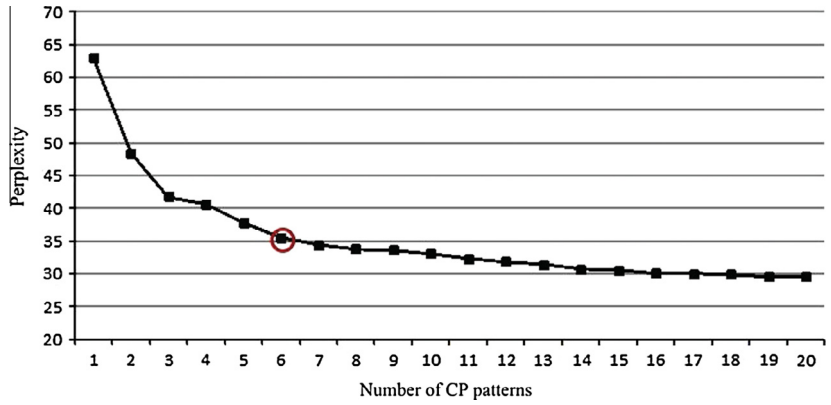


Fig. 11. Perplexity on the oncology log.

4 days in hospital, and other traces take much longer, e.g., 66 days in the hospital (see Table 3).

4.3.1. CP pattern discovery

Constructing CPM is to fit latent CP patterns to the clinical event log. In generating CPM, the Dirichlet prior α, β, and δ are set to 0.1, 0.01, and 0.01 respectively, which are common settings in literature. The number of iterations of Gibbs sampling is set to 10,000. In general, Gibbs sampling converges before 10,000 iterations for the experiments.

Again, we use perplexity as a measure to determine the optimal number of latent CP patterns, K. We computed perplexity for CPM using K values from 1 to 20. For all values of K, initialization was followed by 1000 iterations of the Gibbs sampling algorithm. The perplexity is plot over the number of latent CP patterns in Fig. 11. A drop in perplexity occurs at approximately $K = 6$ CP patterns, after which the perplexity stabilizes. We choose $K = 6$ as the number of latent CP patterns for the oncology log.

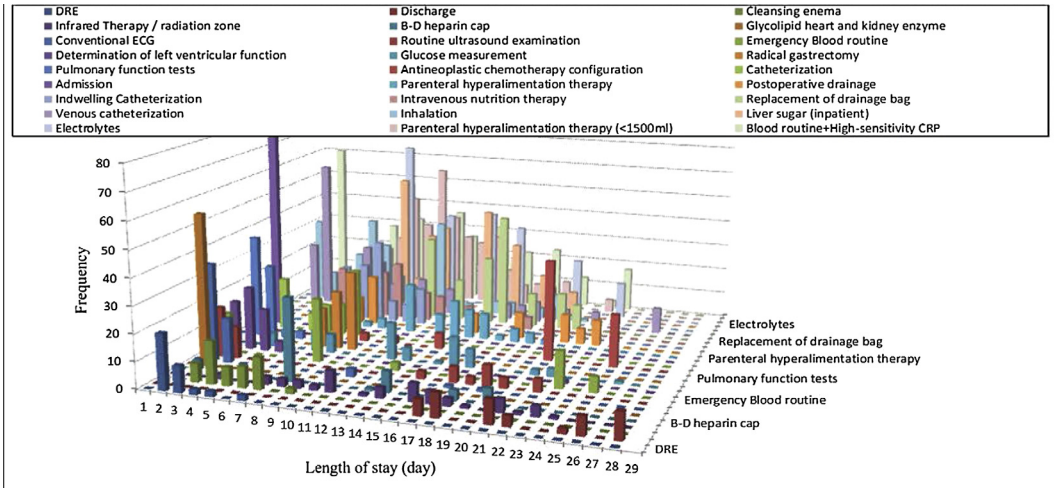
Let's now examine the contents of the patterns, i.e. the learned activity labels and their occurring time stamps that have a high probability of being part of a particular pattern. We examine clinical activities associated with each treatment pattern to evaluate the quality of discovered patterns. For each pattern z, we list all activity labels a with $p(a|z) \geqslant 0.01$. As shown in Figs. 12 and 13, the content often represents a meaningful set of activity labels. We can see that patterns (A)–(E) detect typical treatment behaviors of the patients with gastric cancer, bronchial lung cancer, colon cancer, rectal cancer, and breast cancer respectively. Discovered pattern (F) tends to capture variant treatment behaviors of oncology patients who have serious complications after surgery. As shown in Fig. 13F, major treatment behaviors of pattern (F) occur in the latter part of inpatient LOS (i.e., after surgery), and many of them are emergent treatment activities, e.g., "Order: Calcium determination (Emergency)", "Order: Potassium determination (Emergency)" etc. Note that this is an evidence that patient traces are typical "mixture of CP patterns". Patients who follow the pattern (F) also follow other CP patterns in their careflow, and these patient traces are best represented by a few latent CP patterns.

4.3.2. Patient trace clustering

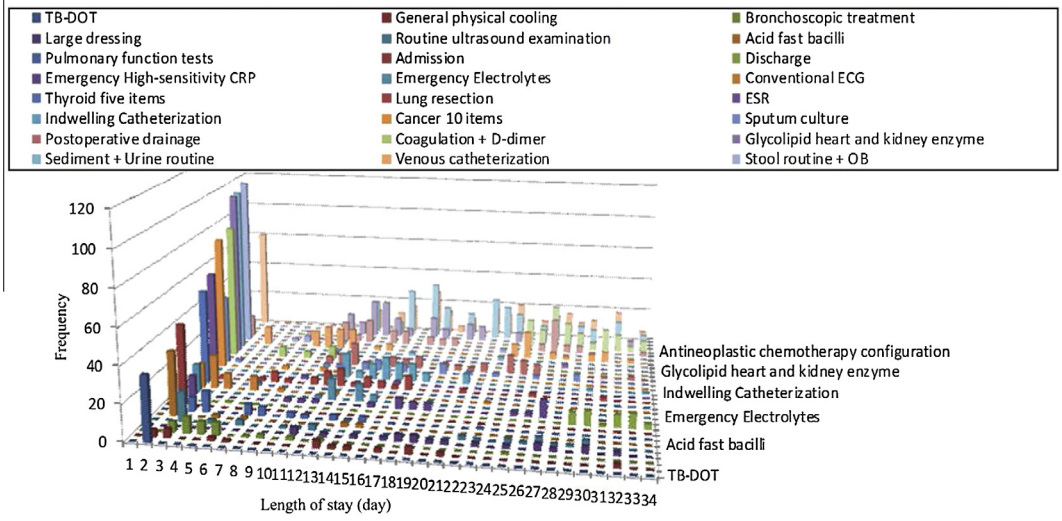
The proposed CPM provides a basis for further tasks in CPA. In this subsection, we introduce a common task of CPA, i.e., patient trace clustering, which helps reveal the underlying characteristics and commonalities among a large collection of patient traces.

A reasonable similarity measure $sim(\sigma, \sigma')$ is critical for the patient trace clustering. Once we have the learned CPM, we can measure the similarity between patient traces. In particular, for a specific trace σ in the log $\mathcal{L}$, we obtain the CP pattern distribution $\overrightarrow{\theta_\sigma} = \{\hat{\theta}_{\sigma,z_1}, \hat{\theta}_{\sigma,z_2}, \cdots, \hat{\theta}_{\sigma,z_K}\}$, where each $\hat{\theta}_{\sigma,z_i}$ is the posterior estimate of θ_{σ,z_i} for the CP pattern z_i ($1 \leqslant i \leqslant K$). Upon this, we are able to calculate the similarity between two traces σ and σ^* ($\sigma, \sigma^* \in \mathcal{L}$) as follows:

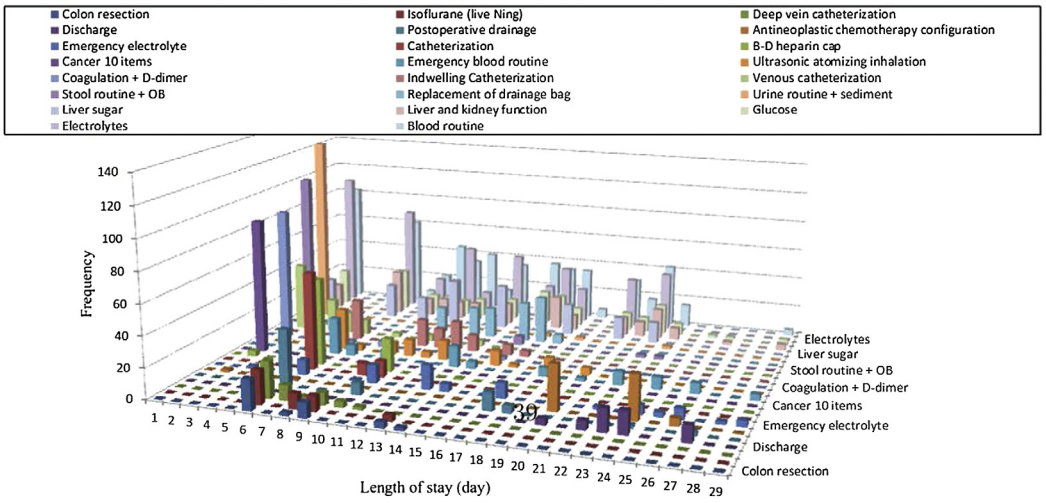
$$sim(\sigma, \sigma^*) = \frac{\sum_{i=1}^{K} \hat{\theta}_{\sigma,z_i} \times \hat{\theta}_{\sigma^*,z}}{\sqrt{\sum_{j=1}^{K} \hat{\theta}^2_{\sigma,z_j}} \sqrt{\sum_{l=1}^{K} \hat{\theta}^2_{\sigma^*,z_l}}} \tag{14}$$



(A) The discovered CP pattern indicates treatment behaviors for patients with gastric cancer



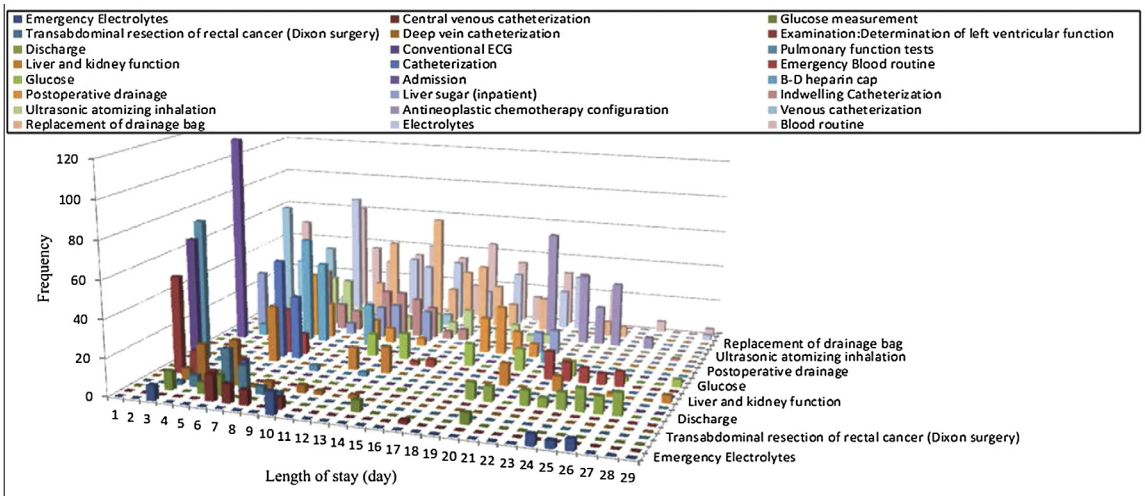
(B) The discovered CP pattern indicates treatment behaviors for patients with bronchial lung cancer



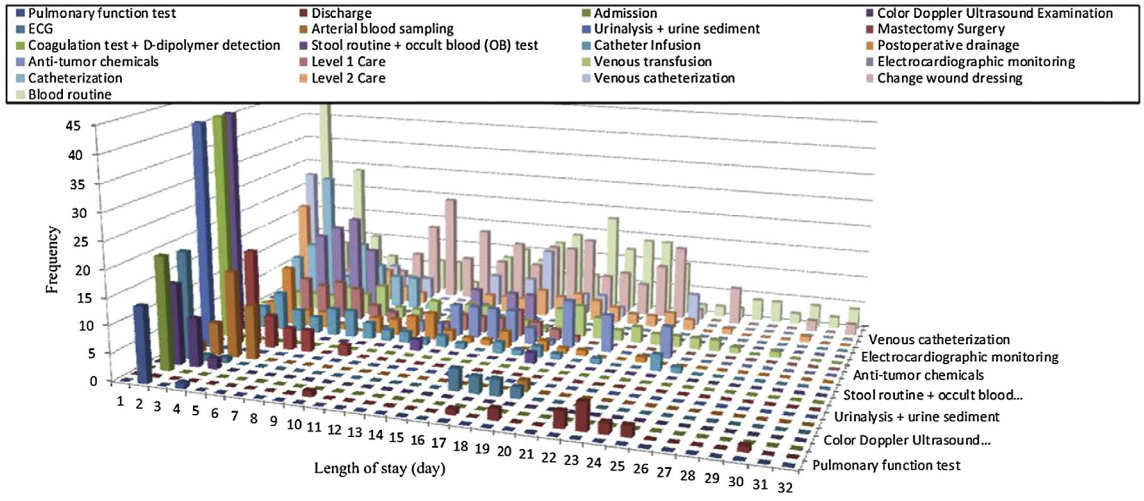
(C) The discovered CP pattern indicates treatment behaviors for patients with colon cancer

Fig. 12. The first three discovered CP patterns from the oncology log.

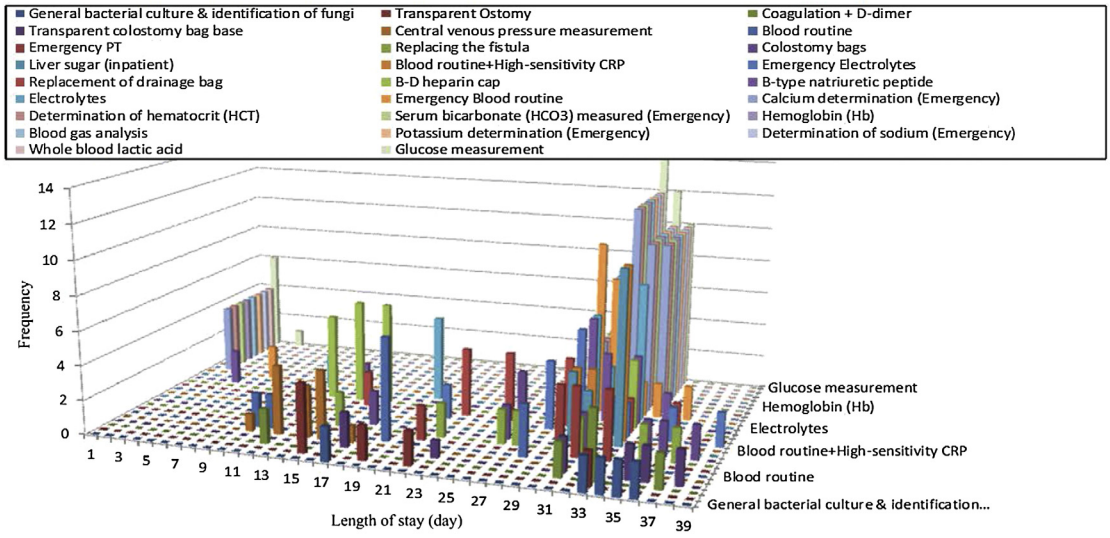
Using Eq. (14), we adopted a hierarchical micro-clustering algorithm [33] to generate partitions of patient traces in the event log. The algorithm iteratively groups two trace clusters with the largest similarity, where the similarity between two clusters is defined as the similarity between the farthest traces in the two clusters. The algorithm terminates when the maximum similarity between



(D) The discovered CP pattern indicates treatment behaviors for patients with rectal cancer



(E) The discovered CP pattern indicates treatment behaviors for patients with breast cancer



(F) The discovered CP pattern indicates variant treatment behaviors after surgery in oncological clinical pathways

Fig. 13. The next three discovered CP patterns from the oncology log.

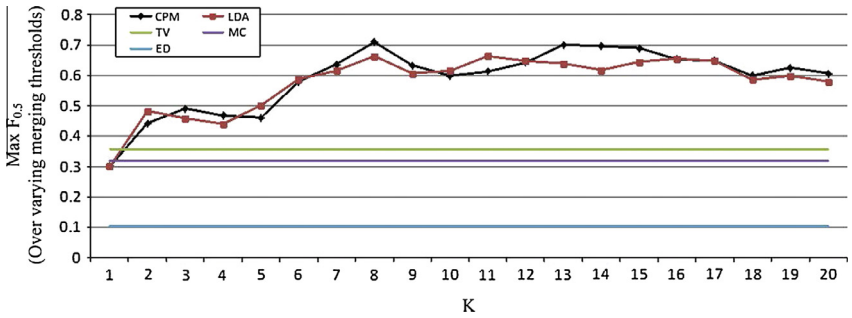


Fig. 14. Performance of clustering using ED, TV, MC, LDA, and CPM on the experimental oncology log. For MC, we set the number of clusters as 5 which is the number of benchmark categories in the log. For ED, TV, LDA, or CPM, we changed the merging threshold and obtained the maximum $F_{0.5}$ for comparison.

clusters becomes smaller than a user-specified threshold ε. The algorithm outputs a set of clusters of patient traces. It guarantees that the similarity between any pairwise traces in the same cluster is larger than ε. For more details, please refer to [33,34].

In the experiments, we compare the generated clusters with the benchmark clusters. The benchmark clusters are identified from the experimental oncology log. In particular, we use the first diagnosis code to category patient traces. As mentioned above, five categories, i.e., bronchial lung cancer, colon cancer, rectal cancer, breast cancer, and gastric cancer, are extracted from the repository.

As to evaluate the patient trace clustering, we first calculate the accuracy of the system on a per-trace basis and then build a global score for all patient traces in the log, i.e., for a patient trace σ. The precision and recall with respect to that trace are calculated as follows:

$$Precision_\sigma = \frac{|A_\sigma \bigcap B_\sigma|}{|A_\sigma|} \tag{15}$$

$$Recall_\sigma = \frac{|A_\sigma \bigcap B_\sigma|}{|B_\sigma|} \tag{16}$$

where A_σ is the generated cluster containing σ, B_σ is the benchmark cluster containing σ, $|A_\sigma \bigcap B_\sigma|$ is the number of patient traces simultaneously appeared in both A_σ and B_σ. And the final precision and recall numbers are calculated as follows:

$$Precision = \frac{1}{|\mathcal{L}|}\sum_{\sigma\in\mathcal{L}} Precision_\sigma \tag{17}$$

$$Recall = \frac{1}{|\mathcal{L}|}\sum_{\sigma\in\mathcal{L}} Recall_\sigma \tag{18}$$

Usually, precision and recall are not used separately, but combined into F_v measure as following:

$$F_v = (1+\beta^2) \times \frac{(Precision \times Recall)}{v^2 \times Precision + Recall} \tag{19}$$

In the experiments, we set $v = 0.5$ to weight precision twice as much as recall. This is because we prefer to have average-size clusters with high precision rather than merge them into a large cluster for higher recall but with low precision.

In order to evaluate the performance of clustering based on the proposed CPM, we compare the presented CPM-based similarity measure with the LDA-based similarity measure, the edit-distance-based similarity measure (ED) (which has been widely used in sequence clustering) [35], the term vector-based TFxIDF schema-employed similarity measure (TV) (being of fledged applications in text document clustering), and the first order Markov Chain-based method (MC) (which has been widely used in business process trace clustering) [16]. For CPM and LDA, we investigate the impact of K on the clustering performance where K ($K = 1,2,3,\ldots,20$) is number of CP patterns. For MC, it needs to set the number of clusters N explicitly. Since the number of benchmark categories is 5, we set the input parameter N of MC as 5 in the experiment.

Using the benchmark clusters, we can evaluate clustering performance on $F_{0.5}$. In particular, by taking the maximum value of $F_{0.5}$ (among different merging thresholds ε from 0.0 to 0.4), we compare the performance of CPM, LDA, ED, TV, and MC on the oncology log. As shown in Fig. 14, when the number of CP patterns is larger than a particular value ($K \geqslant 6$), the curves of both CPM and LDA are quite stable. Certainly, $K \approx 6$ is probably the suitable number of CP patterns for the experimental log, which is also indicated in the measure of Perplexity on the log, as shown in Fig. 11. In comparison with ED, TV and MC, the $F_{0.5}$ achieved by both the proposed CPM and traditional LDA outperform than ED, TV, and MC when K is larger than 1. For example, the $F_{0.5}$ achieved by both the proposed CPM and LDA is 0.579 and 0.587, respectively, when $K = 6$, which is a significant improvement on $F_{0.5}$ achieved by ED, TV, and MC (i.e., 0.104, 0.357 and 0.32, respectively). It indicates that the probabilistic topic models are more appropriate for trace clustering than ED, TV and MC.

4.4. Proof-of-concept prototype

We have implemented a system prototype using Microsoft C# and ASP.net, which provides web services, including upload of clinical event logs and CP pattern analysis using the logs. Both the proposed CPM and LDA have been implemented in the prototype. Some basic information of patient traces in a selected log such as patient ID, department, LOS, etc., are shown in Fig. 15B, while the screen-shot of the probability distributions of the derived CP patterns for a selected patient trace shows on Fig. 15C. The generated patterns indicate the actual treatment behaviors being applied in CPs. Fig. 15D depicts a screen-shot of the probability distributions of treatment activities and their occurring time stamps for a particular CP pattern. Users can observe the derived CP pattern from different angles by adjusting the display parameters shown on the setting panel in the bottom of Fig. 15D.

5. Discussion

The experimental studies present an analysis to the relationships between CP patterns and treatment behaviors from a statistical point of view. We conducted a comparison between the proposed CPM and LDA model. The experimental results demonstrate the effectiveness of our CPM and the potential of using treatment activities, and their occurring time stamps recorded in clinical event logs. The benefits are listed as follows:

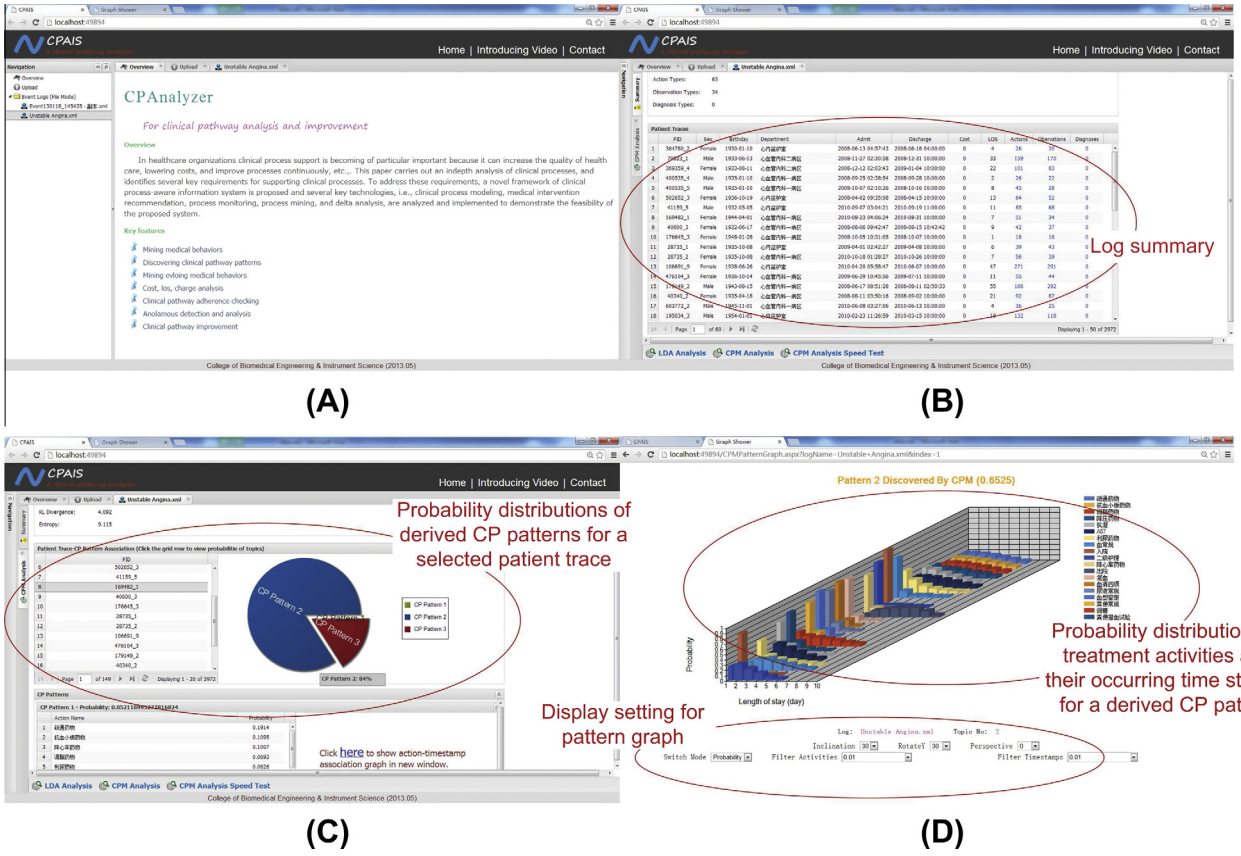


Fig. 15. A screen-shot of the system prototype.

- Discovered CP patterns, representing certain common properties, show our approach can group and classify various treatment activities according to their clinical intention. However, it requires additional information and domain-specific knowledge to find out the exact meanings of the discovered patterns. In order to discern the hidden meanings of a pattern, careful analysis and domain-specific knowledge are required, indicating that the presented method can be a powerful hypothesis generation tool to guide systemic investigation on the relationship between the discovered treatment patterns and clinical intentions.
- Discovered CP patterns, revealing essential/critical treatment behaviors from clinical event logs, form the backbone of CPs, which should be conserved, provide useful insight into the pathway, and can hence be straightforwardly included explicitly as background knowledge for further analytical objectives. As we indicated through this paper, CPs may not perform as desired according to various measures, and have to be redesigned consequently. Now that treatment behaviors in patient careflow can be recorded into clinical event logs through various kinds of hospital information systems, this can be used to verify and analyze medical services. In addition, it effectively reflects the real executing conditions in CPs. Since the constructed CP patterns gather information about the actual treatment behaviors instead of the alleged behaviors in pathways, some differences may be expected between the pathway identified by the logs and the formal description of the pathway provided by available documentation. We can, therefore, rely on them as being good representations of the structure of pathways in reality. Consequently, discovered CP patterns can be used as a feedback tool that helps in auditing and analyzing already enacted CPs, and can also provide a valuable reference for medical staff to redesign and continuously optimize CPs. From technique point of view, the methodology presented within this paper offers a "bottom-up" approach to redesign CPs, which may complement the top-down prescriptive methods. As a direct consequence, such an approach may be very useful to reduce the risks of the complex and expensive CP redesign projects.
- Although the main motivation behind our work is the forensic analysis of clinical event logs, the techniques presented herein can also be applied to business process management domain, in which event logs are commonly recorded by various information systems such that useful business process summaries can be constructed using our methodology.

It should be mentioned that there exist some limitations for the current approach:

- In this study, we exploit partial information of clinical event logs, i.e., activities and their occurring time stamps, to support clinical pathway discovery. Although our study reveals that this partial information is sufficient in discovering latent CP patterns, there are even more complex analysis and evaluation tasks that need to be considered. In the medical field, many activity decisions are based on patient health status [36], which is clearly beyond the support of data used in this study. At present, we are building a cardiovascular clinical database in the cooperation of the cardiology department of the Chinese PLA General hospital. This database contains not only many types of clinical events, but also various kinds of patient health data in their careflow, such as physiological data, pathology

examination results, patient chief complaints, and daily patient medical records, etc. It would be promising to exploit this data in order to provide healthcare organizations with nontrivial knowledge to understand how treatment behaviors are currently being performed on patients, and improve actual practices in alignment with clinical objectives in patient careflow [37,38].

- For simplicity, the time-stamp of clinical events are integer values in this study. In clinical practice, clinical event logs may record the time stamp in different granularity (i.e. some events are recorded in a date-format time-stamps, while other are recorded regarding at which hour or on which day they occur). Although we can transform the collected clinical event logs into a unified time frame, it is highly expected that the proposed CPM can analyze clinical event logs at any time granularity, to strengthen the flexibility and intelligence of CPA. For example, a clinical analyst may examine patient careflow in different time frames, where the time frames specified could be an hour, a day, or both. With richer information to disclose the hidden structure and the content of CPs in any time granularity, the proposed CPM could be more intelligent.
- In this study, perplexity is used as a measure to determine the number of CP patterns on the collected log. However, perplexity is not a "perfect" measure for model selection, since similar resulting CP patterns are not considered in the perplexity computation [39]. It can be solved by extending the proposed CPM to a nonparametric Bayesian model such as the hierarchical Dirichlet process model [40], which can automatically infer the number of patterns on unseen data. We will address it in our future work.

6. Related work

Although IT supported CP has been studied for many years [2], predominant approaches to the analysis of CPs are from an external perspective of CPs [17]. For example, Muluk et al. [41] evaluated the effects of the CP of non-urgent abdominal aortic aneurysm surgery, i.e., charges, length of stay, and mortality rate. Barbieri et al. [42] presented a meta-analysis method to evaluate the CP of hip and knee joint replacements by assessing the major outcomes of in-hospital hip and knee joint replacement processes: postoperative complications, number of patients discharged at home, length of stay, and direct cost, etc. Kul proposed a patient survival analysis for CPs [43]. As valuable as these approaches are, they typically look at aggregated data seen from the measures, e.g., length of stay, mortality, and infection rate, etc. [15], and thus restrict the attention to an external perspective of CP analysis. In clinical settings, CPs are evolving and clinicians typically have an oversimplified and incorrect view of the actual executions of CPs. In this regard, health-care organizations require to provide insights into CPs and enable various types of analysis.

In this context, process mining [21,22], as a general method in business process analysis, is gaining increasing attention in health-care [16,15,17,19]. The idea of process mining is to discover underlying business process models from event logs that record the execution information of business processes. Being transferred into medical settings, process mining methods may be applicable, for example, in retrieving frequent CP patterns from past patient traces, which might be further utilized to refine CP itself [17]. In fact, process mining has already been attempted in clinical environments by some researchers. In [15], Lin et al. reported a technique that was developed to discover the time dependency pattern of CPs for managing brain stroke. In [23], Yang et al. propose a process mining algorithm to facilitate the automatic and systematic detection of health-care fraud and abuse for CPs. In [44], Mans et al. applied process mining to discover how stroke patients are treated in different hospitals. In [16], a methodology of using process mining techniques to support health-care process analysis is thoroughly investigated. In our previous work [17], we developed a new process mining algorithm to discover a set of treatment patterns given an input event log and a minimum support threshold value, such that it can find what critical clinical activities are performed and in what order, and provide comprehensive knowledge about quantified temporal orders of clinical activities in CPs. In [19], we presented an approach to provide a concise and comprehensive summary of CP by segmenting the observed time period of the pathway into continuous and overlapping time intervals, and discovering frequent treatment behaviors in each specific time interval from a clinical event log.

However, as indicated in [45,16], the use of traditional process mining techniques though successful in discovering CP patterns can prove inadequate in CP analysis. We argue that the diversity of medical behaviors in CPs is far higher than that of common business processes. CPs, as typical human-centric processes, always take place in a loosely structured manner. The use of traditional process mining techniques may generate spaghetti-like CP patterns that are difficult to be comprehended by clinicians [16]. These incomprehensible patterns are either not amenable or lack of assistance to efforts of analysis and improvement of CPs. In many cases, the meanings or significance of discovered CP patterns sometimes goes untold using existing process mining techniques [16,46]. In fact, as CPs that deal with a variety of medical problems, it can be assumed that a patient trace is actually guided by multiple underlying treatment patterns [17,47]. For example, a patient who follows the bronchiole lung cancer CP may also be performed specific activities for his/her diabetes treatments. Even for the patients with the same disease, a slight dissimilarity of patient states may result in different treatment behaviors.

Therefore, we argue that there is a critical need to develop new techniques facilitating CP analysis. In our previous work [18], we have employed Latent Dirichlet Allocation to discover CP patterns as a probabilistic combination of clinical activities. The probability distribution derived from LDA surmises the essential features of CP patterns, and CPs can be accurately described by combining different classes of distributions. The advantages of LDA over traditional process mining approaches is: (1) soft clustering of patient traces, and (2) meaningful activity distributions as the representation of treatment patterns. This paper significantly extends our initial work by explicitly incorporating the occurring time information of treatment activities into LDA. Thus, our approach supports the discovery of temporal structure of CP patterns from clinical event logs. Specifically, the proposed CPM results in (1) probabilistic distributions of patient careflow given all treatment patterns whereas traditional process mining approaches, e.g., trace clustering, assigns only one cluster per patient trace, and (2) discriminative patient traces per pattern characterizing treatment behaviors. This information is very useful as we know the precise treatment behavior transitions which characterize treatment patterns as well as the time-stamp, giving the time details of discovered patterns. We presented various experimental results that also document an acceptable generalization by our technique. These results show that we were able to further improve the approach already presented in [18].

7. Conclusion

In this paper, we have presented a new approach for discovering latent CP patterns using probabilistic topic models, continuing from our previous work in [18]. Using a clinical event log collected by 2934 patient traces in the unstable angina CP from the cardiology department of Chinese PLA General hospital, we successfully discover underlying CP patterns using CPM proposed in this paper.

These advantages of the proposed approach have been pointed out in our proposal. In particular, our model treats CP pattern discovery from a statistical perspective. This allows better description, interpretation, understanding and characterization of treatment behaviors in patient careflow. Note that what we need is to collect a clinical event log from hospital information systems and use the collected log on mining and analyzing CPs. Analysis on the collected log is totally unsupervised. It requires small effort of humans for preprocessing patient traces in the log. This is particularly useful when dealing with CPs lacking formal consensus models, where latent CP patterns can be discovered from event logs.

Discovered CP patterns from event logs have been evaluated by hospital managers and clinical experts at the Chinese PLA General hospital, who understand the beneficial effects of discovered CP patterns. They indicate that discovered CP patterns from clinical event logs support CP (re)design and improvement. Despite that the proposed CPM is not a tool for designing CPs, it is evident that a good understanding of the existing careflow is vital for any design and improvement effort. Since event logs are reserved in most hospital information systems, the collected logs can be used to derive CP patterns explaining the events recorded. Besides, discovered patterns are not biased by perceptions, and are useful to confront with the man-made CP specifications. Thus, it might be effective in CP analysis and improvement.

The resulting distributions of clinical events for latent CP patterns reveal the hidden structure of CPs, which can profitably be exploited as a basis for further CP analysis tasks [48], including, grouping and identifying clinical events within the same therapy and treatment intention, and detecting anomalies from normal treatment behaviors, etc. Therefore, the findings provide a foundation for future research using our approach. We will address these tasks by exploiting the potential of the proposed approach, as a crucial advantage over traditional techniques for CP analysis and optimization.

Acknowledgments

This work was supported by the National Nature Science Foundation of China under Grant No 81101126, and the National Hi-Tech R&D Plan of China under Grant No 2012AA02A601.

The authors would like to give special thanks to all experts who cooperated in the evaluation of the proposed method. The authors are especially thankful for the positive support received from the cooperative hospitals as well as to all medical staff involved. And the authors would like to thank the anonymous reviewers for their constructive comments on an earlier draft of this paper.

Appendix A

In this appendix, we give the derivation of Eq. (9)

$$
\begin{aligned}
&P(\mathbf{z},\mathbf{e}|\alpha,\beta,\delta)\\
&=P(\mathbf{z},\mathbf{a},\mathbf{t}|\alpha,\beta,\delta)\\
&=P(\mathbf{z}|\alpha)P(\mathbf{a}|\mathbf{z},\beta)P(\mathbf{t}|\mathbf{z},\mathbf{a},\delta)\\
&=\int P(\mathbf{z}|\Theta)P(\theta|\alpha)d\Theta\int P(\mathbf{a}|\Phi,\mathbf{z})P(\Phi|\beta)d\Phi\int P(\mathbf{t}|\mathbf{z},\mathbf{a},\Psi)P(\Psi|\delta)d\Psi\\
&=\int\prod_{\sigma=1}^{|\mathcal{L}|}\left(\prod_{i=1}^{|\sigma|}P(z_i|\theta_\sigma)P(\theta_\sigma|\alpha)\right)d\Theta\\
&\int\prod_{\sigma=1}^{|\mathcal{L}|}\prod_{i=1}^{|\sigma|}P(e_i\cdot a|\phi_{z_i})\prod_{z=1}^{K}P(\phi_z|\beta)d\Phi\\
&\int\prod_{\sigma=1}^{|\mathcal{L}|}\prod_{i=1}^{|\sigma|}P(e_i\cdot t|\varphi_{z_i,e_i\cdot a})\prod_{z=1}^{K}\prod_{a=1}^{|A|}P(\varphi_{z,a}|\delta)d\Psi\\
&=\int\prod_{\sigma=1}^{|\mathcal{L}|}\prod_{z=1}^{K}\theta_{\sigma,z}^{m_{\sigma,z}}\prod_{\sigma=1}^{|\mathcal{L}|}\left(\frac{\Gamma\left(\sum_{z=1}^{K}\alpha_z\right)}{\prod_{z=1}^{K}\Gamma(\alpha_z)}\prod_{z=1}^{K}\theta_{\sigma,z}^{\alpha_z-1}\right)d\Theta\\
&\int\prod_{z=1}^{K}\prod_{a=1}^{|A|}\phi_{z,a}^{n_{z,a}}\prod_{z=1}^{K}\left(\frac{\Gamma\left(\sum_{a=1}^{|A|}\beta_a\right)}{\prod_{a=1}^{|A|}\Gamma(\beta_a)}\prod_{a=1}^{|A|}\phi_{z,a}^{\beta_a-1}\right)d\Phi\\
&\int\prod_{z=1}^{K}\prod_{a=1}^{|A|}\left(\prod_{t=1}^{|T|}\varphi_{z,a,t}^{q_{z,a,t}}\frac{\Gamma\left(\sum_{t=1}^{|T|}\delta_t\right)}{\prod_{t=1}^{|T|}\Gamma(\delta_t)}\prod_{t=1}^{|T|}\varphi_{z,a,t}^{\delta_t-1}\right)d\Psi\\
&\propto\prod_{\sigma=1}^{|\mathcal{L}|}\frac{\prod_{z=1}^{K}\Gamma(m_{\sigma,z}+\alpha_z)}{\Gamma\left(\sum_{z=1}^{K}(m_{\sigma,z}+\alpha_z)\right)}\prod_{z=1}^{K}\frac{\prod_{a=1}^{|A|}\Gamma(n_{z,a}+\beta_a)}{\Gamma\left(\sum_{a=1}^{|A|}(n_{z,a}+\beta_a)\right)}\\
&\prod_{k=1}^{K}\prod_{a=1}^{|A|}\frac{\prod_{t=1}^{|T|}\Gamma(q_{k,a,t}+\delta_t)}{\Gamma\left(\sum_{t=1}^{|T|}(q_{k,a,t}+\delta_t)\right)}
\end{aligned}
\tag{20}
$$

Using the chain rule and $\Gamma(\alpha)=(\alpha-1)\Gamma(\alpha-1)$, we can obtain the conditional probability conveniently,

$$
\begin{aligned}
&P(z_{\sigma,i}|z_{\sigma,-i},\mathbf{e},\alpha,\beta,\delta)\\
&=\frac{P(z_{\sigma,i},a_{\sigma,i},t_{\sigma,i}|\mathbf{a}_{\sigma,-i},\mathbf{t}_{\sigma,-i},z_{\sigma,-i},\alpha,\beta,\delta)}{P(a_{\sigma,i},t_{\sigma,i}|\mathbf{a}_{\sigma,-i},\mathbf{t}_{\sigma,-i},z_{\sigma,-i},\alpha,\beta,\delta)}\\
&\propto\frac{P(\mathbf{z},\mathbf{a},\mathbf{t}|\alpha,\beta,\delta)}{P(\mathbf{z}_{\sigma,-i},\mathbf{a}_{\sigma,-i},\mathbf{t}_{\sigma,-i}|\alpha,\beta,\delta)}\\
&\propto\frac{m_{\sigma,z}+\alpha}{m_{\sigma,*}+K\alpha}\cdot\frac{n_{z,a}+\beta}{n_{z,*}+|A|\beta}\cdot\frac{q_{z,a,t}+\delta}{q_{z,a,*}+|T|\delta}
\end{aligned}
\tag{21}
$$

References

[1] Wakamiya S, Yamauchi K. What are the standard functions of electronic clinical pathways? Int J Med Inform 2009;78(8):543–50.
[2] Lenz R, Reichert M. IT support for healthcare processes-premises, challenges, perspectives. Data Knowl Eng 2007;61(1):39–58.
[3] Dunn AG, Ong MS, Westbrook JI, Magrabi F, Coiera E, Wobcke W. A simulation framework for mapping risks in clinical processes: the case of in-patient transfers. J Am Med Inform Assoc 2011;18(3):259–66.
[4] Lu X, Huang Z, Duan H. Supporting adaptive clinical treatment processes through recommendations. Comput Methods Programs Biomed 2012;107(3):413–24.
[5] Gooch P, Roudsari A. Computerization of workflows, guidelines and care pathways: a review of implementation challenges for process-oriented health information systems. J Am Med Inform Assoc 2011;18(6):738–48.
[6] Quaglini S, Stefanelli M, Lanzola G, Caporusso V, Panzarasa S. Flexible guideline-based patient careflow systems. Artif Intell Med 2001;22(1):65–80.
[7] Lenz R, Blaser R, Beyer M, Heger O, Biber C, BAumlein M, et al. IT support for clinical pathways-lessons learned. Int J Med Inform 2007;76(3):S397–402.
[8] Hunter B, Segrott J. Re-mappling client journeys and professional identities: a review of the literature on clinical pathways. Int J Nurs Stud 2008;45:608–25.
[9] Weiland DE. Why use clinical pathways rather than practice guidelines? Am J Surg 1997;174:592–5.
[10] Uzark K. Clinical pathways for monitoring and advancing congenital heart disease care. Prog Pediatr Cardiol 2003;18:131–9.
[11] Loeb M, Carusone SC, Goeree R, Walter SD, Brazil K, Krueger P, et al. Effect of a clinical pathway to reduce hospitalizations in nursing home residents with pneumonia. J Am Med Assoc 2006;295:2503–10.
[12] Zand DJ, Brown KM, Konecki UL, Campbell JK, Salehi V, Chamberlain JM. Effectiveness of a clinical pathway for the emergency treatment of patients with inborn errors of metabolism. Pediatrics 2008;122:1191–5.
[13] Huang Z, Lu X, Duan H. Using recommendation to support adaptive clinical pathways. J Med Syst 2012;36(3):1849–60.

[14] Elson RB, Faughnan JG, Connelly DP. An industrial process view of information delivery to support clinical decision making: implications for systems design and process measures. J Am Med Inform Assoc 1997;4(4):266–78.
[15] Lin F, Chen S, Pan S, Chen Y. Mining time dependency patterns in clinical pathways. Int J Med Inform 2001;62(1):11–25.
[16] Rebuge A, Ferreira DR. Business process analysis in healthcare environments: a methodology based on process mining. Inform Syst 2012;37(2):99–116.
[17] Huang Z, Lu X, Duan H. On mining clinical pathway patterns from medical behaviors. Artif Intell Med 2012;56(1):35–50.
[18] Huang Z, Lu X, Duan H. Latent treatment topic discovery for clinical pathways. J Med Syst 2013;37(2):1–10.
[19] Huang Z, Lu X, Duan H, Fan W. Summarizing clinical pathways from event logs. J Biomed Inform 2013;46(1):111–27.
[20] van de Klundert J, Gorissen P, Zeemering S. Measuring clinical pathway adherence. J Biomed Inform 2010;43(6):861–72.
[21] Agrawal R, Gunopulos D, Leymann F. Mining process models from workflow logs. In: Schek HJ, Saltor F, Ramos I, Alonso G, editors. Sixth international conference on extending database technology. London: Springer-Verlag; 1998. p. 469–83.
[22] Cook JE, Wolf AL. Discovering models of software processes from event-based data. ACM Trans Software Eng Methodol 1998;7(3):215–49.
[23] Yang W, Hwang S. A process-mining framework for the detection of healthcare fraud and abuse. Expert Syst Appl 2006;31(1):56–68.
[24] van der Aalst WMP, Weijters AJMM, Maruster L. Workflow mining: discovering process models from event logs. IEEE Trans Knowl Data Eng 2004;16(9):1128–42.
[25] Huang Z, Lu X, Gan C, Duan H. Variation prediction in clinical processes. In: Peleg M, Lavrac N, Combi C, editors. Artificial intelligence in medicine. Lecture notes in computer science, vol. 6747. Berlin (Heidelberg): Springer; 2011. p. 286–95.
[26] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res 2003;3:993–1022.
[27] Newman D, Asuncion A, Smyth P, Welling M. Distributed algorithms for topic models. J Mach Learn Res 2009;10:1801–28.
[28] Griffiths TL. Finding scientific topics. Proc Nat Acad Sci 2004;101:5228–35.
[29] Wang X, McCallum A, Wei X. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In IEEE international conference on data mining; 2007. p. 697–702.
[30] Rosen-Zvi M, Griffiths T, Steyvers M, Smyth P. The author-topic model for authors and documents. In: 20th Conference on uncertainty in artificial intelligence; 2004. p. 487–94.
[31] 2012 Writing Committee Members, Jneid Hani, Anderson Jeffrey L, Scott Wright R, Adams Cynthia D, Bridges Charles R, et al. ACCF/AHA focused update of the guideline for the management of patients with unstable Angina/non-ST-elevation myocardial infarction (updating the 2007 guideline and replacing the 2011 focused update). Circulation 2012;126(7):875–910.
[32] Phung D, Adams B, Venkatesh S, Kumar M. Unsupervised context detection using wireless signals. Pervasive Mobile Comput 2009;5(6):714–33.
[33] Ertoz L, Steinbach M, Kumar V. Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In: Third SIAM international conference on data mining (SDM); 2003 p. 47–58.
[34] Huang Z, Lu X, Duan H. Similarity measuring between patient traces for clinical pathway analysis. In: Peek Niels, Morales Roque Marn, Peleg Mor, editors. Artificial intelligence in medicine. Lecture notes in computer science, vol. 7885. Berlin (Heidelberg): Springer; 2013. p. 268–72.
[35] Gusfield D. Algorithms on strings trees, and sequences. Computer science and computational biology. Cambridge University; 1997.
[36] Kaymak U, Mans R, van de Steeg T, Dierks M. On process mining in health care. In: IEEE international conference on systems, man, and cybernetics (SMC); 2012. p. 1859–64.
[37] Peleg M, Soffer P, Ghattas J. Mining process execution and outcomes position paper. In: Hofstede Arthur, Benatallah Boualem, Paik Hye-Young, editors. Business process management workshops. Lecture notes in computer science, vol. 4928. Berlin (Heidelberg): Springer; 2008. p. 395–400.
[38] Ghattas J, Peleg M, Soffer P, Denekamp Y. Learning the context of a clinical process. In: Rinderle-Ma Stefanie, Sadiq Shazia, Leymann Frank, editors. Business process management workshops. Lecture notes in business information processing, vol. 43. Berlin (Heidelberg): Springer; 2010. p. 545–56.
[39] Farrahi K, Gatica-Perez D. Discovering routines from large-scale human locations using probabilistic topic models. ACM Trans Intell Syst Technol 2011;2(1):3:1–3:27.
[40] Teh YW, Jordan MI, Beal MJ, Blei DM. Hierarchical Dirichlet processes. J Am Statist Assoc 2004;101(476):1566–81.
[41] Muluk SC, Painter L, Sile S, Rhee RY, Makaroun MS, Steed DL, et al. Utility of clinical pathway and prospective case management to achieve cost and hospital stay reduction for aortic aneurysm surgery at a tertiary care hospital. J Vasc Surg 1997;25(1):84–93.
[42] Barbieri A, Vanhaecht K, Van Herck P, Sermeus W, Faggiano F, Marchisio S, et al. Effects of clinical pathways in the joint replacement: a meta-analysis. BMC Med 2009;7(32):1–11.
[43] Kul S. The use of survival analysis for clinical pathways. Int J Care Pathways 2010;14(1):23–6.
[44] Mans R, Schonenberg H, Leonardi G, Panzarasa S, Cavallini A, Quaglini S, et al. Process mining techniques: an application to stroke care. Studies Health Technol Inform 2008;136:573–8.
[45] Lang M, Urkle TB, Laumann S, Prokosch H-U, Process mining for clinical workflows: challenges and current limitations. In: Andersen SK, Klein GO, Schulz S, Aarts J, editors, Proceedings of MIE2008 The XXIst international congress of the European Federation for medical informatics; 2008. p. 229–34.
[46] Peleg M, Mulyar N, van der Aalst WMP. Pattern-based analysis of computer-interpretable guidelines: don't forget the context. Artif Intell Med 2012;54(1):73–4.
[47] Goedertier S, De Weerdt J, Martens D, Vanthienen J, Baesens B. Process discovery in event logs: an application in the telecom industry. Appl Soft Comput 2011;11(2):1697–710.
[48] Huang Z, Dong W, Duan H, Li H. Similarity measure between patient traces for clinical pathway analysis: problem, method, and applications, Biomedical and Health Informatics, IEEE Journal of , vol.PP, no.99, p.1,1, 0 doi: 10.1109/JBHI.2013.2274281.