Data Min Knowl Disc (2015) 29:914–949
DOI 10.1007/s10618-014-0381-y

On mining latent treatment patterns from electronic medical records

Zhengxing Huang · Wei Dong · Peter Bath · Lei Ji · Huilong Duan

Received: 28 February 2014 / Accepted: 21 August 2014 / Published online: 24 September 2014

Abstract Clinical pathway (CP) analysis plays an important role in health-care management in ensuring specialized, standardized, normalized and sophisticated therapy procedures for individual patients. Recently, with the rapid development of hospital information systems, a large volume of electronic medical records (EMRs) has been produced, which provides a comprehensive source for CP analysis. In this paper, we are concerned with the problem of utilizing the heterogeneous EMRs to assist CP analysis and improvement. More specifically, we develop a probabilistic topic model to link patient features and treatment behaviors together to mine treatment patterns hidden in

Responsible editors: Fei Wang, Gregor Stiglic, Ian Davidson and Zoran Obradovic.

Zhengxing Huang and Wei Dong contributed equally to this work.

Z. Huang (✉)· H. Duan
The Key Laboratory of Biomedical Engineering, Ministry of Education, College of Biomedical Engineering and Instrument Science of Zhejiang University, Hangzhou, China
e-mail: zhengxing.h@gmail.com

H. Duan
e-mail: duanhl@zju.edu.cn

W. Dong
Department of Cardiology, Chinese PLA General Hospital, Beijing, China
e-mail: 301dongwei@sina.com

P. Bath
Information School, University of Sheffield, Sheffield, UK
e-mail: p.a.bath@sheffield.ac.uk

L. Ji
IT Department, Chinese PLA General Hospital, Beijing, China
e-mail: jilei_nudt@163.com

EMRs. Discovered treatment patterns, as actionable knowledge representing the best practice for most patients in most time of their treatment processes, form the backbone of CPs, and can be exploited to help physicians better understand their specialty and learn from previous experiences for CP analysis and improvement. Experimental results on a real collection of 985 EMRs collected from a Chinese hospital show that the proposed approach can effectively identify meaningful treatment patterns from EMRs.

Keywords Clinical pathway analysis · Probabilistic topic models · Latent Dirichlet allocation · Pattern discovery · Electronic medical records

1 Introduction

Clinical pathways (CPs) play an important role in clinical environments by delineating the optimal multidisciplinary treatment processes performed by a team of health-care professionals to achieve a specific treatment objective for a particular diagnosis or procedure (Cheah 2000; Lu et al. 2012; Gooch and Roudsari 2011; Yao and Kumar 2013). Different from common business processes in commercial and industry environments, treatment processes are highly dynamic, context sensitive, event driven, and knowledge intensive such that they often bear no relation to the ideal as envisaged by the designers of CPs (Huang et al. 2012). Thus, CPs must be improved continuously (Lu et al. 2012; Huang et al. 2013a, b, 2014). In this regard, CP analysis is vital for health-care management due to its usefulness in capturing the actionable knowledge to administrate, automate, and schedule the best practice for individual patients in the execution of CPs (Huang et al. 2012, 2013a, b, 2014).

Regarding CP analysis, one of the most important aspects is to infer essential/critical treatment behaviors conditioned on individual patient information for a particular pathway (Huang et al. 2012). We might be interested to know, for example, what treatment behaviors should be performed for an unstable angina patient with renal insufficiency. Should Percutaneous Coronary Intervention (PCI) be performed for patients with low-risk levels in the unstable angina pathway? etc. To ensure best practices in CPs, there are two challenges that need to be addressed:

- What are typical patient conditions that commonly exist for patients who follow a particular CP? and
- Which treatment behaviors are most appropriate, given the patient's specific conditions and the execution of the pathway?

In this study, we represent patient conditions as latent categories of patient features that are likely to describe homogeneous patient statuses for a particular CP, and represent treatment behaviors as flexible, transparent, and re-usable pieces of functionality that consist of one or several treatment activities required to set up a clinical solution given specific patient conditions. Furthermore, we represent treatment patterns as a composition of both patient conditions and treatment behaviors, which are underlying in CPs, and form the backbone of CPs.

Note that a patient might have complex clinical condition in the pathway due to complications, comorbidity or infections, etc., which in turn leads to various treatment

behaviors occurring during the pathway execution, and requires the pathway to be a mixture of latent treatment patterns. As a result, treatment patterns are composed of multidisciplinary treatment activities, and the composition of activities has a large variability depending on patient conditions (Rebuge and Ferreira 2012).

Previously, it was difficult to perform CP analysis due to a lack of data that were sufficiently clinical. Recently, with the rapid development of hospital information systems, a large collection of electronic medical records (EMRs) has become available, which provides the opportunity to study medical cases, evidence and knowledge for CP analysis. In particular, various types of patient information (e.g., symptoms, vital signs, lab test results, etc.) and treatment interventions (e.g., medication, surgery, examination, etc.) are recorded in EMRs, which conceal an untapped reservoir of knowledge about particular treatments and the way these are applied for patients with specific conditions. It is therefore possible to mine EMRs, extract non-trivial treatment patterns from EMRs, and exploit these for helping physicians improve and optimize CPs, and make the practice better for the care of individual patients (Peleg 2013).

The key objective in this study is to investigate how to discover latent treatment patterns from the abundant EMRs for the purpose of CP analysis and improvement. We see that an EMR contains heterogeneous clinical information for a patient (e.g., the patient demographics, laboratory tests results, radiological examination reports, etc.) and various treatment interventions (e.g., medications, surgeries, examinations, care activities, etc.) performed during the pathway execution. The different aspects of these types of information are highly correlated and physicians are very interested in the association patterns. While physicians may work through these problems based on their experiences, the emerging EMRs provide us with unique opportunities to learn from previous cases. Nontrivial knowledge can be extracted through mining underlying treatment patterns from a large volume of EMRs, and can be profitably exploited as a basis for further applications, including CP (re)design and optimization, clinical decision support, efficient medical service delivery and business management, etc (Cheah 2000; Lenz et al. 2007; Huang et al. 2012).

In this study, we propose to leverage the power of probabilistic topic models (1) to extract latent treatment patterns from EMRs, and (2) to enable the recognition of patient treatment processes as a composition of such patterns. Discovered treatment patterns consist of both prerequisite patient clinical information and subsequent treatment interventions, and their associations. The proposed approach allows us to infer a number of conditional probabilities for partial patient instances of CPs, e.g., the probabilities of particular treatment behaviors given the observations on patient conditions. The probability distribution derived from the proposed models can surmise the essential features of CPs such that the outcomes of our study can be potentially valuable for CP analysis and redesign. The proposed approach is evaluated on a collection containing 985 EMRs of unstable angina patients collected from the cardiology department of the Chinese PLA general hospital. As well a possible clinical application in terms of treatment recommendation in CPs is given to illustrate the effectiveness of the proposed approach.

The remainder of the paper is organized as follows. Section 2 outlines related work. Section 3 presents preliminary knowledge of the proposed approach. Section 4 describes the proposed approach for discovering underlying treatment patterns from

EMRs. Section 5 presents our experimental results. Section 6 provides an outlook on how the presented approach contributes to CP analysis and improvement. Finally, some conclusions are given in Sect. 7.

2 Related work

CPs have been recognized as a tool to 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 (Quaglini et al. 2001; Lenz and Reichert 2007; Lenz et al. 2007; Hunter and Segrott 2008; Weiland 1997; Uzark 2003; Loeb et al. 2006; Zand et al. 2008; Huang et al. 2012). 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 CPs for patients, and these pathways must be improved continuously (Wakamiya and Yamauchi 2009; Lenz and Reichert 2007; Dunn et al. 2011; Lu et al. 2012; Gooch and Roudsari 2011). 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 health-care 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 (Elson et al. 1997; Lin et al. 2001; Rebuge and Ferreira 2012; Huang et al. 2012, 2013a, 2014).

CP analysis is receiving increasing attention in the field of health-care management (Huang et al. 2013a, b). There have been a large number of studies on CP analysis. For instance, Renholm et al.'s (2002) review of the literature suggested that the use of CPs reduces the cost of care and the length of stay (LOS), and has a positive impact on outcomes (e.g., increased quality of care and patient satisfaction, improved continuity of information, and patient education, etc.). Thomas et al. (2008) performed a systematic review on the effect of using CPs on LOS , hospital costs and patient outcomes. Dy et al. (2005) analyzed 26 surgical critical pathways in a tertiary care center, and their study indicated that CPs' effectiveness in reducing LOS tends to be for procedures with lower patient severity of illness. As valuable as these work are, many of them analyze CPs from an external perspective of CPs, e.g., LOS, mortality, and infection rate, etc (Lin et al. 2001). As a matter of fact, CPs are evolving with the rapid development of medical technologies. While health-care organizations typically have an simplified and incorrect view of the actual situations in CPs (Lu et al. 2012; Huang et al. 2012, 2011), efficient analysis should help them keep tracing essential/critical treatment behaviors of the pathway, and extract potential information, which may substantially improve CPs (Peleg et al. 2012). In this regard, it is necessary to provide insights into CPs at a very refined level (Huang et al. 2012; Rebuge and Ferreira 2012).

There has been work from the business process management community on analyzing CPs (Rebuge and Ferreira 2012; Lenz and Reichert 2007). In particular, process mining (Agrawal et al. 1998; Cook and Wolf 1998), as a general method in business process analysis, has been used to analyze CPs (Huang et al. 2012, 2013a,

2014; Rebuge and Ferreira 2012; Lin et al. 2001). Process mining techniques allow a more "intelligent" kind of analysis by executing data mining algorithms on the warehouse data in order to automatically construct process models explaining the behavior observed from the data. Shifting to the clinical environment, process mining techniques can measure treatment behavior from EMRs, which regularly record CPs' execution information. Note that, with a rigorous mathematical logic and reasoning ability, process mining can be an objective way of analyzing CPs (Rebuge and Ferreira 2012). For instance, Lang et al. (2008) presented their study on the process mining based analysis of the radiology workflow at the clinic of traditional Business Process Analysis (BPA) in the Erlangen University Clinic, in Germany. Mans et al., applied process mining to discover how stroke patients are treated in patient careflow (Mans et al. 2008). Rebuge and Ferreira (2012) presented a systematic methodology for using process mining techniques to support health-care process analysis. Bouarfa and Dankelman proposed a process mining algorithm to derive a consensus model from multiple clinical activity logs, based on which CP outliers can be detected automatically and without prior knowledge from clinical experts (Bouarfa and Dankelman 2012). Work that is closely related to ours is presented in 2013, in which Lakshmanan et al., present a hybrid approach for mining CPs correlated with patient outcomes that involves a combination of clustering, process mining and frequent pattern mining. In particular, their work takes clinical outcome into account in the mining process, and thus could facilitate the improvement of existing CPs. However, the patient-specific information is not included in CP mining in their study.

In our previous work (Huang et al. 2012), we developed a new process mining algorithm to discover treatment behavior patterns from clinical data such that it can reveal what critical clinical activities are performed and in what order, and provide comprehensive knowledge about quantified temporal orders of clinical activities in CPs. In Huang et al. (2013b), 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 discovered frequent treatment behaviors in each specific time interval from a clinical event log. In Huang et al. (2012), we proposed a CP-anomaly detection model to classify a particular patient CP instance to one of the various patient instance clusters of CPs such that anomalous treatment behaviors can be detected in a timely manner and explained clearly with respect to its membership cluster in a maximally-informative manner.

In clinical settings, the complexity and diversity of treatment behaviors in CPs are far higher than that of common business processes (Rebuge and Ferreira 2012; Lenz and Reichert 2007). Although most process mining algorithms can discover business process models in a structured manner (Huang et al. 2013a, 2014), the assumption that the processes take place in a structured fashion is not valid for CPs. CPs are typical human-centric processes, and always take place in an unstructured fashion. In clinical practice, many treatment behaviors can occur arbitrarily without a particular order. Bringing order to the chaos of CPs probably requires a different mining strategy rather than existing process mining algorithms.

To this end, we previously employed Latent Dirichlet Allocation (LDA) to discover treatment patterns as a probabilistic combination of treatment activities (Huang et al.

2013a, 2014). The probability distribution derived from LDA surmises the essential features of treatment patterns, and CPs can be accurately described by combining different classes of distributions such that similarities of pairwise patient instances of CP can be measured based on their underlying behavioral topical features. This also provides a basis for further applications in CP analysis, e.g., patient instance retrieval, clustering, and anomaly detection, etc (Huang et al. 2014). In this paper, we significantly extend our initial work by explicitly incorporating patient-related information into probabilistic topic models to model underlying treatment patterns, which is quite beneficial for CP analysis.

It is worth noting that physicians always refer to patient conditions to work out a patient's status in clinical practice. They then give their diagnosis, on which they base their treatments for the patient during the patient's CP. In general, patients who have the same conditions will be treated with similar treatment behaviors, i.e., patients are given some common medications, medical procedures, etc., while taking some account of known individual specific details, e.g., drug allergies. Using probabilistic topic models, not only can prerequisite patient conditions and subsequent treatment behaviors be discovered from the heterogeneous medical records, but also their associations can be revealed, and thus it has the potential to help physicians learn from previous cases and assist CP analysis and improvement.

3 Preliminaries

Here we introduce our notations and terminologies for the proposed approach. Formally, let $\mathcal{D}$ be a collection of EMRs. Each EMR d in $\mathcal{D}$, corresponding to a particular patient, consists of the descriptions on both the patient clinical information and treatments performed on the patient given his/her clinical conditions and during the pathway execution.

In general, CPs are a mixture of multiple treatment patterns $\mathcal{Z}$. A treatment pattern z ($z \in \mathcal{Z}$) is a latent unobservable variable represented as a multinomial distribution over a set of clinical words W ($W = \{\langle F, V\rangle\} \bigcup E$), which consist of a set of pairs of patient features F and their values V, and a set of clinical events E performed on patients in CPs. Patient features are observable variables, including patient demographics, lab tests results, vital signs, etc., which can have categorical or numerical values. Intuitively, patient conditions are described as a set of pairs of patient features and their values $\langle \mathbf{f}_d, \mathbf{v}_d \rangle$, which are the most important issue physicians aim to figure out for treatments during CPs. Here, $\mathbf{f}_d = \{f_{di}\}_{i=1}^{N_d^f}$ represents patient features that are measured on a particular patient and recorded in patient medical record d, $\mathbf{v}_d = \{v_{di}\}_{i=1}^{N_d^f}$ represents their values, and N_d^f is the number of patient features recorded in d.

After observing the patient's condition, physicians will make a diagnosis and perform the corresponding treatment behaviors on the patient. In this study, we assume that treatment behaviors are a set of clinical events performed on a particular patient during the execution of his/her CP. These clinical events are recorded in the patient's EMR d. Each clinical event is a pair of treatment activity and its occurring time stamp

$e = \langle a, t \rangle$.[1] Here $\mathbf{e}_d = \{e_{dj}\}_{j=1}^{N_d^e}$ represents clinical events performed on a particular patient and recorded in d, $\mathbf{a}_d = \{a_{dj}\}_{j=1}^{N_d^e}$ represents activity types of clinical events $\mathbf{e}_d$, $\mathbf{t}_d = \{t_{dj}\}_{j=1}^{N_d^e}$ represents the occurring time stamps of clinical events $\mathbf{e}_d$, and N_d^e is the number of clinical events in d. Our notation is summarized in Table 1.

4 Method

In this study, we propose a novel approach for extracting latent treatment patterns from EMRs. In particular, we extend our previous work to propose a new model called treatment pattern model (TPM). Figure 1b shows the graphical model for mining latent treatment patterns from EMRs. The graphical model represents dependencies among variables. The shaded and unshaded nodes indicate observed and latent variables respectively. The plate indicates replicates, and the value in the plate indicates the number of replicates. For comparison, we represent a graphical model of the standard topic model, i.e., latent Dirichlet allocation in Fig. 1a.

In Fig. 1b, the proposed TPM can figure out patient conditions that are influenced by both patient features, e.g., patient symptoms, lab tests results, and vital signs, etc., and their values. In general, patient conditions are latent and unobservable. Even for patients with the same disease, they may have different conditions. For example, patients whose first diagnosis is unstable angina may have different conditions due to their individual risk levels (i.e., low-risk, medium-risk, and high-risk scores, etc.).

The generative process for measured patient features is the same as that of standard topic models. Each EMR d in $\mathcal{D}$ has treatment pattern proportions θ_d that are sampled from a Dirichlet distribution with prior α. θ_d stands for the probability of assigning a treatment pattern z to a clinical word generated from patient medical record d. As we mentioned above, clinical words are either pairs of patient features and their values, or clinical events. For each of the N_d^e clinical event, the generative process is similar to the approach proposed in our previous work (Huang et al. 2014, 2013a), i.e., a treatment pattern z is associated with a multinomial distribution ϕ_z over the activity type a of a clinical event e, and a multinomial distribution $\xi_{z,a}$ over the occurring time stamp of e for pattern z.[2] For each of the N_d^f patient features, a treatment pattern z is chosen from the treatment pattern proportions, and then patient feature f is sampled from a treatment pattern-specific multinomial distribution ψ_z.

[1] Note that clinical events could be characterized by various properties, e.g., an event has an occurring time stamp, it corresponds to a treatment 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.

[2] Note that CPs, as standardized inpatient treatment processes, are executed in specific time periods from admission to discharge. During the execution of CPs, treatments of a pathway should be performed in specific time instants. Taking the unstable angina pathway as an example, typical treatment activities, such as lab test, ECG examination, etc., have to be performed in the first days after admission, and PCI surgery have to be performed subsequently, etc. Thus, the occurring time stamps are partially determined by the treatment activities.

Table 1 Notation

Symbol	Description	
$\mathcal{D}$	A collection of patient medical records	Set
$\mathcal{Z}$	Latent treatment patterns	Set
Z	Number of latent treatment patterns	Scalar
D	Number of a collection of patient medical records	Scalar
N	Number of words (i.e., pairs of both patient features and their values, and treatment activities and their occurring time stamps) in $\mathcal{D}$	Scalar
N_d	Number of words in the dth patient medical record, $N_d = N_d^f + N_d^a$	Scalar
N_d^f	Number of pairs of patient features and their values in the dth patient medical record	Scalar
N_d^e	Number of clinical events in the dth patient medical record	Scalar
F	Vocabulary size of patient features in $\mathcal{D}$	Scalar
V	Vocabulary size of values of patient feature in $\mathcal{D}$	Scalar
A	Vocabulary size of treatment activities in $\mathcal{D}$	Scalar
T	Vocabulary size of time stamps in $\mathcal{D}$	Scalar
$V_{z,f}$	Vocabulary size of values of patient feature f given treatment pattern z	Set
$e = \langle a, t \rangle$	A clinical event is a pair of a treatment activity and its occurring time stamp	Pair
$\mathbf{e}_d$	Clinical events in the dth patient medical record	N_d^a-dimensional vector
$\mathbf{a}_d$	Treatment activities in the dth patient medical record	N_d^a-dimensional vector
$\mathbf{t}_d$	The occurring time stamps of treatment activities in the dth patient medical record	N_d^a-dimensional vector
$e_{d,j}$	ith clinical event in the dth patient medical record	jth component of $\mathbf{e}_d$
$\mathbf{V}_{z,f}$	Values of patient feature f that is assigned to treatment pattern z	Set
$\mathbf{z}$	Treatment pattern assignments	N-dimensional vector
$\mathbf{f}_d$	patient features in the dth patient medical record	N_d^f-dimensional vector
$\mathbf{v}_d$	Values of patient features in the dth patient medical record	N_d^f-dimensional vector
$f_{d,i}$	ith patient feature in the dth patient medical record	ith component of $\mathbf{f}_d$
$z_{d,j}$	Treatment pattern assignment for patient feature $f_{d,i}$	ith component of $\mathbf{p}_d$
α	Dirichlet prior	Scalar
β	Dirichlet prior	Scalar
γ	Dirichlet prior	Scalar
η	Dirichlet prior	Scalar
ι	Dirichlet prior	Scalar
x	Shape parameter of Gamma distribution	Scalar

Table 1 continued

Symbol	Description	
y	Inverse scale parameter of Gamma distribution	Scalar
g	Mean parameter of Gaussian distribution	Scalar
k	precision parameter of Gaussian distribution	Scalar
μ_z	Mean value of normal distribution prior of numerical patient features given treatment pattern z	Scalar
λ_z	Precision of normal distribution prior of numerical patient features given treatment pattern z	Scalar
Φ	Probabilities of treatment activities given treatment patterns	$A \times Z$ matrix
ϕ_z	Probabilities of treatment activities given treatment pattern z	Z-dimensional vector
Ψ	Probabilities of patient features given treatment patterns	$F \times Z$ matrix
ψ_z	Probabilities of patient features given treatment pattern z	Z-dimensional vector
Δ	Probabilities of values of categorical patient features given treatment patterns and patient features	$V \times Z \times F$ matrix
$\delta_{z,f}$	Probabilities of values of categorical patient features given treatment pattern z and patient feature f	$Z \times F$ matrix
Ξ	Probabilities of occurring time stamps given treatment patterns and clinical activities	$T \times Z \times A$ matrix
$\xi_{z,a}$	Probabilities of occurring time stamps given treatment pattern z and treatment activity a	$Z \times A$ matrix
Θ	Probabilities of treatment patterns given patient medical records	$Z \times D$ matrix
θ_d	Probabilities of treatment patterns given the dth patient medical record	D-dimensional vector

Note that there are two kinds of patient features, i.e., categorical features and numerical features. For a categorical patient feature $f_{d,i}$, its value, denoted as $v_{d,i}$, is generated from the distribution $\delta_{z_{d,i},f_{d,i}}$. There are, in total, $Z \times F$ prior distributions of patient feature-value pairs, which follow a Dirichlet distribution with prior ι. For numerical patient features, each treatment pattern has its own value distribution for each numerical patient feature, which is assumed to be a Normal distribution, Normal$(\mu_{z,i}, \lambda_{z,i}^{-1})$. To this end, we adopted an approach proposed in Iwata and Sawada (2013) to process numerical patient features. In particular, we set different means and variations of a feature's value depending on the treatment pattern z, and then we can analyze the value of a numerical patient feature and its range that are specific to z. The mean $\mu_{z,i}$ and the precision $\lambda_{z,i}$ are sampled respectively from Normal distributions, which are conjugate priors μ_0 and λ_0 of a Normal distribution. After treatment pattern $z_{d,i}$ and patient feature $f_{d,i}$ are sampled, its value $v_{d,i}$ is determined using Normal$(\mu_{z_{d,i},f_{d,i}}, \lambda_{z_{d,i},f_{d,i}}^{-1})$.

In summary, the proposed TPM assumes the following generative process for the collection of EMRs $\mathcal{D}$:

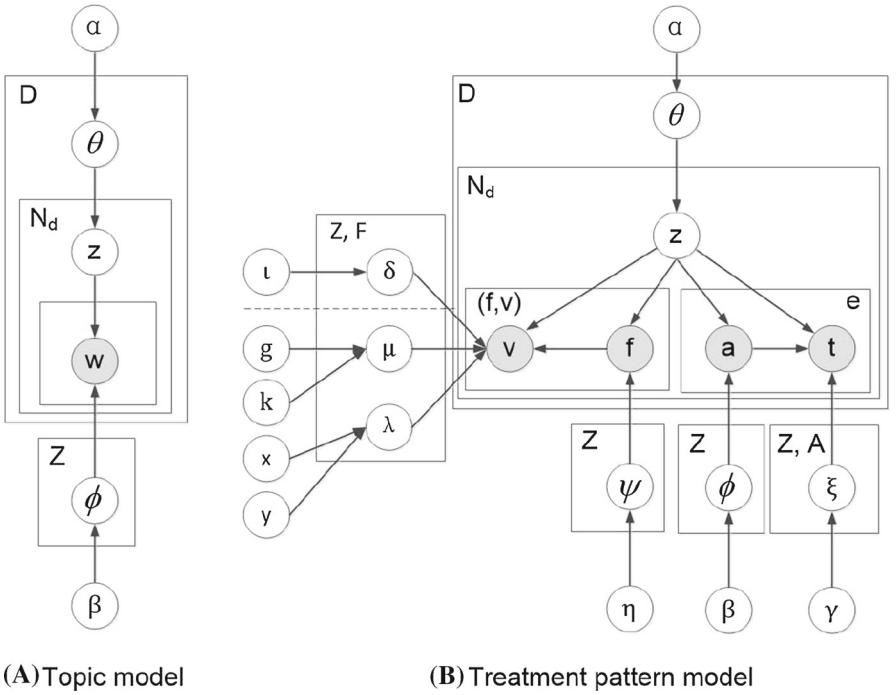

Fig. 1 Graphical model representation of the proposed topic models for mining latent treatment patterns. Z, F, and A denote variables "treatment patterns", "patient features", and "treatment activities" respectively. θ, ϕ, ξ, φ, δ, μ, and λ are distributions of treatment patterns over EMRs, treatment activities over patterns, the occurring time stamps of treatment activities over patterns, patient features over treatment patterns, the values of categorical features over treatment patterns, the mean values of numerical features over treatment patterns, and the precisions of numerical features over treatment patterns, respectively. w represents clinical words which are either a pair of patient feature f and its value v or a clinical event e (a pair of treatment activity a and its occurring time stamp t). N_d represents the number of the occurrences of clinical words for one EMR d. D represents the number of EMRs in the collected data set. The hyperparameters α, β, γ, η, ι, g, k, x, y control dispersions of treatment patterns, activities and their occurring time stamps, patient features, values for categorical features, values' means for numerical features, and values' precisions for numerical features, respectively

1. For each treatment pattern $z = 1, \ldots, Z$, draw patient feature probability $\psi_z \sim$ Dirichlet(η), draw treatment activity probability $\phi_z \sim$ Dirichlet(β);
2. For each clinical word w and each treatment pattern z
   (a) If w is a categorical feature f, draw patient feature's value probability $\delta_{z,f} \sim$ Dirichlet(ι)
   (b) Else if f is a numerical feature
      i. Draw patient feature's value precision $\lambda_{z,f} \sim$ Gamma(x_f, y_f)
      ii. Draw patient feature's value mean $\mu_{z,f} \sim$ Normal($g_f, (k_f \lambda_{z,f})^{-1}$)
   (c) Else if w is a clinical event e, draw activity occurring time stamp probability $\varphi_{z,e.a} \sim$ Dirichlet(δ)
3. For each patient medical record $d = 1, \ldots, D$:
   (a) Draw treatment pattern proportions $\theta_d \sim$ Dirichlet(α)

(b) For each patient feature-value pair $i = 1, \ldots, N_d^f$ in d:
  i. Draw treatment pattern $z_{d,i} \sim \text{Multinomial}(\theta_d)$
  ii. Draw patient feature $f_{d,i} \sim \text{Multinomial}(\psi_{z_{d,i}})$
    A. If $f_{d,i}$ is a categorical patient feature, draw patient feature's value $v_{d,i} \sim \text{Multinomial}(\delta_{z_{d,i},f_{d,i}})$, else
    B. If $f_{d,i}$ is a numerical patient feature, draw patient feature's value $v_{d,i} \sim \text{Normal}(\mu_{d,i}, \lambda_{d,i}^{-1})$

(c) For each clinical event e_i, $i = 1, \ldots, N_d^e$, in d:
  i. Draw treatment pattern $z_{d,i} \sim \text{Multinomial}(\theta_d)$,
  ii. Draw treatment activity $e_{d,i}.a \sim \text{Multinomial}(\phi_{z_i})$, and
  iii. Draw time stamp $e_{d,i}.t \sim \text{Multinomial}(\varphi_{z_i,e_i.a})$.

Note that we assume that θ_d, ϕ_z, $\xi_{z,a}$, ψ_z, and $\delta_{z,f}$ have a symmetric Dirichlet prior with hyper parameters α, β, and γ, respectively, in this study.

Under this generative process, each treatment pattern is drawn independently when conditioned on Θ, each patient feature is drawn independently when conditioned on Ψ and $\mathbf{z}$, each patient categorical feature's value is drawn independently when conditioned on Δ, $\mathbf{z}$ and $\mathbf{f}$, each patient numerical feature's value is drawn independently when conditioned on $\mathbf{g}$, $\mathbf{k}$, $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$ and $\mathbf{f}$, each treatment activity is drawn independently when conditioned on Φ and $\mathbf{z}$, and each treatment activity's occurring time stamp is drawn independently when conditioned on Ξ, $\mathbf{z}$ and $\mathbf{a}$. The joint distribution of patient features $\mathbf{f}$ and their values $\mathbf{v}$, treatment activities $\mathbf{a}$ and their occurring time stamps $\mathbf{t}$, and latent treatment patterns $\mathbf{z}$ is described as follows:

$$\begin{aligned} &P(\mathbf{f}, \mathbf{v}, \mathbf{a}, \mathbf{t}, \mathbf{z}|\alpha, \beta, \gamma, \eta, \iota, \mathbf{g}, \mathbf{k}, \mathbf{x}, \mathbf{y}) \\ &= P(\mathbf{z}|\alpha)P(\mathbf{f}|\mathbf{z}, \eta)P(\mathbf{v}|\mathbf{f}, \mathbf{z}, \iota, \mathbf{g}, \mathbf{k}, \mathbf{x}, \mathbf{y})P(\mathbf{a}|\mathbf{z}, \beta)P(\mathbf{t}|\mathbf{z}, \mathbf{a}, \gamma) \end{aligned} \tag{1}$$

where $\mathbf{x} = \{x_i\}_{i=1}^F$ and $\mathbf{y} = \{y_i\}_{i=1}^F$ are shape and inverse scale parameters of Gamma distributions for each numerical patient feature f_i, respectively, and $\mathbf{g} = \{g_i\}_{i=1}^F$ and $\mathbf{k} = \{k_i\}_{i=1}^F$ are mean and precision parameters of Gaussian distributions for each patient feature f_i, respectively.

Regarding Eq. 1, we can integrate out multinomial parameters in the first and second factors by using Dirichlet distributions. For $P(\mathbf{z}|\alpha)$, we have:

$$P(\mathbf{z}|\alpha) = \prod_{d=1}^{D} \frac{\prod_{z=1}^{Z} \Gamma(C_{z,d} + \alpha)}{\Gamma(C_{d,*} + \alpha Z)} \tag{2}$$

where $\Gamma(\cdot)$ is the gamma function, $C_{z,d}$ is the count of observing that patient medical record d is assigned to the treatment pattern z, and $*$ denotes the summing operation at the corresponding index, e.g., $C_{d,*} = \sum_{z=1}^{Z} C_{z,d}$.

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

$$P(\mathbf{f}|\mathbf{z}, \eta) = \prod_{z=1}^{Z} \frac{\prod_{f=1}^{F} \Gamma(C_{z,f} + \eta)}{\Gamma(C_{z,*} + \eta F)} \tag{3}$$

where $C_{z,f}$ is the count of observing that patient feature f is assigned to z, and $*$ denotes the summing operation at the corresponding index, e.g., $C_{z,*} = \sum_{f=1}^{F} C_{z,f}$.

For the third factor $P(\mathbf{v}|\mathbf{f}, \mathbf{z}, \iota, \mathbf{g}, \mathbf{k}, \mathbf{x}, \mathbf{y})$, we have:

$$P(\mathbf{v}|\mathbf{f}, \mathbf{z}, \iota, \mathbf{g}, \mathbf{k}, \mathbf{x}, \mathbf{y}) = \begin{cases} \prod_{z=1}^{Z} \prod_{f=1}^{F} \frac{\prod_{v=1}^{V_{z,f}} \Gamma(\iota + C_{z,f,v})}{\Gamma(\iota V_{z,f} + C_{z,f,*})} & : f \text{ is a categorical feature} \\ \prod_{z=1}^{Z} \prod_{f=1}^{F} (2\pi)^{-\frac{C_{z,f}}{2}} \frac{\Gamma(x_{z,f})}{\Gamma(x_f)} \frac{{y_f}^{x_f}}{{b_{z,f}}^{x_{z,f}}} (\frac{k_f}{k_{z,f}})^{\frac{1}{2}} & : f \text{ is a numerical feature} \end{cases} \tag{4}$$

where $V_{z,f}$ is the size of value set of patient feature f which is assigned to the treatment pattern z, $C_{z,f,v}$ is the count of observing that f and its categorical value v is assigned to p, and $*$ denotes the summing operation at the corresponding index, e.g., $C_{z,f,*} = \sum_{v=1}^{V_{z,f}} C_{z,f,v}$. Note that, for numerical patient feature f and its value v, $g_{z,f}$ and $k_{z,f}$ are hyperparameters of posterior distributions for mean $\mu_{z,f}$, $x_{z,f}$ and $y_{z,f}$ are hyperparameters of posterior distributions for precision $\lambda_{z,f}$. They are given as follows (Iwata and Sawada 2013):

$$g_{z,f} = \frac{k_f g_f + \sum_{v \in \mathbf{V}_{z,f}} v}{k_f + C_{z,f}} \tag{5}$$

$$k_{z,f} = k_f + C_{z,f} \tag{6}$$

$$x_{z,f} = x_f + \frac{C_{z,f}}{2} \tag{7}$$

$$y_{z,f} = y_f + \frac{\sum_{v \in \mathbf{V}_{z,f}} v^2}{2} + \frac{k_f g_f^2}{2} - \frac{k_{z,f} g_{z,f}^2}{2} \tag{8}$$

where $\mathbf{V}_{z,f}$ is the set of values of patient feature f which is assigned to treatment pattern z.

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

$$P(\mathbf{a}|\mathbf{z}, \beta) \propto \prod_{z=1}^{Z} \frac{\prod_{a=1}^{A} \Gamma(\beta_a + C_{z,a})}{\Gamma(A\beta + C_{z,*})} \tag{9}$$

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

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

$$P(\mathbf{t}|\mathbf{z}, \mathbf{a}, \gamma) \propto \prod_{z=1}^{Z} \prod_{a=1}^{A} \frac{\prod_{t=1}^{T} \Gamma(\gamma + C_{z,a,t})}{\Gamma(T\gamma + C_{z,a,*})} \tag{10}$$

where $C_{z,a,t}$ 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., $C_{z,a,*} = \sum_{t=1}^{T} C_{z,a,t}$.

Our objective is to derive the conditional Gibbs distribution $P(z_{d,i} = z|\mathbf{z}_d^{-i}, \mathbf{f}, \mathbf{v}, \mathbf{a}, \mathbf{t}, \alpha, \beta, \gamma, \eta, \iota, \mathbf{g}, \mathbf{k}, \mathbf{x}, \mathbf{y})$, where $\mathbf{z}_d^{-i} = \mathbf{z}/\{z_{d,i}\}$ denotes the set of remaining treatment pattern variables except at the ith observation of clinical word in the medical record d. Substituting Eqs. (2)–(4), (9) and (10) into Eq. (1), and using symmetric Dirichlet distribution, we obtain the conditional Gibbs distribution as follows:

$$P(z_{d,i} = z|\mathbf{f}, \mathbf{v}, \mathbf{a}, \mathbf{t}, \mathbf{z}_d^{-i}, \alpha, \beta, \gamma, \eta, \iota, \mathbf{g}, \mathbf{k}, \mathbf{x}, \mathbf{y})$$
$$\propto \frac{C_{z,d}^{-i} + \alpha}{C_{d,*}^{-i} + \alpha Z} \cdot \begin{cases} \frac{C_{z,a_i}^{-i} + \beta}{C_{z,*}^{-i} + A\beta} \cdot \frac{C_{z,a_i,t_i}^{-i} + \gamma}{C_{z,a_i,*}^{-i} + T\gamma} & : i \text{ is a clinical event} \\ \frac{C_{z,f_i}^{-i} + \eta}{C_{z,*}^{-i} + \eta F} \cdot \begin{cases} \frac{C_{z,f_i,v_i}^{-i} + \iota}{C_{z,f_i,*}^{-i} + V_{z,f_i}\iota} & : f_i \text{ is a categorical feature} \\ \frac{\Gamma(x_{z,f_i})}{\Gamma(x_{z,f_i}^{-i})} \cdot \frac{{y_{z,f_i}^{-i}}^{x_{z,f_i}^{-i}}}{y_{z,f_i}^{x_{z,f_i}}} \cdot \left(\frac{k_{z,f_i}^{-i}}{k_{z,f_i}}\right)^{\frac{1}{2}} & : f_i \text{ is a numerical feature} \end{cases} & : i \text{ is a patient feature} \end{cases} \tag{11}$$

where $(\cdot)^{-i}$ represents the count or hyperparameter when excluding sample i. The hyperparameters of Gaussian-Gamma distributions excluding sample i are calculated using Eqs. (12)–(15) as suggested in Iwata and Sawada (2013). In all cases, the current sampling position i is always excluded during counting. Details of the derivation are in Appendix.

$$g_{z,f_i}^{-i} = \frac{k_{z,f_i} g_{z,f_i} - v_i}{k_{z,f_i} - 1} \tag{12}$$

$$k_{z,f_i}^{-i} = k_{z,f_i} - 1 \tag{13}$$

$$x_{z,f_i}^{-i} = x_{z,f_i} - \frac{1}{2} \tag{14}$$

$$y_{z,f_i}^{-i} = y_{z,f_i} - \frac{v_i^2}{2} + \frac{k_{z,f_i} g_{z,f_i}^2}{2} - \frac{k_{z,f_i}^{-i} {g_{z,f_i}^{-i}}^2}{2} \tag{15}$$

Consider Eq. (11), which computes a probability of a certain treatment pattern for the present clinical word i (i.e., either a patient feature with its value or a clinical event) observed in EMR d.

We can estimate parameters $\theta_{z,d}$, $\psi_{z,f}$, $\delta_{z,f,v}$, $\mu_{z,f}$, $\lambda_{z,f}$, $\phi_{z,a}$ and $\xi_{z,a,t}$ using sampled latent treatment patterns $\mathbf{z}$ as follows:

$$\hat{\theta}_{d,z} = \frac{C_{d,z} + \alpha}{C_{d,*} + Z\alpha} \tag{16}$$

$$\hat{\psi}_{z,f} = \frac{C_{z,f} + \eta}{C_{z,*} + F\eta} \tag{17}$$

$$\hat{\delta}_{z,f,v} = \frac{C_{z,f,v} + \iota}{C_{z,f,*} + V_{z,f}\iota} \tag{18}$$

$$\hat{\mu}_{z,f} = k_{z,f} \tag{19}$$

$$\hat{\lambda}_{z,f} = \frac{x_{z,f}}{y_{z,f}} \tag{20}$$

$$\hat{\phi}_{z,a} = \frac{C_{z,a} + \beta}{C_{z,*} + A\beta} \tag{21}$$

$$\hat{\xi}_{z,a,t} = \frac{C_{z,a,t} + \gamma}{C_{z,a,*} + T\gamma} \tag{22}$$

With all the parameters derived above, we can apply the proposed TPM to various applications. For example, we can use the estimated treatment pattern proportions $\{\hat{\theta}_{d,z}\}_{z=1}^{Z}$ to describe a particular patient CP instance, including both the patient condition and treatment behaviors given that condition. We can use $\{\hat{\psi}_{z,f}, \hat{\delta}_{z,f,v}, \hat{\mu}_{z,f}, \hat{\lambda}_{z,f}\}_{f=1}^{F}$ to analyze the characteristics of patient condition w.r.t pattern z. As well, associations between patient condition and treatment behaviors can be revealed using the estimated treatment pattern proportions. It can assist to estimate the probability of a treatment activity a, and its occurring time stamp t given a particular patient by integrating out the latent treatment patterns z, i.e., $P(a, t|d) = \sum_{z=1}^{Z} \hat{\theta}_{d,z}\hat{\phi}_{z,a}\hat{\xi}_{z,a,t}$. Etc.

5 Experiments and results

In this section, we present experimental results on a real collection of EMRs of unstable angina patients to evaluate the efficiency and effectiveness of the proposed approach.

5.1 Experimental setup

The experimental data were collected from the cardiology department at the Chinese PLA General hospital. The CP of unstable angina was 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 poor blood flow in the coronary vessels caused by atherosclerosis and a 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. The symptoms for unstable angina range from exertional stable angina, to acute myocardium infarction and sudden death. While the risk of unstable angina is high, the population of unstable angina is huge, especially for elderly people and those with associated diseases such as hypertension and diabetes (Writing Committee Members 2012; Dong et al. 2014). Thus, the discovery of underlying patterns in the unstable angina CP will be of significant value and interest. Any discovered patterns could provide the user with explicit suggestions for treatment actions to influence medical behaviors for the patient's benefit.

In this case study, 985 patient medical records following the unstable angina CP were selected from the Department of Cardiology to demonstrate the ability of the proposed method in discovering latent patterns of the unstable angina CP. These patient medical records have 79 patient features, and 62,047 clinical events within 320 treatment activity types. 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#.

The case study was performed in the Cardiology Department at the Chinese PLA General hospital. Prior approval was obtained from the data protection committee of the hospital to conduct the study. We state that the patient data was anonymized in this study and in this paper.

5.2 Treatment pattern discovery

An input required for the proposed model is the number of topics to be discovered, i.e., the number of treatment patterns. In the case study, 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 geometrical 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 the literature (Blei et al. March 2003; Griffiths 2004; Wang et al. 2007; Rosen-Zvi et al. 2004). It is defined as the reciprocal geometric mean of the likelihood of a collection of EMRs given TPM,

$$\text{Perplexity}_{\text{TPM}} = \exp\left[-\frac{\sum_{d=1}^{D} log\, P(\langle \mathbf{f}_d, \mathbf{v}_d\rangle, \mathbf{e}_d|\mathcal{M})}{\sum_{d=1}^{D} N_d}\right] \tag{23}$$

where $\mathcal{M}$ is the proposed TPM, and $\mathbf{f}_d$, $\mathbf{v}_d$, and $\mathbf{e}_d$ are the set of unseen patient features, their values, and clinical events in the patient medical record d, respectively.

Regarding latent treatment pattern discovery by the proposed TPM, we set Dirichlet prior α, β, γ, η, and ι of TPM as 0.1, 0.01, 0.01, 0.01 and 0.01, respectively, which are common settings in the literature (Blei et al. March 2003; Griffiths 2004). For numerical patient features, we used $g_f = \bar{\mu}_f$, $k_f = 1$, $x_f = 1$ and $y_f = 1 + \bar{\sigma}_f^2$, where $\bar{\mu}_f$ and $\bar{\sigma}_f^2$ are the empirical mean and variance of the value of patient feature f. The number of iterations of the Markov chain for Gibbs sampling is set to 1,000. Note that Gibbs sampling usually converges before 1,000 iterations for the collected log.

Figure 2 shows the perplexity curve with respect to the number of treatment patterns, using the proposed TMP. The lower the perplexity, the better the derived model fits with the collected data-set. 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 over-explain the data-set, and it requires more sampling computation and storage as well (Phung et al. 2009). Thus, it needs to choose a balance between simplicity of the model and the degree of fitness.

To select the appropriate number of treatment patterns, we examined the discovered patterns by TPM with different value of Z by a simple way; that is, if the reducing ratio of perplexity is less than τ, we do not select a larger Z. In practice, we set τ to be 3 % according to experiment analysis. In this study, we empirically choose the number of patterns $Z = 3$ for the experimental data set, where the perplexity seems to decrease rapidly and appear to settle down.

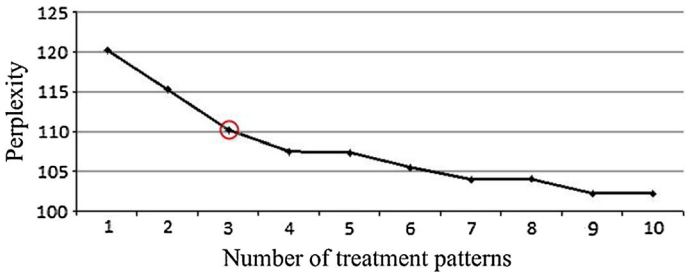

Fig. 2 Choosing number of treatment patterns using perplexity for the unstable angina data-set

For a discovered treatment pattern z, we looked at those clinical words which were assigned to z, with high probability. In the experiments, we selected a set of representative patient features $\{f|\forall z \in Z, P(f|z) > 0.001\}$ and treatment activities $\{a|\forall z \in Z, P(a|z) > 0.01\}$ to represent the discovered patterns.

Table 2 visualizes patient features and their values for discovered treatment patterns with the number of patterns set to 3. From the discovered treatment patterns, we can also see that different patterns might result in different patient features and values. For example, the average value of age in treatment pattern 1 is 68.885, and that in treatment pattern 2 is 75.216. We can determine this from the value range of patient features.

In addition, the discovered treatment patterns show the latent correlations between patient features, from which we can find some interesting phenomena. In fact, all three patterns show that there exist latent correlations between *unstable angina* and *Hypertension*, which indicates that there is a large probability for patients with *unstable angina* to have the complication *Hypertension* at the same time.

To offer a deeper understanding on treatment pattern discovery, Figs. 3, 4, and 5 show the top 50 treatment activities and their occurring time stamps for the discovered patterns (ranked by $\phi_{z,a}$). Clearly, therefore, different treatment behaviors exist for different patient conditions. A closer analysis shows that pattern 1 contains typical treatment behaviors (e.g., 'Coronary angiography', 'Stent placement', 'PTCA', etc.) for unstable angina. There is little variation occurred and common treatment activities are carried out smoothly. In clinical practice, patients who follow this dominant pattern have shorter LOS (on average 13 days) than others, and almost all physical examinations (e.g., 'CT', 'Routine blood test', etc.) are performed on the first days after admission. Pattern 2 contains typical conservative treatments of unstable angina. In clinical practice, patients who follow treatment pattern 2 have either low risks or specific physical problems, e.g., coronary stenosis such that they prefer conservative treatments instead of PCI surgery. In general, the LOS of patients who follow "Pattern 2" is greater than 13 days. Moreover, it is interesting to see that "Pattern 3", as shown in Fig. 5, has captured typical treatment behaviors of unstable angina patients who have more complex conditions than others such that many complications can be found in this pattern, e.g., hypertension, diabetes, insufficient kidney, etc. Note that this variant pattern is a bit normal in the unstable angina CP (4 % patient in the collected EMRs). Some of the patients who follow this pattern were transferred to the Cardiovascular Surgery Department to take 'Coronary Artery Bypass Grafting' (CABG), which was not generated for the other patterns. It should be mentioned that we found that most

Table 2 Patient conditions of discovered treatment patterns

Feature	Value	Prob	Feature	Value	Prob
Treatment pattern 1					
Patient condition					
Attack of angina in 24 h	true	0.0087	Sodium	Normal	0.0086
Lactate dehydrogenase	Normal	0.0082	Creatine kinase	Normal	0.0082
Platelet counts	Normal	0.0079	Mean corpuscular hemoglobin amount	Normal	0.0079
Low-density lipoprotein cholesterol	Normal	0.0077	Creatinine	Normal	0.0073
Hemoglobin measurement	Normal	0.0072	Creatine kinase isoenzyme	Normal	0.0070
Mean platelet volume measurement	Normal	0.0068	Gender	Male	0.0067
Total cholesterol	Normal	0.0066	Hypertension	True	0.0065
Triglycerides	Normal	0.0062	History of CHD	True	0.0053
High-density lipoprotein cholesterol	Normal	0.0046	Glucose	Normal	0.0045
Qualitative urine glucose test	895.543, 256.623	0.0042	Troponin T	Normal	0.0042
Age	68.885,10.515	0.0039	Post-PCI	True	0.0034
High-density lipoprotein cholesterol	Low	0.0031	Serum ferritin	Normal	0.0029
Quantitative determination of creatine kinase isoenzyme	Normal	0.0025	Gender	Female	0.0024
Diabetes	True	0.0022	Triglycerides	High	0.0021
Glucose	High	0.0019	Brain natriuretic peptide precursor	High	0.0015
Hyperlipidemia	True	0.0016	Tumor	True	0.0015
Total cholesterol	Low	0.0013	Platelet volume distribution width	13.254, 2.038	0.0011
Cardiac insufficiency	True	0.0010	Platelet volume measurement	0.208,0.083	0.0010
Hemoglobin measurement	Low	0.0010	Creatinine	High	0.0010

Table 2 continued

Feature	Value	Prob	Feature	Value	Prob
Treatment pattern 2					
Patient condition					
Age	75.216, 9.94	0.0121	Qualitative urine glucose test	583.636, 371.735	0.0044
Platelet volume measurement	0.192, 0.0564	0.0036	Platelet volume distribution width	12.74, 2.533	0.0034
Brain natriuretic peptide precursor	High	0.0024	Hypertension	true	0.0020
Quantitative determination of creatine kinase isoenzyme	Normal	0.0019	Blood glucose	6.846, 5.016	0.0018
History of CHD	True	0.0016	Post-PCI	True	0.0013
Creatinine	High	0.0012	Tumor	True	0.0012
Cardiac insufficiency	true	0.0011	Diabetes	True	0.0011
Treatment pattern 3					
Patient condition					
Age	78.542, 8.916	0.0179	Qualitative urine glucose test	409.406, 313.454	0.0069
Platelet volume measurement	0.182, 0.051	0.0062	Platelet volume distribution width	12.495, 1.881	0.0059
Blood glucose	7.242, 4.647	0.0038	Thromboelastography TPI- the index of thrombosis	52.1, 2.842E-14	0.0022
History of CHD	True	0.0014	Cardiac insufficiency	True	0.0014
Insufficiency of kidney function	True	0.0014	Hypertension	True	0.0013
Pulmonary disease	Tru	0.0012	Tumor	True	0.0012
Post-PCI	True	0.0011	Hyperlipidemia	True	0.0010

Patient condition of each pattern is depicted by a set of most related patient feature-value pairs, and their probabilities

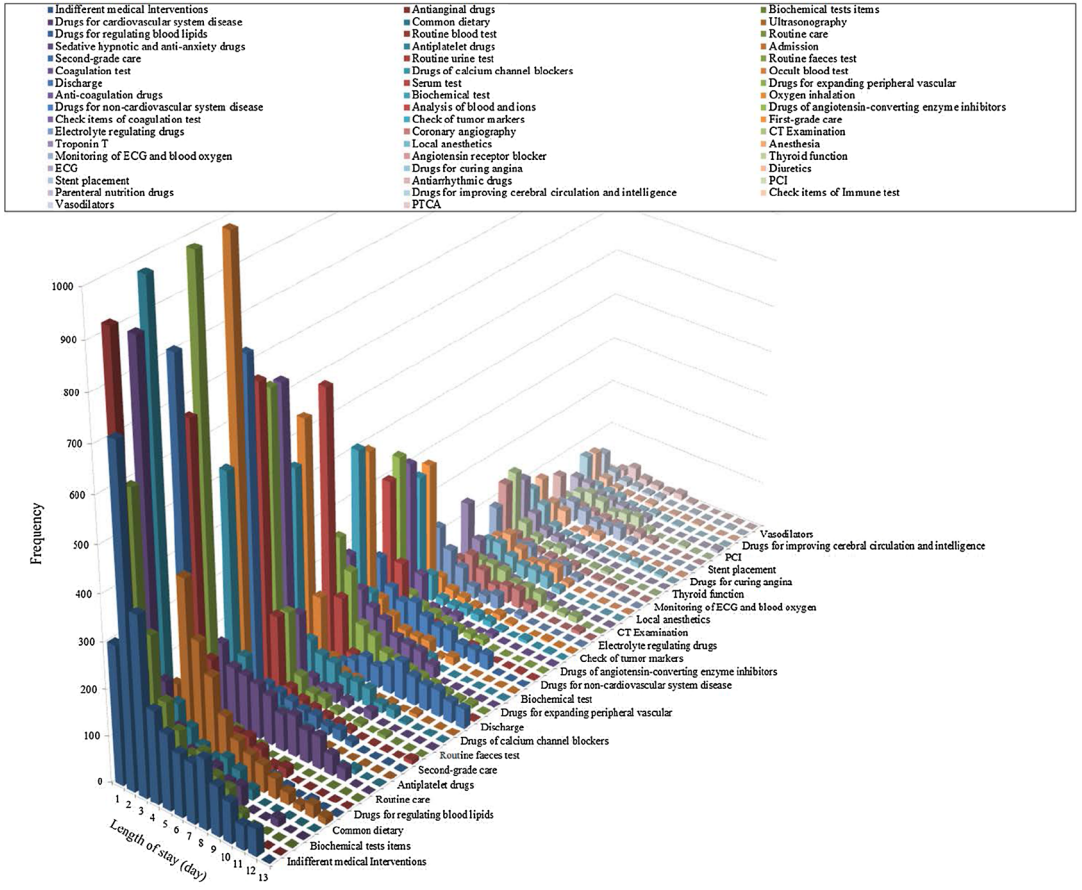

Fig. 3 Typical treatment behaviors of the discovered pattern 1 from the unstable angina EMRs

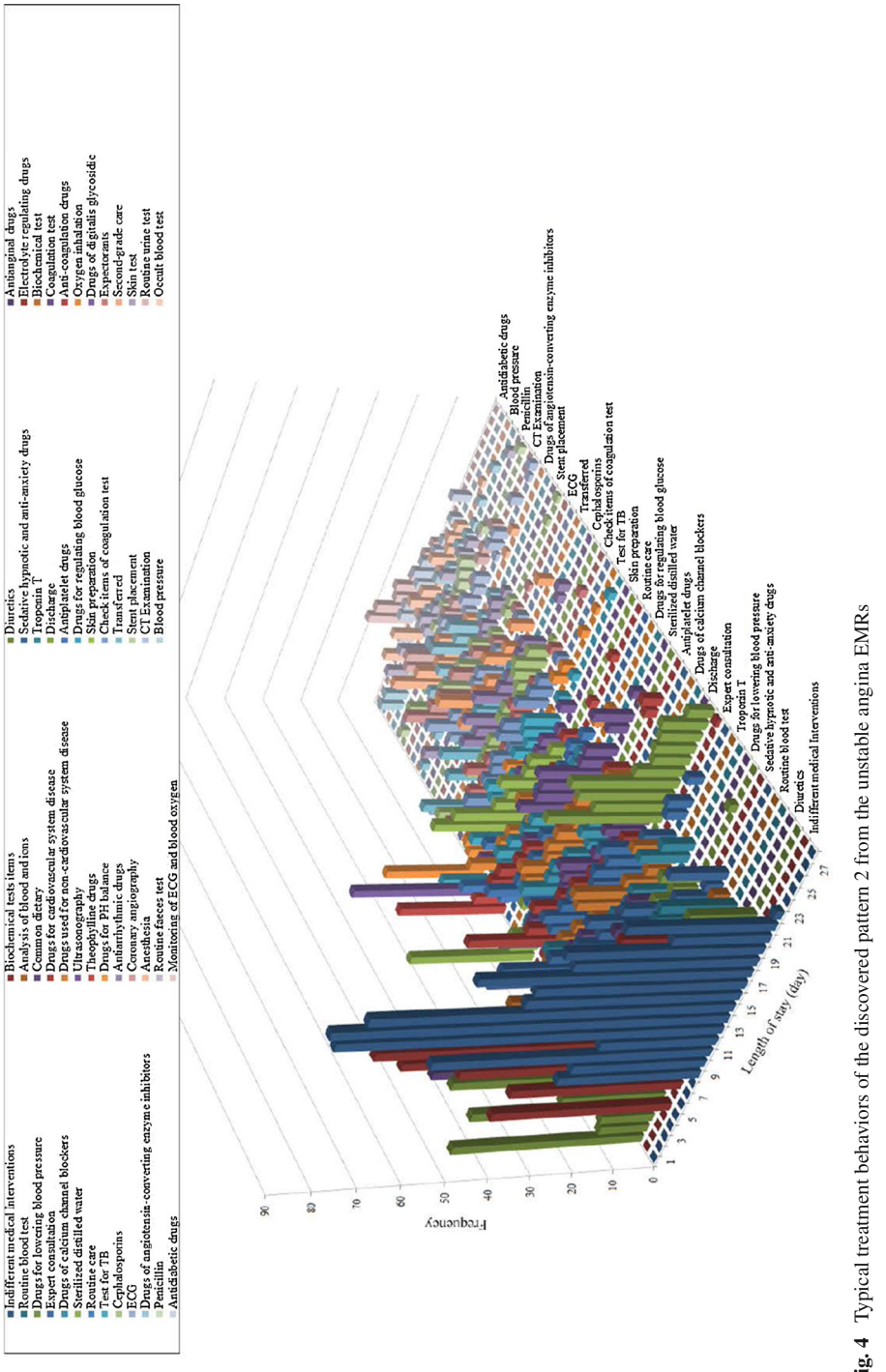

Fig. 4 Typical treatment behaviors of the discovered pattern 2 from the unstable angina EMRs

Fig. 5 Typical treatment behaviors of the discovered pattern 3 from the unstable angina EMRs

patients who follow "Pattern 3" also take "Pattern 1" or "Pattern 2" as well. Typically, they are best represented as a mixture of treatment patterns.

5.3 Treatment pattern analysis w.r.t patient demographics

Furthermore, we wanted to study the characteristics of treatment patterns w.r.t patient demographics, e.g., what is the average LOS of patients for each pattern? Do the patients of each pattern have similar risk scores? Does the gender of patients have an impact on their treatment processes? To this end, we classified patients into 3 clusters by checking the distribution of treatment patterns on EMRs, i.e., $\theta_{d,z}$. Each cluster is associated with a specific treatment pattern. If $\theta_{d,z} \geq 0.25$, we assume the patient whose medical record is d belongs to the cluster z. Apparently, some patients may belong to more than one cluster.

Figure 6a shows the TIMI risk scores of patients of each cluster. TIMI risk score (Antman et al. 2000) is a simple prognostication scheme that categorizes a patient's risk of death and ischemic events such that it indicates patient demographics implicitly. From Fig. 6a, we find that patients in cluster 3 have the highest TIMI risk scores. As shown in Table 2, most of the indicators for TIMI score calculation are observed in the patient feature list of treatment pattern 3, such as the average age of patients (i.e., 78.542, which is larger than 65), history of CHD, cardiac insufficiency, hypertension, diabetes, and hypercholesterolemia, etc. In comparison, few risk indicators can be found in treatment pattern 1, of which patients have smaller TIMI risk scores than other patterns. It also confirms our assumption that the discovered treatment patterns can reveal the actual patients' conditions and provide a basis for therapeutic decision making.

Figure 6b shows the gender distribution for the discovered patterns. It seems that male patients are more likely to have unstable angina than female patients in all three patterns. It is quite a surprise since most literatures indicate that female patients have more risks for unstable angina than do male patients (Writing Committee Members 2012). We will investigate this finding using a larger volume of EMRs in our future work. Furthermore, Fig. 6c and d show that patients with treatment pattern 3 are older than others, and have larger LOS than others. As shown in Fig. 6a, patients with treatment pattern 3 have higher TIMI risk scores than others. It, therefore, indicates that high-risk patients may be older and have larger LOS than low-risk patients.

In many real cases, patients probably follow more than one pattern, i.e., they are typically a mixture of treatment patterns. To this end, we further studied the impact of treatment pattern combinations on patient conditions. Figure 7a shows the curve of the number of patients who follow 1–3 patterns, which indicates that the patient population declines with the size of pattern combinations increasing. In other words, the patient population with more patterns is smaller than the one with fewer. Figure 7b shows the TIMI risk score curve changing with the size of treatment pattern combinations, and demonstrates that the TIMI risk score has an obvious growth trend when the number of patterns increases. Thus, we conclude that the high-risk patients are likely to follow more treatment patterns simultaneously than low-risk patients. Figure 7c shows the

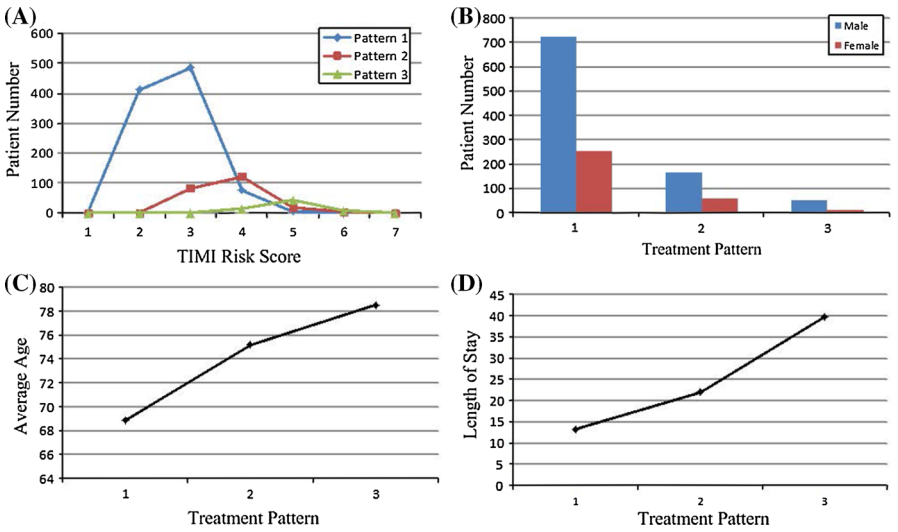

Fig. 6 The discovered treatment patterns w.r.t patient demographics

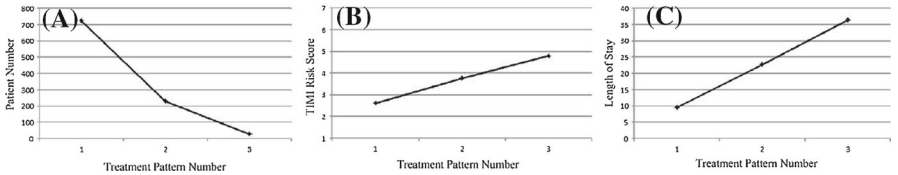

Fig. 7 The impact of treatment pattern combinations on patient conditions

average LOS curve changing with the size of pattern combinations, and demonstrates that the average LOS increases as the size of pattern combination increases.

Table 3 lists the top-ranked treatment activities of 3 patterns learned by the standard LDA topic model with the same setting of pattern numbers. It is easy to find that the learned patterns not only enumerate regular treatment behaviors that are expected to occur in CPs, but also disclose the correlations between patient-specific information and treatment behaviors. Thus, the proposed TPM is more effective in discovering one or more patient condition-dependent treatment patterns from EMRs.

5.4 Treatment pattern analysis w.r.t clinical outcomes

As stated earlier, the proposed TPM provides a basis for further CP analysis tasks. In this subsection, we set out to answer an interesting question using the proposed TPM, i.e., how do the treatment patterns discovered characterize the set of patients and lead to specific clinical outcomes in CPs? To this end, we propose a Bayesian model, as shown in Fig. 8, to derive the correlations between the discovered treatment patterns and multiple clinical outcomes. This differs from previous approaches in that each time the model focuses on finding patterns correlated to one outcome label only (Huang et al. 2013), the proposed model uses all EMRs and finds the patterns that are

Table 3 Treatment pattern discovery results with LDA

Pattern id	Significant activities
1	Indifferent medical interventions, antianginal drugs, biochemical tests items, sedative & hypontic and anti-anxiety drugs, ultrasonography, drugs for cardiovascular system disease, draw off blood, meal, routine blood test, routine care, drugs for regulating blood lipids, coagulation test, routine faces test, second-grade care, discharge, routine urine test, admission, drugs for lowering blood pressure, drugs of calcium channel blockers, occult blood test
2	Indifferent medical interventions, antiplatelet drugs, meal, drug replacement, antianginal drugs, anti-coagulation drugs, biochemical tests items, drugs for regulating blood lipids, skin preparation, ultrasonography, second-grade care, routine care, coronary angiography, discharge, routine blood test, admission, drugs for lowering blood pressure, anesthesia, coagulation test, routine urine test
3	Diuretics, biochemical tests items, draw off blood, indifferent medical interventions, routine blood test, electrolyte regulating drugs, analysis of blood and ions, antianginal drugs, drugs for regulating blood glucose, drug replacement, biochemical test, oxygen inhalation, troponin T, meal, sedative & hypnotic and anti-anxiety drugs, drugs for lowering blood pressure, coagulation test, cephalosporins, drugs of calcium channel blockers, expert consultation

Top treatment activities are listed for these patterns, ranked by $P(a|z)$

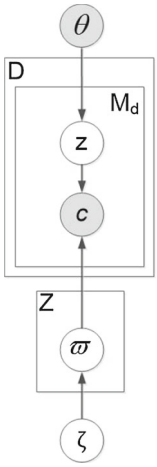

Fig. 8 Treatment pattern-clinical outcome model. Z denotes the variable "treatment patterns", "patient features". θ is the learned distribution of treatment patterns over EMRs based on TPM. ω is the distribution of clinical outcomes over treatment patterns. c represents clinical outcomes. M_d represents the number of clinical outcomes for one EMR d. D represents the number of EMRs in the collected data set. The hyperparameter ζ controls dispersions of clinical outcomes

specific to the target clinical outcome label and are helpful in differentiating EMRs of different clinical outcome labels.

Formally, let C be a domain of clinical outcomes. For each piece of EMR d, there is a group clinical outcomes $\mathbf{c}_d$ ($\mathbf{c}_d \subseteq C$), including a vast range of descriptions on patient health states (e.g., mortality, transfer, or normal discharge, etc.), physiologic measures (e.g., heart attack, etc.), and patient-reported health states (e.g., Length of stay in 7 days, readmission in one month, etc.). For clinical outcomes, ω denotes the $C \times Z$ matrix of clinical outcome-treatment pattern distributions, with a multinomial distribution over C outcome classes for each of Z treatment patterns drawn independently from a Dirichlet(ζ) prior.

Table 4 The probability of clinical outcomes conditioned on each treatment pattern ($P(c|z)$)

	Treatment pattern 1	Treatment pattern 2	Treatment pattern 3
LOS $\leq$ 7 days	0.308	0	0
7 days $<$ LOS $\geq$ 14 days	0.006	0.258	0
14 days $<$ LOS $\geq$ 28 days	0	0.081	0.219
LOS $>$ 28 days	0	0	0.223
Readmission in 1 month	0.003	0.014	0
Readmission in 6 months	0.006	0.026	0
Readmission larger than 6 months or without readmission	0.333	0.309	0.248
Normal discharge	0.331	0.276	0
Transfer	0.013	0.036	0.297
Death	0	0	0.013

After generating all EMRs $\mathcal{D}$ by the proposed TPM, the posterior distribution for $\hat{\theta}_d$ is further used to generate clinical outcomes as follows:

1. Choose a treatment pattern $z_{d,j}$ from $\hat{\theta}_d$,
2. Choose a clinical outcome $c_{d,j} \sim Multinomial(\omega_{z_{d,j}})$.

For each clinical outcome c, the following conditional posterior distribution is used

$$P(z_{d,i} = z|c_{d,i} = c, z_d^{-i}, c_d^{-i}, \zeta) \sim \hat{\theta}_{d,z} \cdot \frac{M_{z,c}^{-i} + \zeta}{M_{z,*}^{-i} + C\zeta} \tag{24}$$

where $M_{z,c}$ is the number of times treatment pattern z is assigned to a clinical outcome c, and $M_{z,*}$ is the pattern-outcome sum.

In the experiments, we set out 10 clinical outcome variables, based on the suggestions of clinical experts, to investigate the patient benefits from the pathway. Each EMR in the collection can be categorized into one or several classes using these labels. As shown in Table 4, the representative samples of clinical outcomes are generated by TPM. The probability of clinical outcomes conditioned on each treatment pattern is estimated by Eq. (24).

The results shown in Table 4 indicate that the proposed TPM can be exploited to find the correlations between latent treatment patterns and specific clinical outcomes. For example, patients who follow the treatment pattern-1 may have less than 7 days of LOS, and may be normal discharged. In general, patients who follow pattern-1 have higher probabilities to be readmitted after six months or without readmission (0. 333) than in one month (0. 003) or in six months (0.006). In addition, Table 4 shows that our model can distinguish the patterns evoking strong clinical outcomes from background treatment patterns. For example, pattern-1 and pattern-3 trigger different clinical outcomes, such as different LOS (i.e., "LOS $>$ 28 days", and "LOS $\leq$ 7 days"), and different discharge states (i.e., "Normal discharge" and "Transfer"). Furthermore, treatment patterns may result in multiple typical clinical outcomes. For example, "LOS

$>$ 28 days" and "Transfer" are highly correlated for the pattern-3. Last but not the least, our model can discover some rare but interesting findings from clinical outcome samples. For example, as shown in Table 4, LOS of patients who follow the treatment pattern-1 and pattern-2 are generally less than 14 days. It is not surprised since both patterns are generally applied to low-risk patients, as indicated in Fig. 6, who are expected to have small LOS. However, there are few patients who follow pattern-1 and pattern-2 readmit to the hospital in six months (0.006 and 0.026, respectively). It indicates that the treatments for these patients do not achieve the expected quality in CPs. Clinical experts may have interests in analyzing the demographics of these patients, and investigating if it is appropriate to perform the treatment behaviors of the patterns on these patients, so as to adjust and optimize the pathway.

As we mentioned above, an important parameter to the proposed TPM is the number of latent treatment patterns, which indicates how many aspects of EMRs can be derived. To this end, the accuracy of clinical outcome classification can be employed as the indicator of performance with respect to different number of patterns. Formally, given an unlabeled patient record d, its truth clinical outcome set is $C_{d,b}$, including $|C_{d,b}|$ clinical outcomes, and its $|C_{d,p}|$ top-ranked predicted outcome set is $C_{d,p}$, where $|C_{d,b}| \equiv |C_{d,p}|$. Then, accuracy is computed through dividing the number of correctly predicted clinical outcomes by the total number of outcomes obtained in EMRs, i.e.,

$$Accuracy = \frac{\sum_{d \in \mathcal{D}} |C_{d,b} \bigcap C_{d,p}|}{\sum_{d \in \mathcal{D}} |C_{d,b}|} \tag{25}$$

To calculate *Accuracy*, we split the experimental dataset into a training set and a testing set, and evaluate the performance by 5-fold cross-validation. In addition, we take clinical outcomes as multi labels, and employ three widely-used multi-label classification algorithms (Tsoumakas and Katakis 2007), i.e., Binary Relevance (BR), RAndom k-labELsets (RAKEL) and Multi-label k-nearest neighbors (ML-kNN), for comparison.[3] The results are shown in Fig. 9.

The experimental results show that the performance of the proposed TPM-based method converges with a relatively small number of treatment patterns. The average accuracy of our method is 86.08 %, and the standard deviation of the accuracy is 0.024, indicating that the performance of the proposed method is quite stable, in particular, when the number of treatment patterns is larger than 3. It is consistent with the experimental conclusion in the view of perplexity.

In addition, the proposed method always outperforms the baseline BR, RAkEL, and ML-kNN. Note that BR, RAkEL, ML-kNN are discriminative models, and the performances of BR, RAkEL, ML-kNN is to classify EMRs regarding their clinical outcomes. On the other hand, the proposed TPM-based method, as a typical generative method, can not only classify EMRs with respect to their clinical outcomes, but also reveal the details of underlying treatment patterns for CPs.

[3] We used a well-known toolkit, i.e., MEKA (http://meka.sourceforge.net/), for the task of multi-label classification.

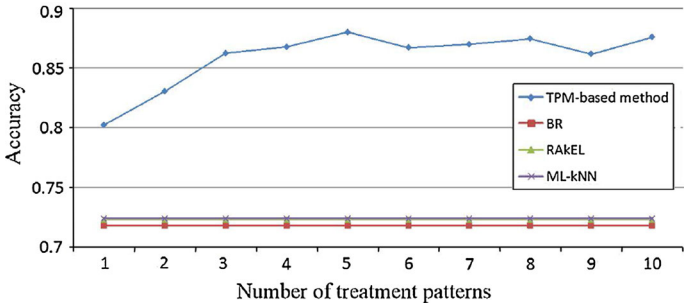

Fig. 9 The performance of clinical outcome classification with different number of treatment patterns

5.5 Treatment recommendation based on the proposed TPM

The proposed approach can not only provide a basis for further CP analysis but also be utilized to support clinical decision-making. In this subsection, we present a promising application of the proposed TPM, i.e., a treatment recommendation service, which can be used to guide physicians in CPs by providing recommendations on possible next steps based on the measurement of the target patient conditions and medical knowledge from completed clinical cases.

As illustrated above, the probability of a treatment activity a given a particular patient by integrating out the latent treatment patterns z, i.e.,

$$P(a|d) = \sum_{z=1}^{Z} \hat{\theta}_{d,z} \hat{\phi}_{z,a} \tag{26}$$

Based on Eq. (26), we can test the performance of treatment recommendation of our model quantitatively. To this end, we randomly selected 100 EMRs as a testing data-set and used the rest records to train the model. W.r.t the ground-truth, we asked three experienced physicians in the Cardiology department of the Chinese PLA General hospital to select the top-10 and -50 most important types of treatment activities from the test EMRs adopting a majority voting.[4] Then, we checked the consistency of the possible treatment interventions suggested by our model with the ground-truth. More specifically, we recommended the top 10 and 50 types of treatment activities for each test piece of EMR and checked if these are actually equal with ones in the top 10 or 50 types of treatment activities ranked by physicians. In this sense, we utilize two measurements which are "mean precision at top 10", and "mean precision at top 50".

In comparison with our model, a K-nearest neighbor (kNN) model was employed by using the weighted combination of treatment interventions of the k (k = 10 in this study) nearest neighbors as the suggested treatments. In addition, the standard LDA (Huang et al. 2013a), and an extension of LDA proposed in our previous work,

[4] Note that in clinical settings, the given treatments are biased. Even for the same patient, different physicians may have different opinions on patient conditions so as to give different treatment interventions.

Table 5 The results of treatment recommendation

	Mean precision at top 10	Mean precision at top 50
TPM	0.606	0.545
kNN	0.292	0.234
CPM	0.535	0.497
LDA	0.549	0.505

i.e., CPM (Huang et al. 2014), were employed. Table 5 shows the experimental results.

As we can see from Table 5, the LDA-based models outperform kNN significantly in terms of treatment recommendations. In addition, TPM achieves the best performance among the three LDA-based models, which indicates that the effectiveness of treatment recommendations can be significantly improved if the patient-specific information is incorporated into the model.

5.6 Proof-of-concept prototype

We have implemented a system prototype using Microsoft C# and ASP.net, which provides web services, including upload of EMRs and treatment pattern analysis using EMRs. The proposed TPM has been implemented in the prototype. Some basic information of patient traces in a selected EMR such as patient ID, department, LOS, etc., are shown in Fig. 10a, while the screen-shot of the probability distributions of the derived patterns for a selected EMR shows on Fig. 10b. The generated patterns indicate the actual patient condition and treatment behaviors being applied given that condition. Figure 10c depicts a screen-shot of the probability distributions of treatment activities and their occurring time stamps for a particular treatment pattern. Users can observe the derived pattern from different angles by adjusting the display parameters shown on the setting panel in the bottom of Fig. 10c.

6 Discussion

The experimental results demonstrate the effectiveness of our approach and the potential of using the discovered knowledge for CP analysis and improvement. The benefits are listed relating to the following aspects:

– The discovered treatment patterns have been evaluated by clinical experts from Chinese PLA general hospital, who indicate that the mining results of our approach: (1) disclose the correlations between patient conditions and treatment activities; (2) allow treatment activities to be clearly spread along the time-line of CPs with specific occurring probabilities; and (3) let a treatment pattern enumerate regular treatment behaviors that are expected to occur in CPs, which serve as checkpoints for the performance of CPs. In general, physicians from the hospital are satisfied with the mined results. The evaluations received indicate that the proposed

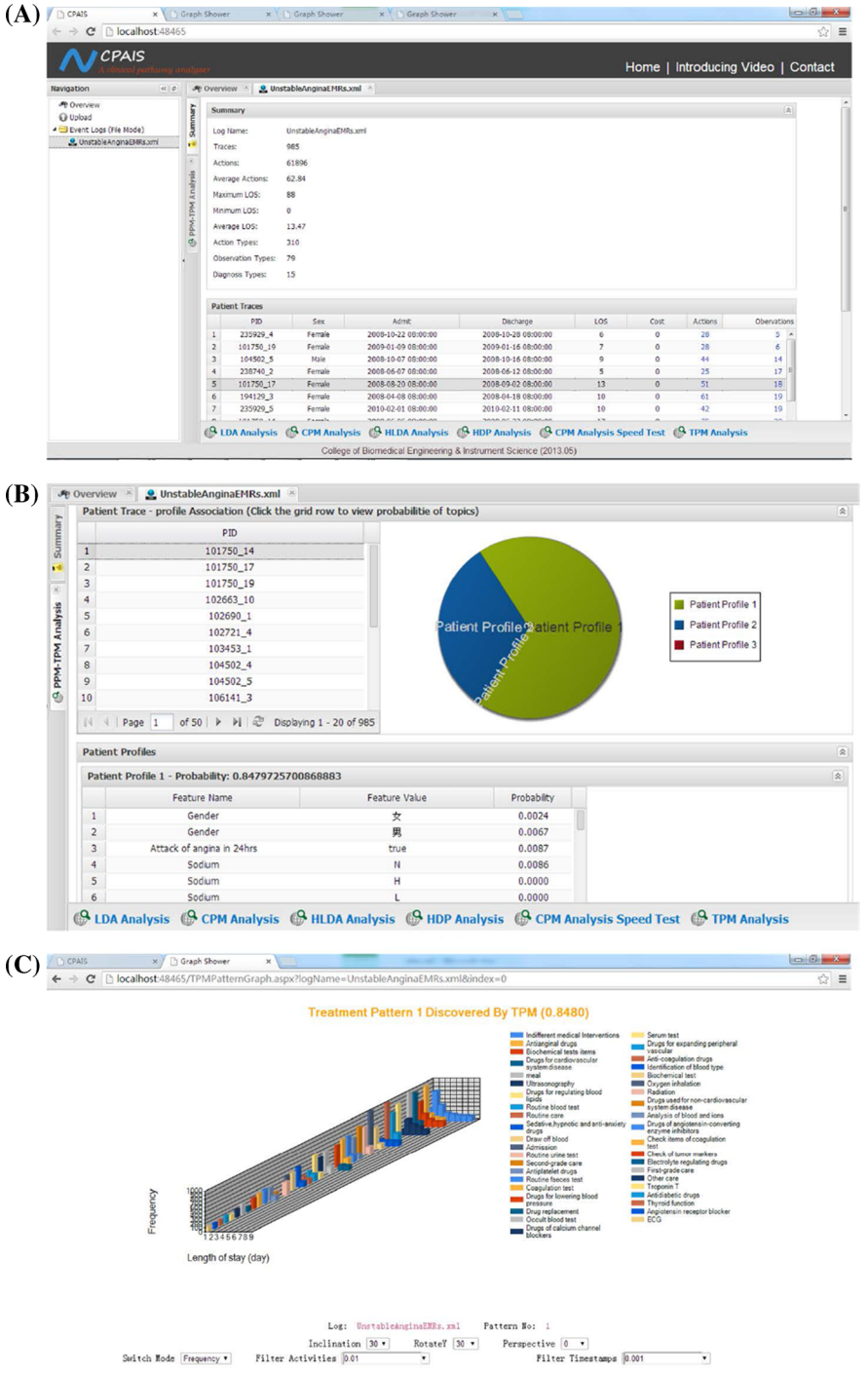

Fig. 10 A screen-shot of the system prototype

approach has the ability to find a clear characterization of possible treatment patterns for particular diseases.
- The discovered treatment patterns form the backbone of CPs, and thus can be utilized to support CP redesign and improvement. On the one hand, antecedent patient conditions of the discovered patterns provide domain-specific information to find out the exact meanings of these patterns. On the other hand, subsequent treatment behaviors of discovered patterns demonstrate that, according to different patient conditions the treatment behaviors are different. Thus, they provide useful insights into the pathway, and can hence be straightforwardly included explicitly as background knowledge for further analytical objectives. Since CPs may not perform as desired according to various measures, and have to be subsequently redesigned, actionable knowledge discovered from our models 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 optimize CPs continuously.
- The discovered treatment patterns can be utilized to support clinical decision-making and aid clinical diagnosis. For instance, we can analyze the characteristics of patient conditions for their demographics to aid physicians to create patient-specific CPs. It is also applicable to clinical decision support systems that recommend proper treatment behaviors matching with the specific patient conditions. This might guide physicians in CPs by giving recommendations on possible next steps based on the measurement of the target patient conditions and medical knowledge, which would be learned from completed clinical cases.

Although our results have been encouraging to date, the proposed approach could be further improved in a number of ways:

- We could automatically infer the number of treatment patterns by extending the proposed probabilistic topic models to a nonparametric Bayesian model such as hierarchical Dirichlet processes (Whye Teh et al. 2004). In addition, for the proposed TPM, we could incorporate additional information (such as resources of radiological examinations, quantities and costs of medications, etc) other than treatment activities and their occurring time stamps into the model to allow us to analyze treatment behaviors at a very refined level. Note that EMRs might contain fruitful information of patient features and treatment behaviors, and thus it would be promising to exploit this data to provide health-care 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 CPs (Peleg et al. 2008; Ghattas et al. 2010).
- In this study, the patient conditions of the discovered treatment patterns are generated using the data collected at the admission stage of CPs. However, the dynamic nature of patient conditions is often essential to treatment behaviors adopted in CPs. During the executions of CPs, patient conditions could be changed so as to influence physicians' diagnostic conclusions. To this end, the selection of treatment behaviors may depend on the time at which the selection is made. Besides, new evidence often becomes available at time-points, which could cause the vari-

ant treatment behaviors as well. Apparently, by incorporating richer execution information into the proposed models to disclose the dynamic features of CPs, our method could be more intelligent.

– This study proposes a probabilistic generative model to simultaneously capture the correlations between patient-specific information and treatment interventions from the heterogeneous medical records. However, the causal relationship and the the interaction between the two were not taken into consideration by the model. Causal analysis can be useful to clinical analysts to find out unexpected changes of patient status and thus appropriate treatment behaviors could be performed on patients in CPs. Association rule mining could be a possible choice to address this challenge (Wang et al. 2014). As well, sequence pattern mining could also be used to classify and analyze interactions among treatment events in CPs (Huang et al. 2012). However, the interesting question that remains address the issues of how to design efficient algorithms for mining the causal relationships and the interactions among treatment events in CPs, as well as, how to explain the discovered causal relationships and the interactions in a maximum-informative manner. Much research is still needed to make such mining both effective and efficient.

7 Conclusion

In this paper, we present a new approach of discovering underlying treatment patterns from EMRs. In detail, a new probabilistic topic model is developed to link patient features and treatment behaviors together to mine treatment patterns hidden in EMRs. Experimental results on a collection of EMRs of unstable angina patients from the Chinese PLA general hospital demonstrate intrinsic patient characteristics and meaningful treatment patterns discovered by our models.

The discovered treatment patterns have been evaluated by hospital managers and clinical experts at the Chinese PLA General hospital, who understand the beneficial effects of our study. They indicated that the discovered knowledge from EMRs support CP (re)design and improvement. Despite that, the proposed approach is not a tool for designing CPs, it is evident that a good understanding of the existing patient treatment processes is vital for any design and improvement effort (Huang et al. 2013a). Since a large collection of EMRs becomes available in hospitals nowadays, they can be meaningfully used to derive nontrivial knowledge explaining treatment intentions and behaviors in CPs. Besides, discovered knowledge is not biased by perceptions, and is useful to confront with the man-made CP specifications. Thus, it might be effective in CP analysis and improvement.

Although our study reveals that the proposed approach is effective in discovering efficient patterns, there are even more complex analysis and evaluation tasks that need to be considered. In fact, our clinical collaborators from Chinese PLA general hospital have indicated that, even though our approach is efficient for mining precise and complete set of regular treatment patterns in CPs, there are still a number of infrequent behaviors that are missing in the discovered patterns. Note that many of these infrequent behaviors are correlated with the treatments of the comorbidities of patients, and thus need to be discovered and analyzed. From this perspective, the interesting

questions that remain address the issues of how to design efficient algorithms for mining and detecting variants in CPs, as well as, how to explain these variants in a maximally-informative manner. Much research is still needed to make such mining both effective and efficient.

The issue of meaningful or secondary use of EMRs represents a promising and interesting research direction in health informatics. Our study indicates the feasibility of exploring EMRs to support CP-oriented and patient-specific clinical decision making. There are many potential applications of our work, such as a patient-specific treatment recommendation service in CPs, treatment behavior grouping and identification within the same therapy and treatment intention, and anomaly detection from normal treatment behaviors, etc. As for future work, we are planning to evaluate our approach with a larger scale of EMR collections, and 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, the National Hi-Tech R&D Plan of China under Grant No 2012AA02A601, and the Fundamental Research Funds for the Central Universities under Grant No 2014QNA5014. 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. The authors would like to thank the anonymous reviewers for their constructive comments on an earlier draft of this paper.

Appendix

In this appendix, we give the derivation of Eq. (11)

$$\begin{aligned} &P(\mathbf{f}, \mathbf{v}, \mathbf{a}, \mathbf{t}, \mathbf{z}|\alpha, \beta, \gamma, \eta, \iota, \mathbf{g}, \mathbf{k}, \mathbf{x}, \mathbf{y}) \\ &= P(\mathbf{z}|\alpha)P(\mathbf{f}|\mathbf{z}, \eta)P(\mathbf{v}|\mathbf{z}, \mathbf{f}, \iota, \mathbf{g}, \mathbf{k}, \mathbf{x}, \mathbf{y})P(\mathbf{a}|\mathbf{z}, \beta)P(\mathbf{t}|\mathbf{z}, \mathbf{a}, \gamma) \end{aligned} \tag{27}$$

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

$$\begin{aligned} P(\mathbf{z}|\alpha) &= \int P(\mathbf{z}|\Theta)P(\theta|\alpha)d\Theta \\ &= \int \prod_{d=1}^{D} \left(\prod_{i=1}^{N_d} P(z_{d,i}|\theta_d)P(\theta_d|\alpha) \right) d\Theta \\ &= \int \prod_{d=1}^{D} \prod_{z=1}^{Z} \theta_{d,z}^{C_{d,z}} \prod_{d=1}^{D} \left(\frac{\Gamma(Z\alpha)}{\Gamma(\alpha)^Z} \prod_{z=1}^{Z} \theta_{d,z}^{\alpha-1} \right) d\Theta \\ &\propto \prod_{d=1}^{D} \frac{\prod_{z=1}^{Z} \Gamma(C_{d,z} + \alpha)}{\Gamma(\sum_{z=1}^{Z}(C_{d,z} + \alpha))} \end{aligned} \tag{28}$$

For $P(\mathbf{f}|\mathbf{z}, \eta)$, we have

$$\begin{aligned} P(\mathbf{f}|\mathbf{z}, \eta) &= \int P(\mathbf{f}|\Psi, \mathbf{z}) P(\Psi|\eta) d\Psi \\ &= \int \prod_{d=1}^{D} \prod_{i=1}^{N_d^f} P(f|\psi_{z_{d_i}}) \prod_{z=1}^{Z} P(\psi_{z_{d_i}}|\eta) d\Psi \\ &= \int \prod_{z=1}^{Z} \prod_{f=1}^{F} \psi_{z,f}^{C_{z,f}} \prod_{z=1}^{Z} \left(\frac{\Gamma(F\beta)}{\Gamma(\beta)^F} \prod_{f=1}^{F} \psi_{z,f}^{\beta-1} \right) d\Psi \\ &\propto \prod_{z=1}^{Z} \frac{\prod_{f=1}^{F} \Gamma(C_{z,f} + \beta)}{\Gamma(\sum_{f=1}^{F} (C_{z,f} + \beta))} \end{aligned} \tag{29}$$

For $P(\mathbf{v}|\mathbf{z}, \mathbf{f}, \iota, \mathbf{g}, \mathbf{k}, \mathbf{x}, \mathbf{y})$, we have

$$\begin{aligned} &P(\mathbf{v}|\mathbf{z}, \mathbf{f}, \iota, \mathbf{g}, \mathbf{k}, \mathbf{x}, \mathbf{y}) \\ &= \int P(\mathbf{v}|\mathbf{z}, \mathbf{f}, \Delta) P(\Delta|\iota) d\Delta \prod_{z=1}^{Z} \prod_{f=1}^{F} P(V_{z,f}|\mathbf{g}, \mathbf{k}, \mathbf{x}, \mathbf{y}) \\ &= \begin{cases} \int \prod_{d=1}^{D} \prod_{i=1}^{N_d^f} P(v_i|\delta_{z,f}) \prod_{z=1}^{Z} \prod_{f=1}^{F} P(\delta_{z,f}|\iota) d\Delta & : f \text{ is a categorical feature} \\ \prod_{z=1}^{Z} \prod_{f=1}^{F} \int \int \prod_{v \in \mathbf{v}_{z,f}} P(v|\mu_{z,f}, \lambda_{z,f}^{-1}) & \\ P(\mu_{z,f}|g_f, (k_f \lambda_{z,f})^{-1}) P(\lambda_{z,f}|x_i, y_i) d\mu_{z,f} d\lambda_{z,f} & : f \text{ is a numerical feature} \end{cases} \\ &= \begin{cases} \int \prod_{z=1}^{Z} \prod_{f=1}^{F} \left(\prod_{v=1}^{V_{z,f}} \delta_{z,f,v}^{C_{z,f,v}} \frac{\Gamma(\iota V_{z,f})}{\Gamma(\iota)^{V_{z,f}}} \prod_{v=1}^{V_{z,f}} \delta_{z,f,v}^{\iota-1} \right) d\Delta & : f \text{ is a categorical feature} \\ \prod_{z=1}^{Z} \prod_{f=1}^{F} (2\pi)^{-\frac{C_{z,f}}{2}} \frac{\Gamma(x_f)}{\Gamma(x_f)} \frac{{y_f}^{x_f}}{{y_{z,f}}^{x_{z,f}}} \left(\frac{k_f}{k_{z,f}}\right)^{\frac{1}{2}} & : f \text{ is a numerical feature} \end{cases} \\ &\propto \begin{cases} \prod_{z=1}^{Z} \prod_{f=1}^{F} \frac{\prod_{v=1}^{V_{z,f}} \Gamma(C_{z,f,v} + \iota)}{\Gamma(\sum_{v=1}^{V_{z,f}} (C_{z,f,v} + \iota))} & : f \text{ is a categorical feature} \\ \prod_{z=1}^{Z} \prod_{f=1}^{F} (2\pi)^{-\frac{C_{z,f}}{2}} \frac{\Gamma(x_f)}{\Gamma(x_f)} \frac{{y_f}^{x_f}}{{y_{z,f}}^{x_{z,f}}} \left(\frac{k_f}{k_{z,f}}\right)^{\frac{1}{2}} & : f \text{ is a numerical feature} \end{cases} \end{aligned} \tag{30}$$

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

$$\begin{aligned} P(\mathbf{a}|\mathbf{z}, \beta) &= \int P(\mathbf{a}|\Phi, \mathbf{z}) P(\Phi|\beta) d\Phi \\ &= \int \prod_{d=1}^{D} \prod_{i=1}^{N_d^e} P(e_i.a|\phi_{z_i}) \prod_{z=1}^{Z} P(\phi_z|\beta) d\Phi \\ &= \int \prod_{z=1}^{Z} \prod_{a=1}^{A} \phi_{z,a}^{C_{z,a}} \prod_{z=1}^{Z} \left(\frac{\Gamma(A\beta)}{\Gamma(\beta)^A} \prod_{a=1}^{A} \phi_{z,a}^{\beta-1} \right) d\Phi \\ &\propto \prod_{z=1}^{Z} \frac{\prod_{a=1}^{A} \Gamma(C_{z,a} + \beta)}{\Gamma(\sum_{a=1}^{A} (C_{z,a} + \beta))} \end{aligned} \tag{31}$$

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

$$\begin{aligned} P(\mathbf{t}|\mathbf{z}, \mathbf{a}, \gamma) &= \int P(\mathbf{t}|\mathbf{z}, \mathbf{a}, \Xi) P(\Xi|\gamma) d\Xi \\ &= \int \prod_{d=1}^{D} \prod_{i=1}^{N_d^e} P(e_i.t|\xi_{z_i,e_i.a}) \prod_{z=1}^{Z} \prod_{a=1}^{A} P(\xi_{z,a}|\gamma) d\Xi \\ &= \int \prod_{z=1}^{Z} \prod_{a=1}^{A} \left(\prod_{t=1}^{T} \xi_{z,a,t}^{C_{z,a,t}} \frac{\Gamma(T\gamma)}{\Gamma(\gamma)^T} \prod_{t=1}^{T} \xi_{z,a,t}^{\gamma-1} \right) d\Xi \\ &\propto \prod_{z=1}^{Z} \prod_{a=1}^{A} \frac{\prod_{t=1}^{T} \Gamma(C_{z,a,t} + \gamma)}{\Gamma(\sum_{t=1}^{T} (C_{z,a,t} + \gamma))} \end{aligned} \tag{32}$$

Substituting Eqs. (28)–(32) into Eq. (27), and using the chain rule and $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$, we can obtain the conditional probability conveniently,

$$\begin{aligned} & P(z_{d,i} = z|\mathbf{f}, \mathbf{v}, \mathbf{a}, \mathbf{t}, \mathbf{z}_d^{-i}, \alpha, \beta, \gamma, \eta, \iota, \mathbf{g}, \mathbf{k}, \mathbf{x}, \mathbf{y}) \\ &= \begin{cases} \frac{P(z_{d,i}, f_{d,i}, v_{d,i}|\mathbf{f}_d^{-i}, \mathbf{v}_d^{-i}, \mathbf{z}_d^{-i}, \alpha, \eta, \iota, \mathbf{g}, \mathbf{k}, \mathbf{x}, \mathbf{y})}{P(f_{d,i}, v_{d,i}|\mathbf{f}_d^{-i}, \mathbf{v}_d^{-i}, \mathbf{z}_d^{-i}, \alpha, \eta, \iota, \mathbf{g}, \mathbf{k}, \mathbf{x}, \mathbf{y})} & : i \text{ is a patient feature} \\ \frac{P(z_{d,i}, p_{d,i}.a_{d,i}.t_{d,i}|\mathbf{a}_d^{-i}, \mathbf{t}_d^{-i}, \mathbf{z}_d^{-i}, \alpha, \beta, \gamma)}{P(a_{d,i}.t_{d,i}|\mathbf{a}_d^{-i}, \mathbf{t}_d^{-i}, \mathbf{z}_d^{-i}, \alpha, \beta, \gamma)} & : i \text{ is a clinical event} \end{cases} \\ &\propto \begin{cases} \frac{P(\mathbf{z}, \mathbf{f}, \mathbf{v}|\alpha, \eta, \iota, \mathbf{g}, \mathbf{k}, \mathbf{x}, \mathbf{y})}{P(\mathbf{z}_d^{-i}, \mathbf{f}_d^{-i}, \mathbf{v}_d^{-i}|\alpha, \eta, \iota, \mathbf{g}, \mathbf{k}, \mathbf{x}, \mathbf{y})} & : i \text{ is a patient feature} \\ \frac{P(\mathbf{z}, \mathbf{a}, \mathbf{t}|\alpha, \beta, \gamma)}{P(\mathbf{z}_d^{-i}, \mathbf{a}_d^{-i}, \mathbf{t}_d^{-i}|\alpha, \beta, \gamma)} & : i \text{ is a clinical event} \end{cases} \\ &\propto \frac{C_{z,d}^{-i} + \alpha}{C_{d,*}^{-i} + \alpha Z} \cdot \begin{cases} \frac{C_{z,f_i}^{-i} + \eta}{C_{z,*}^{-i} + \eta F} \cdot \begin{cases} \frac{C_{z,f_i,v_i}^{-i} + \iota}{C_{z,f_i,*}^{-i} + V_{z,f_i}\iota} & : f_i \text{ is a categorical feature} \\ \frac{\Gamma(x_{z,f_i})}{\Gamma(x_{z,f_i}^{-i})} \cdot \frac{y_{z,f_i}^{-i\ x_{z,f_i}^{-i}}}{y_{z,f_i}^{x_{z,f_i}}} \cdot \left(\frac{k_{z,f_i}^{-i}}{k_{z,f_i}}\right)^{\frac{1}{2}} & : f_i \text{ is a numerical feature} \end{cases} & : i \text{ is a patient feature} \\ \frac{C_{z,a}^{-i} + \beta}{C_{z,*}^{-i} + A\beta} \cdot \frac{C_{z,a,t}^{-i} + \gamma}{C_{z,a,*}^{-i} + T\gamma} & : i \text{ is a clinical event} \end{cases} \end{aligned} \tag{33}$$

References

Agrawal R, Gunopulos D, Leymann F (1998) Mining process models from workflow logs. In HJ Schek, F Saltor, I Ramos, G Alonso (eds) Sixth international conference on extending database technology. Springer-Verlag, London, pp 469–483

Antman EM, Cohen M, Bernink PM et al (2000) The TIMI risk score for Unstable Angina/Non-ST elevation MI: a method for prognostication and therapeutic decision making. J Am Med Assoc 284(7):835–842

Blei DM, Ng AY, Jordan MI (March 2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022

Bouarfa L, Dankelman J (2012) Workflow mining and outlier detection from clinical activity logs. J Biomed Inform 45(6):1185–1190

Cheah J (2000) Development and implementation of a clinical pathway programme in an acute care general hospital in singapore. Int J Qual Health Care 12:403–412

Cook JE, Wolf AL (1998) Discovering models of software processes from event-based data. ACM Transactions on Software Engineering and Methodology 7(3):215–249

Dong W, Huang Z, Ji L, Li H (2014) A genetic fuzzy system for unstable angina risk assessment. BMC Med Inform Decis Mak 14:12

Dunn AG, Ong MS, Westbrook JI, Magrabi F, Coiera E, Wobcke W (2011) A simulation framework for mapping risks in clinical processes: the case of in-patient transfers. J Am Med Inform Assoc 18(3):259–266

Dy SM, Garg P, Nyberg D, Dawson PB, Pronovost PJ, Morlock L, Rubin H, Wu AW (2005) Critical pathway effectiveness: assessing the impact of patient, hospital care, and pathway characteristics using qualitative comparative analysis. Health Serv Res 40(2):499–516

Elson RB, Faughnan JG, Connelly DP (1997) An industrial process view of information delivery to support clinical decision making: implications for systems design and process measures. J Am Med Inform Assoc 4(4):266–278

Ghattas J, Peleg M, Soffer P, Denekamp Y (2010) Learning the context of a clinical process. In: Stefanie R-M, Shazia S, Leymann F (eds) Business process management workshops, vol 43. Lecture Notes in Business Information Processing. Springer, Berlin, pp 545–556

Gooch P, Roudsari A (2011) Computerization of workflows, guidelines, and care pathways: a review of implementation challenges for process-oriented health information systems. J Am Med Inform Assoc 18(6):738–748

Griffiths TL (2004) Finding scientific topics. Proc Natl Acad Sci USA 101:5228–5235

Huang Z, Lu X, Gan C, Duan H (2011) Variation prediction in clinical processes. In: Peleg M, Lavrac N, Combi C (eds) Artificial intelligence in medicine, vol 6747., Lecture notes in Computer ScienceSpringer, Berlin/Heidelberg, pp 286–295

Huang Z, Lu X, Duan H (2012) Using recommendation to support adaptive clinical pathways. J Med Syst 36(3):1849–1860

Huang Z, Lu X, Duan H (2012) On mining clinical pathway patterns from medical behaviors. Artif Intell Med 56(1):35–50

Huang Z, Juarez JM, Duan H, Li H (2013) Length of stay prediction for clinical treatment process using temporal similarity. Expert Syst Appl 40(16):6330–6339

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

Huang Z, Lu X, Duan H, Fan W (2013) Summarizing clinical pathways from event logs. J Biomed Inform 46(1):111–127

Huang Z, Dong W, Duan H, Li H (2014) Similarity measure between patient traces for clinical pathway analysis: problem, method, and applications. IEEE J Biomed Health Inform 18(1):4–14

Huang Z, Dong W, Ji L, Gan C, Lu X, Duan H (2014) Discovery of clinical pathway patterns from event logs using probabilistic topic models. J Biomed Inform 47:39–57

Huang Z, Lu X, Duan H (2012) Anomaly detection in clinical processes. In AMIA Annu Symp Proc, pp 370–379

Hunter B, Segrott J (2008) Re-mappling client journeys and professional identities: a review of the literature on clinical pathways. Int J Nurs Stud 45:608–625

Iwata T, Sawada H (2013) Topic model for analyzing purchase data with price information. Data Min Knowl Discov 26(3):559–573

Lakshmanan GT, Rozsnyai S, Wang F (2013) Investigating clinical care pathways correlated with outcomes. In: Daniel F, Wang J, Weber B (eds) Business process management, vol 8094. Lecture Notes in Computer Science.Springer, Berlin, pp 323–338

Lang M, Burkle TB, Laumann S, Prokosch HU (2008) Process mining for clinical workflows: challenges and current limitations. In SK Andersen, GO Klein, S Schulz, J Aarts (eds) Proceedings of MIE2008 the XXIst international congress of the European federation for medical informatics, pp 229–234

Lenz R, Blaser R, Beyer M, Heger O, Biber C et al (2007) IT support for clinical pathways-lessons learned. Int J Med Inform 76(3):S397–S402

Lenz R, Reichert M (2007) IT support for healthcare processes-premises, challenges, perspectives. Data Knowl Eng 61(1):39–58

Lin F, Chen S, Pan S, Chen Y (2001) Mining time dependency patterns in clinical pathways. Int J Med Inform 62(1):11–25

Loeb M, Carusone SC, Goeree R, Walter SD, Brazil K, Krueger P et al (2006) Effect of a clinical pathway to reduce hospitalizations in nursing home residents with pneumonia. J Am Med Assoc 295: 2503–2510

Lu X, Huang Z, Duan H (2012) Supporting adaptive clinical treatment processes through recommendations. Comput Methods Programs Biomed 107(3):413–424
Mans R, Schonenberg H, Leonardi G, Panzarasa S, Cavallini A, Quaglini S (2008) Process mining techniques: an application to stroke care. Stud Health Technol Inform 136:573–578
Peleg M, Mulyar N, van der Aalst WMP (2012) Pattern-based analysis of computer-interpretable guidelines: don't forget the context. Artif Intell Med 54(1):73–74
Peleg M (2013) Computer-interpretable clinical guidelines: a methodological review. J Biomed Inform 46(4):744–763
Peleg M, Soffer P, Ghattas J (2008) Mining process execution and outcomes—position paper. In: Arthur H, Benatallah B, Paik H-Y (eds) Business process management workshops, vol 4928. Lecture Notes in Computer Science. Springer, Berlin, pp 395–400
Phung D, Adams B, Venkatesh S, Kumar M (2009) Unsupervised context detection using wireless signals. Pervasive Mobile Comput 5(6):714–733
Quaglini S, Stefanelli M, Lanzola G, Caporusso V, Panzarasa S (2001) Flexible guideline-based patient careflow systems. Artif Intell Med 22(1):65–80
Rebuge A, Ferreira DR (2012) Business process analysis in healthcare environments: a methodology based on process mining. Inform Syst 37(2):99–116
Renholm M, Leino-Kilpi H, Suominen T (2002) Critical pathways: a systematic review. J Nurs Adm 32(4):196–202
Rosen-Zvi M, Griffiths T, Steyvers M, Smyth P (2004) The author-topic model for authors and documents. In 20th conference on uncertainty in artificial intelligence, pp 487–494
Rotter T, Kugler J, Koch R, Gothe H, Twork S, van Oostrum JM, Steyerberg EW (2008) A systematic review and meta-analysis of the effects of clinical pathways on length of stay, hospital costs and patient outcomes. BMC Health Serv Res 8:265
Tsoumakas G, Katakis I (2007) Multi-label classification: an overview. Int J Data Warehous Min 3(3):1–13
Uzark K (2003) Clinical pathways for monitoring and advancing congenital heart disease care. Progr Pediatr Cardiol 18:131–139
Wakamiya S, Yamauchi K (2009) What are the standard functions of electronic clinical pathways? Int J Med Inform 78(8):543–550
Wang X, McCallum A, Wei X (2007) Topical n-grams: phrase and topic discovery, with an application to information retrieval. In IEEE international conference on data mining, pp 697–702
Wang F, Zhang P, Cao N, Hu J, Sorrentino R (2014) Exploring the associations between drug side-effects and therapeutic indications. J Biomed Inform. doi:10.1016/j.jbi.2014.03.014
Weiland DE (1997) Why use clinical pathways rather than practice guidelines? Am J Surg 174:592–595
2012 Writing Committee Members, Jneid H, Anderson JL, Wright RS, Adams CD, Bridges CR, Casey DE, Ettinger SM, Fesmire FM, Ganiats TG, Lincoff AM, Peterson ED, Philippides GJ, Theroux P, Wenger NK, Zidar JP (2012) 2012 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 126(7):875–910
Whye Teh Y, Jordan MI, Beal MJ, Blei DM (2004) Hierarchical Dirichlet processes. J Am Stat Assoc 101(476):1566–1581
Yao W, Kumar A (2013) Conflexflow: integrating flexible clinical pathways into clinical decision support systems using context and rules. Decis Support Syst 55(2):499–515
Zand DJ, Brown KM, Konecki UL, Campbell JK, Salehi V, Chamberlain JM (2008) Effectiveness of a clinical pathway for the emergency treatment of patients with inborn errors of metabolism. Pediatrics 122:1191–1195