Journal of Biomedical Informatics 59 (2016) 227–239

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Incorporating comorbidities into latent treatment pattern mining for clinical pathways

Zhengxing Huang [a,d,*], Wei Dong [b], Lei Ji [c], Chunhua He [d], Huilong Duan [a]

[a] *College of Biomedical Engineering and Instrument Science, Zhejiang University, China*
[b] *Department of Cardiology, Chinese PLA General Hospital, China*
[c] *IT Department, Chinese PLA General Hospital, China*
[d] *College of Medical Engineering Technology, Xinjiang Medical University, China*

ARTICLE INFO

Article history:
Received 28 August 2015
Revised 29 November 2015
Accepted 15 December 2015
Available online 21 December 2015

Keywords:
Clinical pathway
Treatment pattern mining
Latent Dirichlet allocation
Comorbidity

ABSTRACT

In healthcare organizational settings, the design of a clinical pathway (CP) is challenging since patients following a particular pathway may have not only one single first-diagnosis but also several typical comorbidities, and thus it requires different disciplines involved to put together their partial knowledge about the overall pathway. Although many data mining techniques have been proposed to discover latent treatment information for CP analysis and reconstruction from a large volume of clinical data, they are specific to extract nontrivial information about the therapy and treatment of the first-diagnosis. The influence of comorbidities on adopting essential treatments is crucial for a pathway but has seldom been explored. This study proposes to extract latent treatment patterns that characterize essential treatments for both first-diagnosis and typical comorbidities from the execution data of a pathway. In particular, we propose a generative statistical model to extract underlying treatment patterns, unveil the latent associations between diagnosis labels (including both first-diagnosis and comorbidities) and treatments, and compute the contribution of comorbidities in these patterns. The proposed model extends latent Dirichlet allocation with an additional layer for diagnosis modeling. It first generates a set of latent treatment patterns from diagnosis labels, followed by sampling treatments from each pattern. We verify the effectiveness of the proposed model on a real clinical dataset containing 12,120 patient traces, which pertain to the unstable angina CP. Three treatment patterns are discovered from data, indicating latent correlations between comorbidities and treatments in the pathway. In addition, a possible medical application in terms of treatment recommendation is provided to illustrate the potential of the proposed model. Experimental results indicate that our approach can discover not only meaningful latent treatment patterns exhibiting comorbidity focus, but also implicit changes of treatments of first-diagnosis due to the incorporation of typical comorbidities potentially.

1. Introduction

Although overhyped, and sharply criticized repeatedly over the past 20 years, clinical pathways (CPs) have remained on the agenda of many healthcare organizations [1,2]. The simple reason is that CPs, as defined set of therapy and treatment activities that represent the steps required to achieve a specific treatment objective for a particular disease, are recognized as one of the most useful tools to increase the quality of care services in an unfavorable economic scenario and under the financial pressure by governments [1–5].

Since 2010, Chinese government has published over 200 CP specifications, which are required to be implemented in more than one hundred pilot hospitals [6,7]. Each CP specification is a standardized/normalized treatment guideline of a particular disease, and is assumed to provide actionable knowledge to schedule the best practice for individual patients in their hospitalizations [5]. However, actual treatment activities are extremely complex, with numerous variations across various stages in the execution of CPs, and they often bear no relation to the ideal as envisaged by the designers of CPs [8–10]. In particular, since a CP specification is generally designed for a single first-diagnosis, the treatments on the typical comorbidities are easily neglected in the pathway design phrase [11]. For example, the unstable angina CP describes the guideline-conform and consented management of unstable angina patients with ST-segment elevation infarct,

[*] Corresponding author at: College of Biomedical Engineering and Instrument Science, Zhejiang University, China.

E-mail address: zhengxinghuang@zju.edu.cn (Z. Huang).

http://dx.doi.org/10.1016/j.jbi.2015.12.012
1532-0464/© 2015 Elsevier Inc. All rights reserved.

non-ST-segment elevation infarct, and Troponin negative unstable angina [12]. However, treatment interventions for typical comorbidities of unstable angina, e.g., *Diabetes*, *Hypertensions*, etc., are not well explained in the pathway specification. Note that there are more than 70 comorbidities of unstable angina patients, and many comorbidities (e.g., *Kidney deficiency*) afflict the unstable angina pathway seriously, and incur deviations from the predefined CP specification [13]. According to a survey [39] we have performed at the Cardiology department of Chinese PLA general hospital, the compliance rate of the unstable angina pathway is about 7.2%, which is significantly below the well-recognized threshold value (i.e., 40%,) for the meaningful use of CPs [14]. Clinicians indicate that the main reason for this is that the predesigned unstable angina pathway specification lacks the support for the treatments of comorbidities and complications [1,15–17]. As a result, patient care-flow may not go well towards the expected direction, and there are deviations from the pre-defined CP specifications [18,38].

To design feasible CPs, it is increasingly important to study actual treatments between a particular first-diagnosis and its common comorbidities to extract nontrivial knowledge in CPs, and to help clinical analysts to redesign/refine CPs. Recently, the fast development of healthcare information systems produces a large volume of electronic clinical data to record actual execution situations in CPs and thus provides a comprehensive source for exploratory analysis and statistics to benefit many real applications for CP analysis and redesign [19,20]. In fact, many data mining and machine learning techniques have been proposed to utilize these execution data to extract useful information for CP analysis and redesign [20,21]. This, also called process mining, has been recognized as an objective way of analyzing CPs as it is not biased by perceptions or normative behaviors [4,22]. In particular, it can provide insight about what is actually happening, and ultimately the knowledge extracted from clinical data can be used for effective improvement of clinical pathways and of their supporting systems [22].

However, most existing process mining approaches analyze treatment behaviors from the first-diagnosis perspective, and seldom studied the influence of typical comorbidities on treatment adoption in CPs. Note that, in clinical practice, the different aspects of treatment information are highly correlated, and clinicians are very interested in the latent associations between comorbidities and treatments. In particular, they wonder what kind of treatment activities should be given to patients who have some specific comorbidities. And how will the different comorbidities influence individual patients' treatment processes? Etc.

A straightforward method to address this problem is to manually build a dictionary of treatment events for each typical comorbidity. But building such a treatment dictionary is not only labor consuming, but also unable to quantize the connection strengths between treatments and comorbidities. As an alternative, discriminative methods, such as Decision Tree, and SVM, can provide a principled way to estimate comorbidity-treatment associations using their concurrence counts in a clinical dataset [20]. However, since such approaches treat each treatment event individually, most relevant treatments extracted for each comorbidity are usually mixed with background and the first-diagnosis-oriented treatment events. Moreover, it is more sensible to associate comorbidities with a specific comorbidity-incorporated treatment pattern instead of only a single treatment event.

In this study, we explore the problem of comorbidity-incorporated treatment pattern discovery using the execution data of CPs. In particular, we proposed a generative statistical model named Diagnosis Treatment Model (DTM) to mine underlying treatment patterns from data. The proposed DTM, as an extension of Latent Dirichlet Allocation (LDA) [23], can be applied to the task of comorbidity-incorporated treatment pattern mining, and allows us to infer the contributions of comorbidities on the treatments adoption in CPs [19,24,25]. It first generates a treatment pattern from a multinomial distribution conditioned on diagnosis labels, and then generates treatment activities and their occurring time stamps from other multinomial distributions based on latent treatment patterns. As a complete generative model, the proposed model is capable of distinguishing between different therapy purposes of the same treatment activity, and discovering meaningful treatment patterns resulted from typical comorbidities, since the patient care-flow with specific comorbidities is characterized by its latent treatment patterns rather than individual treatment activities. We test the proposed model on a real clinical dataset gathered from Chinese PLA General Hospital. The remainder of this paper is organized as follows: Section 2 discusses related work; Section 3 presents the proposed model for comorbidity-incorporated latent treatment pattern mining. The dataset, experimental results and discussions are illustrated in Section 4. Finally, we draw conclusions in Section 5.

2. Related work

The most related direction to our work is healthcare process mining/analysis [22,26,27]. Process mining, as a valuable set of techniques, has been widely studied in business process management domain [26]. It uses process event logs to record business process execution information, to mine the actual behaviors in business processes, and discover business process patterns [22]. Based on process event logs, with its logic and reasoning ability, process mining guarantees integrity, objectivity and universality of the discovered process patterns [22].

Process mining techniques have gained gradual attentions in the healthcare domain, and has already been attempted by some researchers [22,28–32,36,37]. Work that is closely related to ours is presented in 2013, in which Lakshmanan et al., presented a hybrid approach for mining CPs correlated with patient outcomes that involves a combination of clustering, process mining and frequent pattern mining [33]. Their work takes clinical outcomes into account in mining treatment processes. In particular, they describe an algorithm to mine the structure of a clinical pathway using frequent pattern mining, and rank the frequent patterns according to the degree of their correlation with a patient outcome. In this sense, an individual clinical pathway mined from real patient event data in the form of a frequent pattern may contain a small subset of the overall possible set of events that could be applied to treat a particular disease. As valuable as their approach, the comorbidity-oriented issues, such as the correlations between the mined patterns and comorbidities, and the implicit changes of treatments of the first-diagnosis due to the incorporation of typical comorbidities, were not addressed in their work.

The use of traditional process mining techniques though successful in discovering latent treatment knowledge can prove inadequate in CP analysis [20–22]. Note that the complexity, dynamic, and ad hoc natures of CPs are far higher than that of common business processes [22]. Using the traditional process mining techniques, it may generate incomprehensible and even spaghetti-like treatment patterns [21,24]. Note that CPs deal with a variety of medical problems. Therefore, it can be assumed that a patient trace is actually guided by multiple underlying treatment patterns [19,25]. For example, a patient who follows the unstable angina CP may also be performed specific activities for his/her diabetes treatments. To this end, it is advocated to develop new process mining techniques facilitating CP analysis. In our previous work, we developed a generative topic model to derive latent treatment patterns hidden in clinical data [19]. Discovered treatment patterns, as actionable knowledge representing the best practice for most patients in most time of their treatment journeys, 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 redesign [20].

Despite the success of previous work on healthcare process mining and analysis, existing approaches usually discover underlying treatment patterns for the therapy of first-diagnosis, such that the influence of typical comorbidities in CPs seems to be overlooked in the literature. This also prevents clinical analysts from further understanding the contributions of typical comorbidities on the therapy and treatment in CPs, because the comorbidities may not only lead to newly and unexpected comorbidity-oriented treatments but also change the original treatment behaviors w.r.t the first-diagnosis in CPs. In fact, it is no wonder that patients with a specific type of comorbidity may have a different treatment strategy other than the patients without that comorbidity in clinical practice. In this sense, comorbidity-incorporated treatment pattern mining is crucial for CP analysis and redesign.

3. Methodology

In this section, we present a generative statistical model for comorbidity-incorporated treatment pattern mining, namely, the Diagnosis Treatment Model (DTM). The problem is firstly defined. Then, we present our model in detail.

3.1. Problem definition

In this study, our goal is to extract latent associations between comorbidities and treatments about a particular CP from the execution data. These execution data are also called clinical event log, which conceals an untapped reservoir of knowledge about the way of specific therapy and treatment activities being performed on particular patients in their hospitalizations. Formally, let A be the treatment activity domain, and T the time domain, and C the disease type domain. We assume that a clinical event log L consists of $|L|$ patient traces. Each patient trace σ ($\sigma \in L$) includes both a set of diagnosis labels $C_\sigma \subseteq C$, and a finite non-empty set of treatment events $< e_1, e_2, \ldots, e_{|\sigma|}\}$ performed on a particular patient during his/her hospitalization. A treatment event e_i is represented as $e_i = (a_i, t_i)$, where a_i is the activity type of e_i ($a_i \in A$), and t_i is the occurring time of e_i ($t_i \in T$). For convenience, let $e_i \cdot a$ and $e_i \cdot t$ be the treatment activity type and the occurring time stamp of e_i, respectively. A treatment event is a performed treatment activity at a particular time stamp.

For example, Fig. 1 shows a clinical event log example consisting of 5 patient traces. The predefined disease label domain contains one first-diagnosis ("*Unstable angina*"), and seven typical comorbidities (i.e., "*Ulcer*", "*Anterior myocardial infraction*", "*One degree atrioventricular block*", "*Carotid atherosclerosis*", "*Anemia*", "*Diabetes*", and "*Hypertension*"). Thus, $|L|$ and $|C|$ respectively equal to 5 and 7. Each patient trace is a set of clinical events and has been assigned particular disease labels: C_{σ_1} = {"*Unstable angina*", "*Ulcer*"} and C_{σ_2} = {"*Unstable angina*", "*Anterior myocardial infraction*"}, etc. Our objective is to accurately model the associations between diagnosis labels and treatment events, and improve the performance of latent treatment pattern discovery by taking the first-diagnosis and typical comorbidities into account. Table 1 summarizes the notations of these frequently-used variables.

In our previous work, we have presented a generative statistic model, i.e., Clinical Pathway Model (CPM), to model the whole

A0:Admission	A14:Routine blood test	A28:Anti-arrhythmic examination	A42:Diuretic
A1:ECG examination	A15:Transfer	A29:Local anesthesia	A43:Puncture
A2:Ultrasonography examination	A16:Coronary angiography	A30:Diabetes examination	A44:Peripheral vasodilator
A3:Cardiovascular treatment	A17:Discharge	A31:Thyroid function tests	A45:Plasma concentration
A4:β-adrenergic receptor blockers	A18:Radiation	A32:Anesthesia	A46:CT examination
A5:Antianginal treatment	A19:Occult blood test	A33:H2 receptor antagonist	A47:Tumor markers check
A6:Anticoagulation treatment	A20:Stent implantation	A34:Renal arteriography	A48:First-level care
A7:Antiplatelet treatment	A21:PTCA	A35:Angiotensin receptor antagonists	A49:Routine care
A8: Biochemical examination	A22:Angiotensin-converting enzyme inhibitors	A36:Antihypertensive treatment	A50:Specific meal
A9:Routine urine test	A23:Lipid regulation	A37:Blood sugar regulation	A51:Second-level care
A10:Routine stool test	A24:Coagulation examination	A38:Troponin T	A52:Oxygen inhalation
A11:Thrombosis examination	A25:Proton pump inhibitors	A39:Consultation	A53:Monitoring of blood oxygen
A12:Blood typing	A26:X-ray	A40:Anti-diabetes treatment	A54:Multifunctional monitors
A13:Blood serum test	A27:Calcium channel blockers	A41:Calcium regulation	

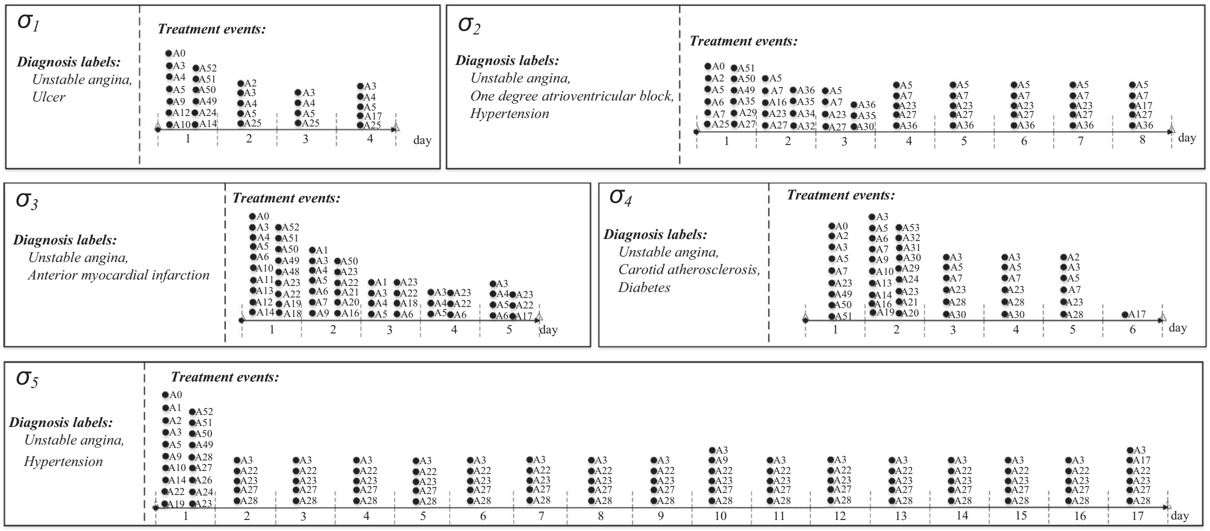

Fig. 1. A clinical event log example.

Table 1
Notations of frequently-used variables in this paper.

Symbol	Description
L	Clinical event log
C	Diagnosis labels in L
σ	Patient trace in L
N_σ	Number of treatment events in patient trace σ
C_σ	Set of diagnosis labels in patient trace σ
Z	Universe of treatment patterns
A	The domain of treatment activity types in L
T	The domain of occurring time stamps of treatment events in L
$e_{\sigma,i}$	The ith treatment event in patient trace σ, i.e., $e_{\sigma,i} = \langle a, t \rangle$, where $a \in A$, and $t \in T$
α	Dirichlet prior of θ
β	Dirichlet prior of ϕ
γ	Dirichlet prior of φ
θ	Multinomial distribution of treatment patterns given the diagnosis labels in L
ϕ	Multinomial distribution of treatment activity types to treatment patterns in L
φ	Multinomial distribution of occurring time stamps of treatment activities to treatment patterns in L

event log by Z latent treatment patterns [24]. As shown in Fig. 2(A), the notations θ_σ, ϕ, and φ are used to present the patient trace-treatment pattern, treatment pattern-activity, and treatment pattern-activity-time stamp distributions, respectively. The hyperparameters are denoted by α, β, and γ. In particular, the hyperparameter α is a Dirichlet prior of θ, which can be interpreted as the prior observation counts for the number of times the topic was sampled from patient trace before any treatment event is observed. The hyperparameter β is a Dirichlet prior of ϕ, which can be treated as the prior observation counts for the number of times treatment events with particular activity types were sampled from treatment pattern before any actual clinical event has been observed. The hyperparameter γ is a Dirichlet prior of φ, which can be interpreted as the prior observation counts for the number of occurring time stamps of clinical events with particular activity types were sampled from treatment patterns before any treatment event has been observed. While CPM utilizes the contextual information within the patient traces, it fails to utilize the diagnosis information to guide the treatment pattern generation.

This work can be seen as building on earlier ideas from our previous work in probabilistic modeling for latent treatment pattern mining in CPs. We describe in the next subsection the DTM. As a complete generative model, the proposed DTM allows us to associate each treatment pattern with both treatment events and diagnosis labels jointly, and to infer the influence of particular comorbidities on the therapy and treatment of patients in CPs.

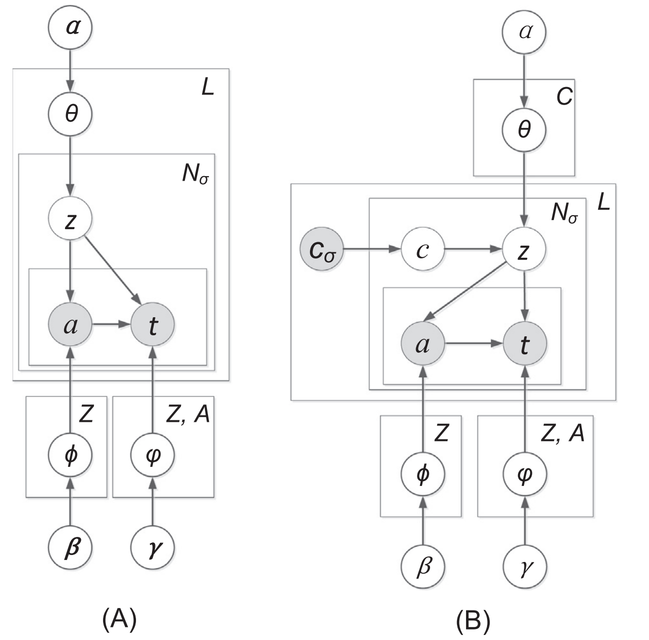

Fig. 2. (A) Clinical Pathway Model (CPM) [24], and (B) Diagnosis Treatment Model (DTM).

3.2. Diagnosis-treatment model

CPM discloses the underlying treatment patterns in patient traces of a clinical event log. However, CPM does not identify comorbidities of a patient trace nor the association of comorbidities to each treatment pattern of a patient trace. To this end, we propose the Diagnosis Treatment model, an extension of CPM, that models the treatment events of a patient trace and the contributions of both first-diagnosis and typical comorbidities. As illustrated in Fig. 2(B), the proposed DTM can also associate treatment events and diagnosis labels jointly, and infer the contributions of comorbidities on treatment behaviors of patient traces. The generative process for DTM is described as follows:

1. For each diagnosis label c, choose $|C|$ multinomials $\theta_c \sim Dir(\alpha)$.
2. For each patient trace σ in the event log L.
    (1) For each clinical event $e_{\sigma,i}$ in σ.
        (a) Choose a diagnosis label $c_i \sim Uniform(c_\sigma)$.
        (b) Choose a treatment pattern $z_i \sim Multinomial(\theta_c)$.
        (c) Choose a treatment activity type $e_{\sigma,i} \cdot a \sim Multinomial(\phi_z)$.
        (d) Choose an occurring time stamp $e_{\sigma,i} \cdot t \sim Multinomial(\varphi_{z,a})$.

The procedure begins by choosing a diagnosis label c, randomly at uniform, form the set of diagnosis labels c_σ of σ. Afterward, the multinomial distribution θ_c, from the Dirichlet distribution α, is picked, and this distribution determines which treatment patterns are most likely to be assigned to the diagnosis label c. Next, a single treatment pattern z is sampled for each treatment event e_i in σ, from the multinomial distribution θ_c associated with the diagnosis label c for that event. Finally, the event type $e_i \cdot a$ and the occurring time stamp $e_{io} \cdot t$ of e_i are generated based on the multinomial distribution ϕ_z, and $\varphi_{z,a}$, respectively, where ϕ_z is generated from the Dirichlet distribution β for each treatment pattern z, and $\varphi_{z,a}$ is generated from the Dirichlet distribution γ form each treatment pattern z and each treatment activity type a.

The joint probability of all the random variables for a clinical event log is as follows:

$$P(\mathbf{e}, \mathbf{c}, \mathbf{z}, \theta, \phi, \varphi | C, \alpha, \beta, \gamma) = P(\theta|\alpha)P(\phi|\beta)P(\varphi|\gamma)P(\mathbf{z}|\alpha, \mathbf{c})P(\mathbf{c}|C) \\ P(\mathbf{e} \cdot \mathbf{a}|z, \phi)P(\mathbf{e} \cdot \mathbf{t}|z, \mathbf{e} \cdot \mathbf{a}, \varphi). \quad (1)$$

Since it is intractable to perform an exact inference, we use the approximate inference based on Gibbs sampling to estimate the parameters [34]. Specifically, for each treatment event, we estimate the posterior distribution on diagnosis label c and treatment pattern z based on the following conditional probabilities, which can be derived by marginalizing the above joint probabilities in Eq. (1).

$$P(z_i = z, c_i = c|\boldsymbol{e}, \boldsymbol{z}_{-i}, c_{-i}, C, \alpha, \beta, \gamma) \propto \frac{n_{c_{-i},z} + \alpha}{\sum_{z \in Z} n_{c_{-i},z} + |Z|\alpha} \times \frac{n_{z,e_{-i} \cdot a} + \beta}{\sum_{a' \in A} n_{z,e_{-i} \cdot a'} + |A|\beta} \times \frac{n_{z,e_{-i} \cdot a,e_{-i} \cdot t} + \gamma}{\sum_{t' \in T} n_{z,e_{-i} \cdot a,e_{-i} \cdot t'} + |T|\gamma}, \qquad (2)$$

where z_i and c_i are the candidate treatment pattern and diagnosis label that e_i is assigned to, $\boldsymbol{z}_{-i}$ refers to the pattern assignments of all other treatment events, c_{-i} refers to the diagnosis label assignments of all other treatment events, $n_{c,z}$ is the number of instances of treatment pattern z that has been assigned to diagnosis label c, $n_{z,e \cdot a}$ is the number of instances of treatment activity type $e \cdot a$ that has been assigned to treatment pattern z, and $n_{z,e \cdot a,e \cdot t}$ is the number of instances of the occurring time stamp $e \cdot t$ of the activity type $e \cdot a$ that has been assigned to treatment pattern z. The suffix $-i$ means the number that does not include the current assignment of treatment event e_i.

With the sampled treatment patterns available, it is easy to estimate the probability of treatment pattern z conditioned on diagnosis label c, the probability of activity type a conditioned on treatment pattern z, and the probability of occurring time stamp t conditioned on treatment pattern z and clinical activity type a, by the following equations:

$$\theta_{c,z} = \frac{n_{c,z} + \alpha}{\sum_{z' \in Z} n_{c,z'} + |Z|\alpha}, \qquad (3)$$

$$\phi_{z,a} = \frac{n_{z,a} + \beta}{\sum_{a' \in A} n_{z,a'} + |A|\beta}, \qquad (4)$$

$$\varphi_{z,a,t} = \frac{n_{z,a,t} + \gamma}{\sum_{t' \in T} n_{z,a,t'} + |T|\gamma}, \qquad (5)$$

With all the parameters derived above, the proposed DTM can be utilized for various applications. For example, associations between diagnosis labels (including both first-diagnosis and comorbidities) 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 stamps t (a.k.a a treatment event) given a set of comorbidities of a specific particular patient in his/her treatment journey:

$$P(a,t|C_\sigma) = \sum_{c \in C_\sigma} P(a,t|c) = \sum_{c \in C_\sigma} \sum_{z \in Z} P(t|a,z)P(a|z)P(z|c) = \sum_{c \in C_\sigma} \sum_{z \in Z} \varphi_{z,a,t} \phi_{z,a} \theta_{c,z}. \qquad (6)$$

The worst case time complexity of each iteration of the Gibbs sampler is $O(ZATC_{max})$, where Z is the number of treatment patterns, A is the number of activity types, T is the number of occurring time stamps, and C_{max} is the maximum number of diagnosis labels that can be associated with a single patient case. As the complexity is linear in AT, Gibbs sampling can be efficiently carried out on large data sets.

4. Experiments

In this section, we firstly evaluate the performance of the proposed model on comorbidity-incorporated treatment pattern mining in a quantificational manner. The experiments are, firstly, designed to analyze the influence of latent treatment pattern numbers. We then conduct a qualitative investigation on the generated treatment patterns using the proposed models, to study the effect of the proposed model on mining latent associations between comorbidities and treatments.

4.1. Experiment design

To test the effectiveness of the proposed model, we have collected an experimental log from the cardiology department of Chinese PLA General Hospital. The CP of unstable angina is selected in this case study. Unstable angina is a kind of chest discomfort or pain that occurs in a continuous and unpredictable way. The cause of angina is commonly the poor blood flow in coronary vessels caused by atherosclerosis and the lack of oxygen supply to the myocardium. While the risk of unstable angina is high, the population of unstable angina is huge, especially for aged people and those with associated disease such as hypertension and diabetes [12,13]. Thus, the discovery of latent associations between comorbidities and treatments and the underlying treatment patterns in the unstable angina CP will be of significant value and interest.

In this case study, 12,120 patient traces following the unstable angina CP were selected from the cardiology department to demonstrate the ability of the proposed method to discover comorbidity-incorporated treatment patterns. Each patient trace in the experimental log consists of one or a set of diagnosis types including the first diagnosis "*Unstable angina*" and a set of comorbidities, e.g., *Hypertension*, *Diabetes*, etc. Table 2 summarizes the statistics of the collected event log.

In contrast to the clinical datasets typically utilized in research, the event logs collected from real clinical settings contain not only tens of thousands of treatment events, but also hundreds of diagnosis labels indicating various comorbidities of patients following particular pathways. In addition, the frequencies of diagnosis label in clinical event logs tend to have highly skewed frequency-distributions with power-law statistics. Fig. 3 illustrates this point for a real clinical event log pertaining to the unstable angina CP. It contains 80 unique comorbidities, and the total number of diagnosis labels is plotted as a function of label-frequency on a log–log scale (i.e., more precisely, number of unique labels [y-axis] that have been assigned to M patient traces in the event log is plotted as a function of M [x-axis]). Of note is the power-law like distribution of label frequencies for the experimental log, in which the vast majority of labels are associated with very few patient traces, and there are relatively few labels that are assigned to a large number of patient traces. For example, roughly 55 diagnosis labels are only assigned to less than 10 patient traces (<0.1% of patient traces in the experimental log), and the mean diagnosis label-frequencies is 2.5.

To evaluate the proposed DTM, we also developed a Naïve Bayes based model to derive latent treatment patterns from the experimental log. Fig. 4 illustrates the generation process of the Naïve Bayes based model generates each clinical event e_i in three sampling steps, i.e., sample a diagnosis label c_i according to the uniform distribution from c_σ, sample the activity type $e_i \cdot a$ of e_i given the diagnosis label under the conditional probability $P(e_i \cdot a|c_i)$, and sample the occurring time stamp $e_i \cdot t$ of e_i given

Table 2
The details of the experimental event log.

Number of patient traces	Number of events	Number of activity types	Number of comorbidity types	Minimum length of stay (day)	Maximum length of stay (day)	Average length of stay (day)
12,120	706,348	617	76	1	94	9.9

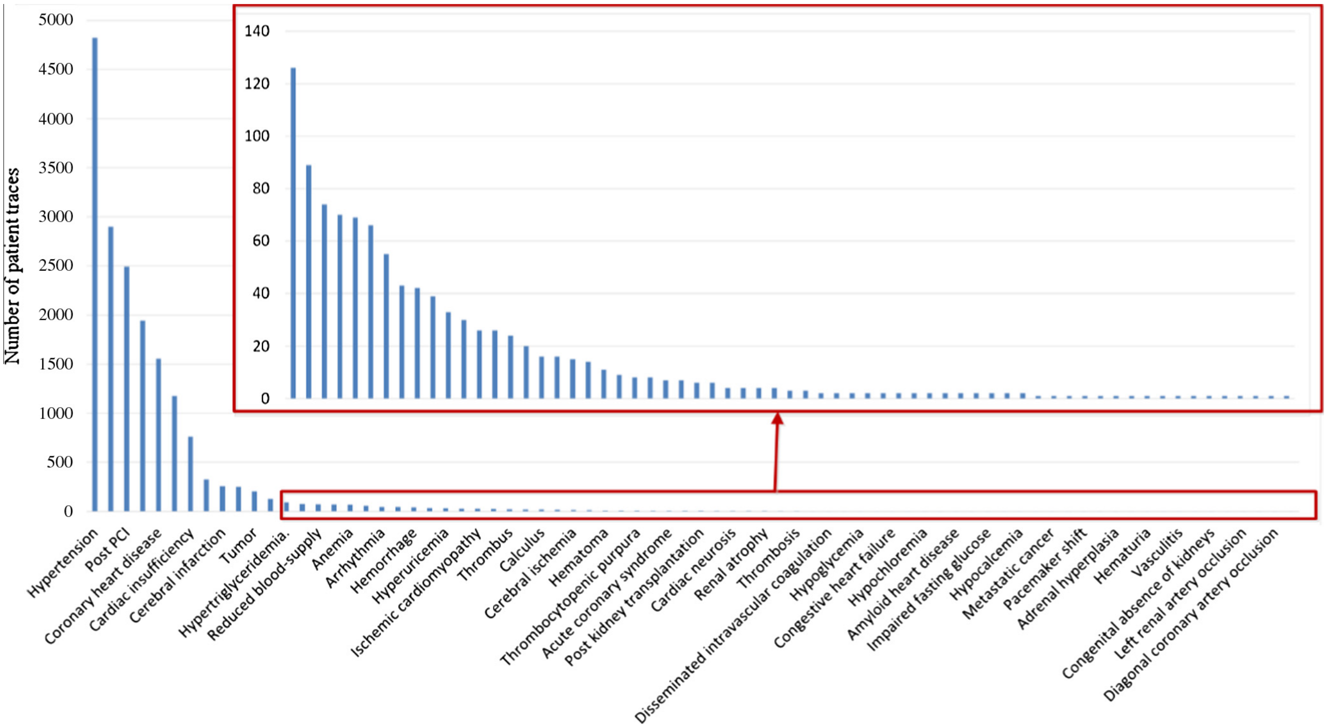

Fig. 3. The number of training patient traces for each unique diagnosis label in the experimental log.

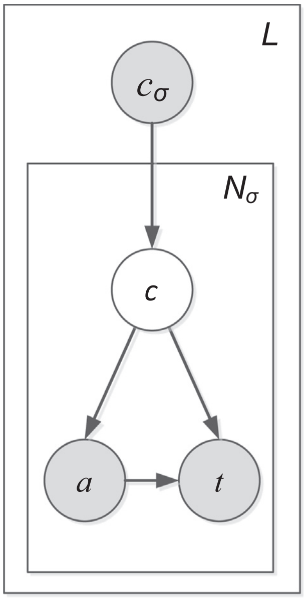

Fig. 4. A Naïve Bayes method for comorbidity-incorporated treatment pattern mining.

the diagnosis label and the treatment activity type $e_i \cdot a$ under the conditional probability $P(e_i \cdot t | c_i, e_i \cdot a)$. Note that the model parameters can be learned by maximum-likelihood estimation. In particular, the conditional probability of a treatment activity type a given a diagnosis label c, and the conditional probability of an occurring time stamp t given a diagnosis label c and a treatment activity type a, can be estimated as follows:

$$P(a|c) = \frac{n_{a,c}}{\sum_{a' \in A} n_{a',c}}, \tag{7}$$

$$P(t|c,a) = \frac{n_{a,t,c}}{\sum_{t' \in T} n_{a,t',c}}, \tag{8}$$

where $n_{a,c}$ is the co-occurrence count between treatment activity a and diagnosis label c for all patient traces in the experimental log, and $n_{a,t,c}$ is the co-occurrence count between treatment activity a, its occurring time stamp t and diagnosis label c for all patient traces in the experimental log, respectively.

4.2. Treatment pattern number selection

The number of treatment patterns Z indicates how many latent aspects of patient traces can be derived, which may influence the performance of the proposed model. Note that the performance of the Naïve Bayes based model does not change with a different number of treatment patterns since it does not consider the latent aspects in patient traces. In this subsection, we focus on selecting the number of latent treatment patterns based on topic modeling performance, which is measured by conditional perplexity. Perplexity is a standard measure to compare probability models. In this study, we calculate the perplexity on generating treatment event e from diagnosis label c as follows:

$$Perp(L) = \exp\left\{-\frac{\sum_{\sigma \in L}\sum_{e_i \in \sigma} \ln P(e_i|c_\sigma)}{\sum_{\sigma \in L}|\sigma|}\right\}, \tag{9}$$

where $P(e_i|c_\sigma)$ is the probability of generating a particular clinical event e_i conditioned on a set of diagnosis labels c_σ, which can be estimated as

$$P(e_i|c_\sigma) = \sum_{z \in Z} P(e_i \cdot a|z) P(e_i \cdot t|e_i \cdot a, z) P(z|c_\sigma) \propto \sum_{c \in c_\sigma}\sum_{z \in Z} \phi_{z,e_i \cdot a} \varphi_{z,e_i \cdot a,e_i \cdot t} \theta_{c,z} \tag{10}$$

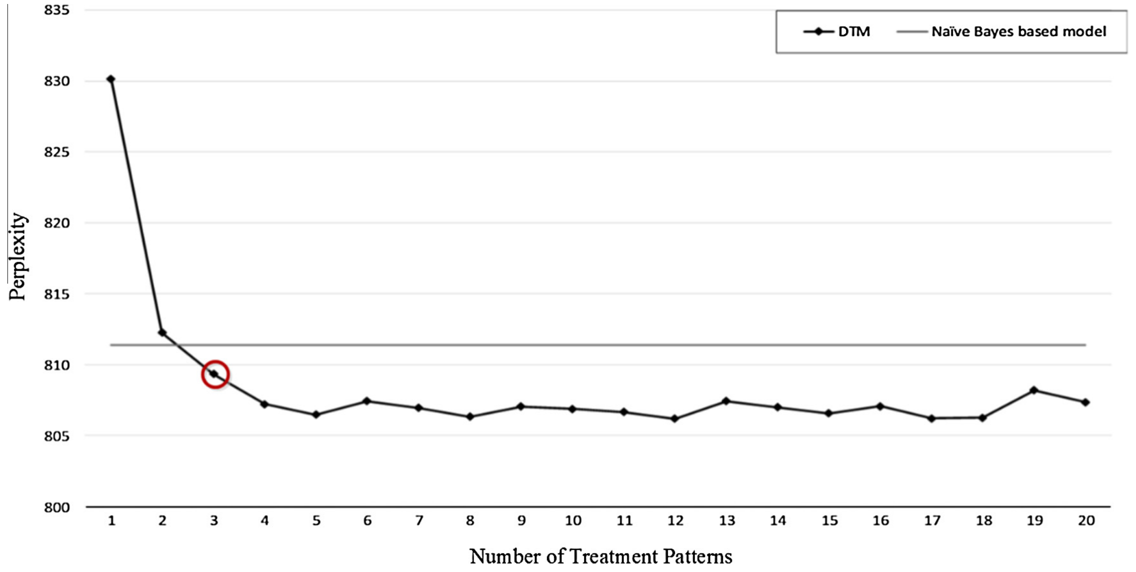

Fig. 5. Perplexity over number of latent treatment patterns with iteration number of 1000.

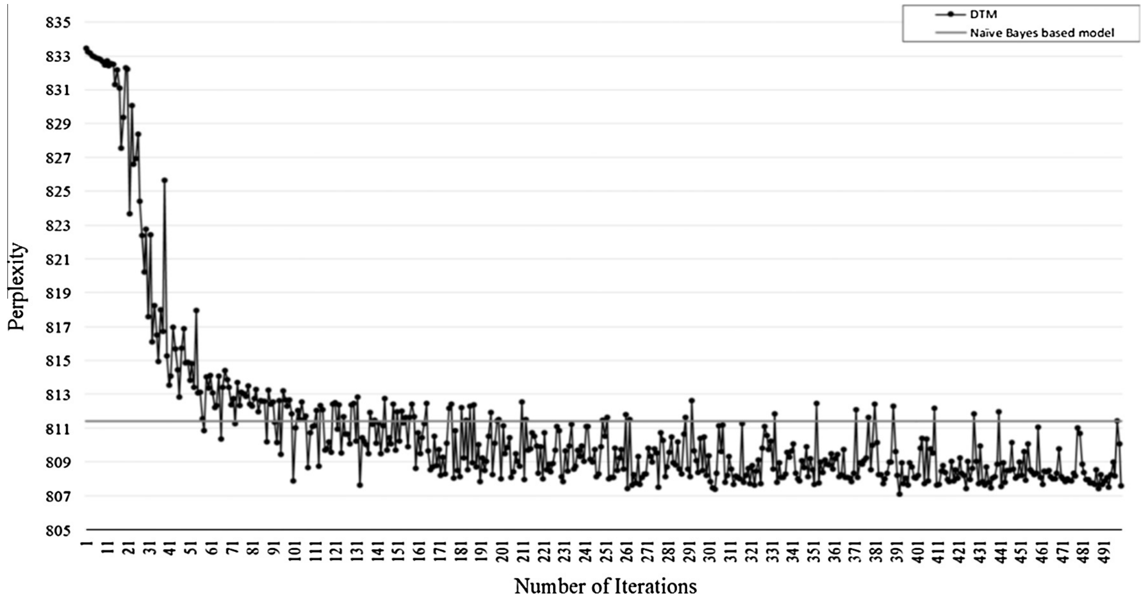

Fig. 6. Perplexity over number of Gibbs sampling iterations with treatment pattern number of 3.

For the Naïve Bayes based model, $P(e_i|c_\sigma)$ can be estimated as

$$P(e_i|c_\sigma) = \sum_{c \in c_\sigma} P(e_i \cdot a|c) P(e_i \cdot t|e_i \cdot a, c), \tag{11}$$

where $P(e_i \cdot a|c)$ and $P(e_i \cdot t|e_i \cdot a, c)$ can be calculated using Eqs. (7) and (8).

Fig. 5 shows the perplexity curves of both DTM and Naïve Bayes based model with a growing number of latent treatment pattern numbers. Note that, because the Naïve Bayes based model does not consider latent treatment patterns, its perplexity does not change with a growing number of treatment patterns. As shown in Fig. 5, the proposed DTM converges to its asymptote within a relative small number of treatment patterns, and becomes less than the baseline when the pattern number is more than 3. 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 dataset, and it requires more sampling computation and storage as well [35]. Thus, it needs to choose a balance between simplicity of the model and the degree of fitness. To this end, we examined the discovered patterns by DTM with different value of Z by a simple way: i.e., if the reducing ratio of perplexity is less than a particular threshold value ε, we do not select a larger Z. In this study, we set ε to be 3%, and 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.

Furthermore, we investigate the impact of the number of iterations of Gibbs sampling on the experimental results. Fig. 6 shows the perplexity curves of both DTM with 3 latent treatment patterns and the Naïve Bayes based model against the number of Gibbs sampling iterations. As depicted in Fig. 6, the proposed DTM converges

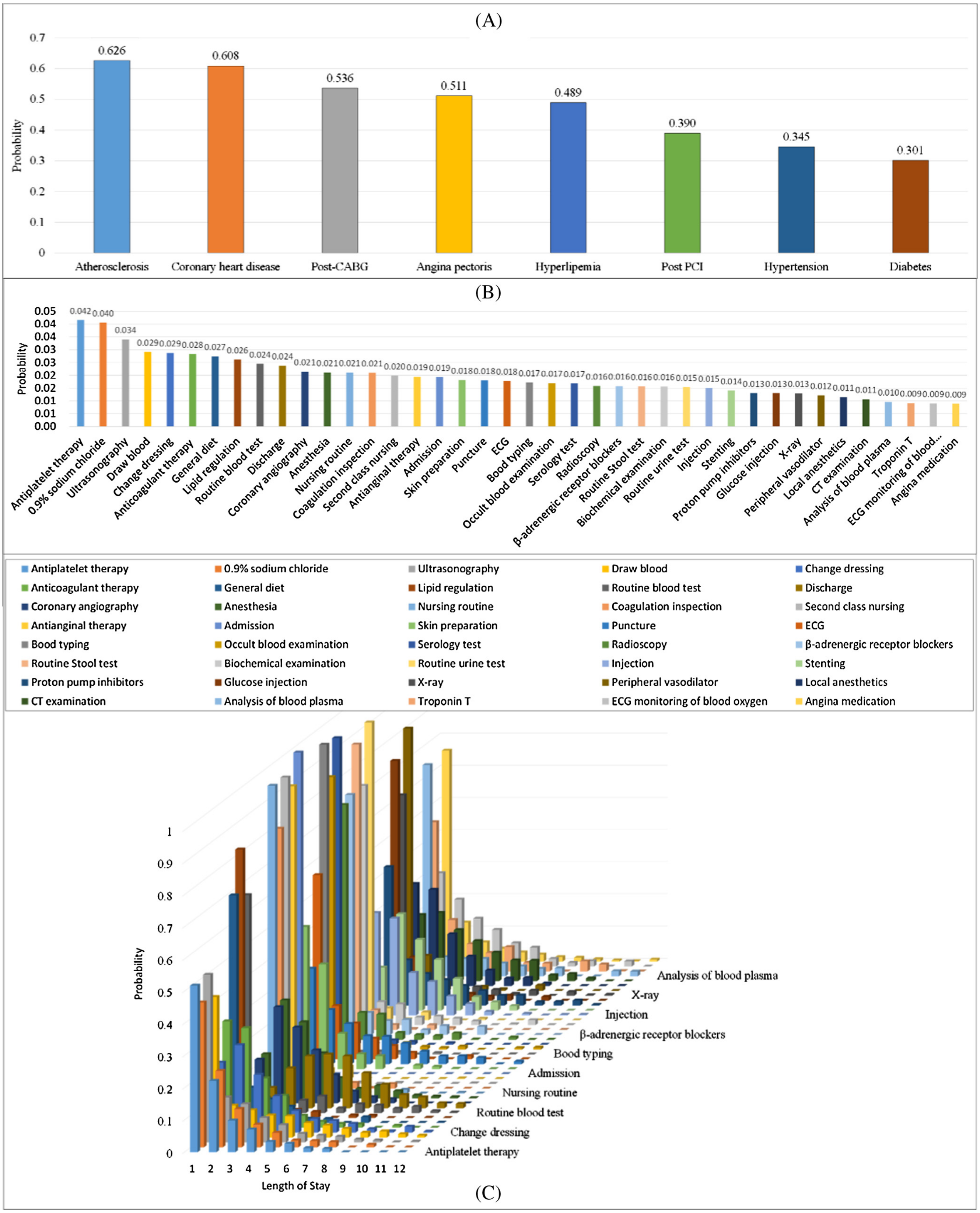

Fig. 7. The generated comorbidity-incorporated treatment Pattern-1. (A) Top 8 most related comorbidities and their probabilities, (B) top 40 most related treatment activities and their probabilities, and (C) the probabilities of occurring time stamps of top 40 most related treatment activities.

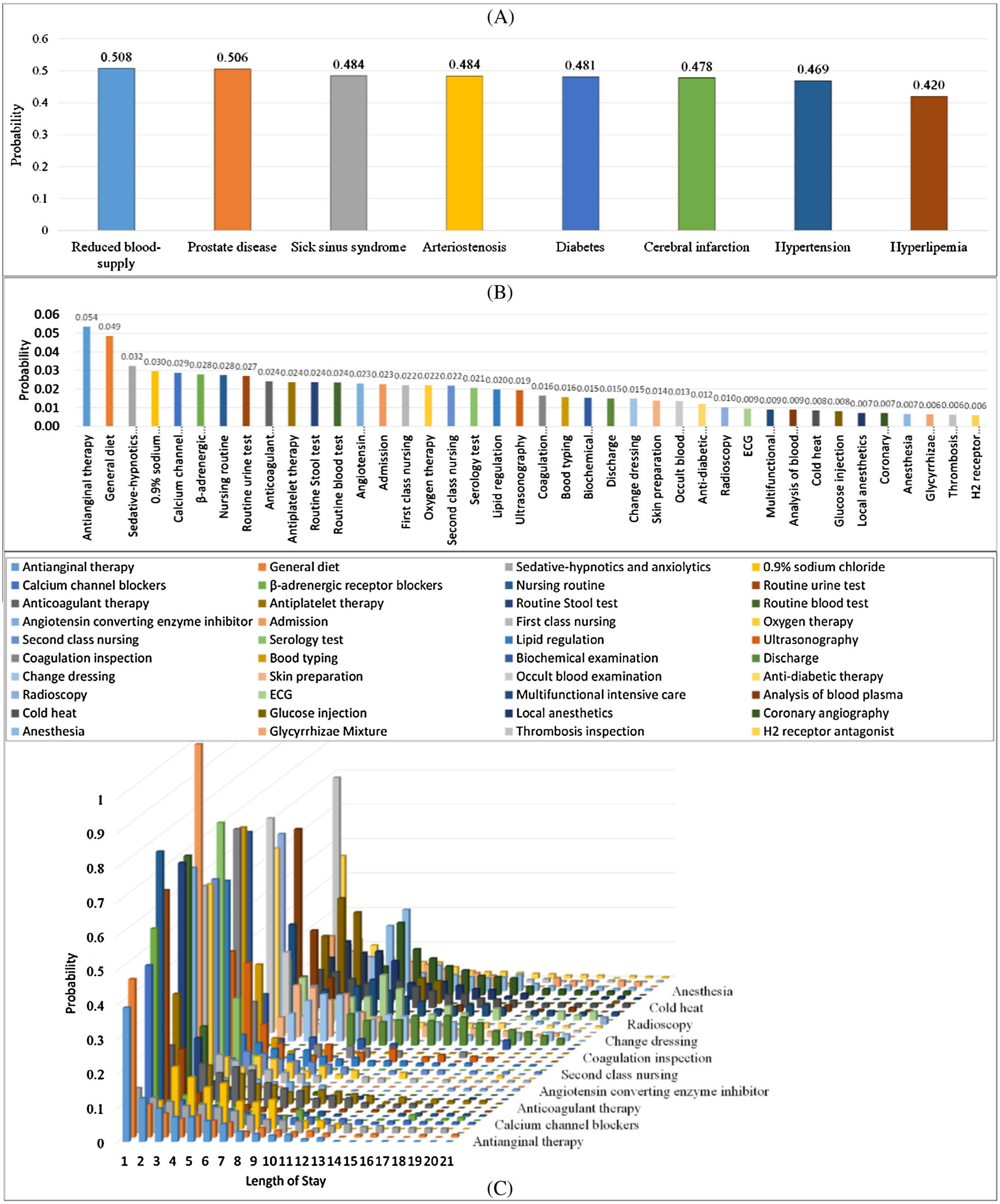

Fig. 8. The generated comorbidity-incorporated treatment Pattern-2. (A) Top 8 most related comorbidities and their probabilities, (B) top 40 most related treatment activities and their probabilities, and (C) the probabilities of occurring time stamps of top 40 most related treatment activities.

in less than 200 iterations. After its convergence, the perplexity of DTM is noticeably less than that of the Naïve Bayes based model.

Based on above observations, we choose 3 latent treatment patterns and 200 iterations as the default setting in the following experiments. For the other hyperparameters α, β and γ, they are set to symmetric Dirichlet priors with values of $50/Z$, 0.1, and 0.1, respectively, which are similar to the previous work [24].

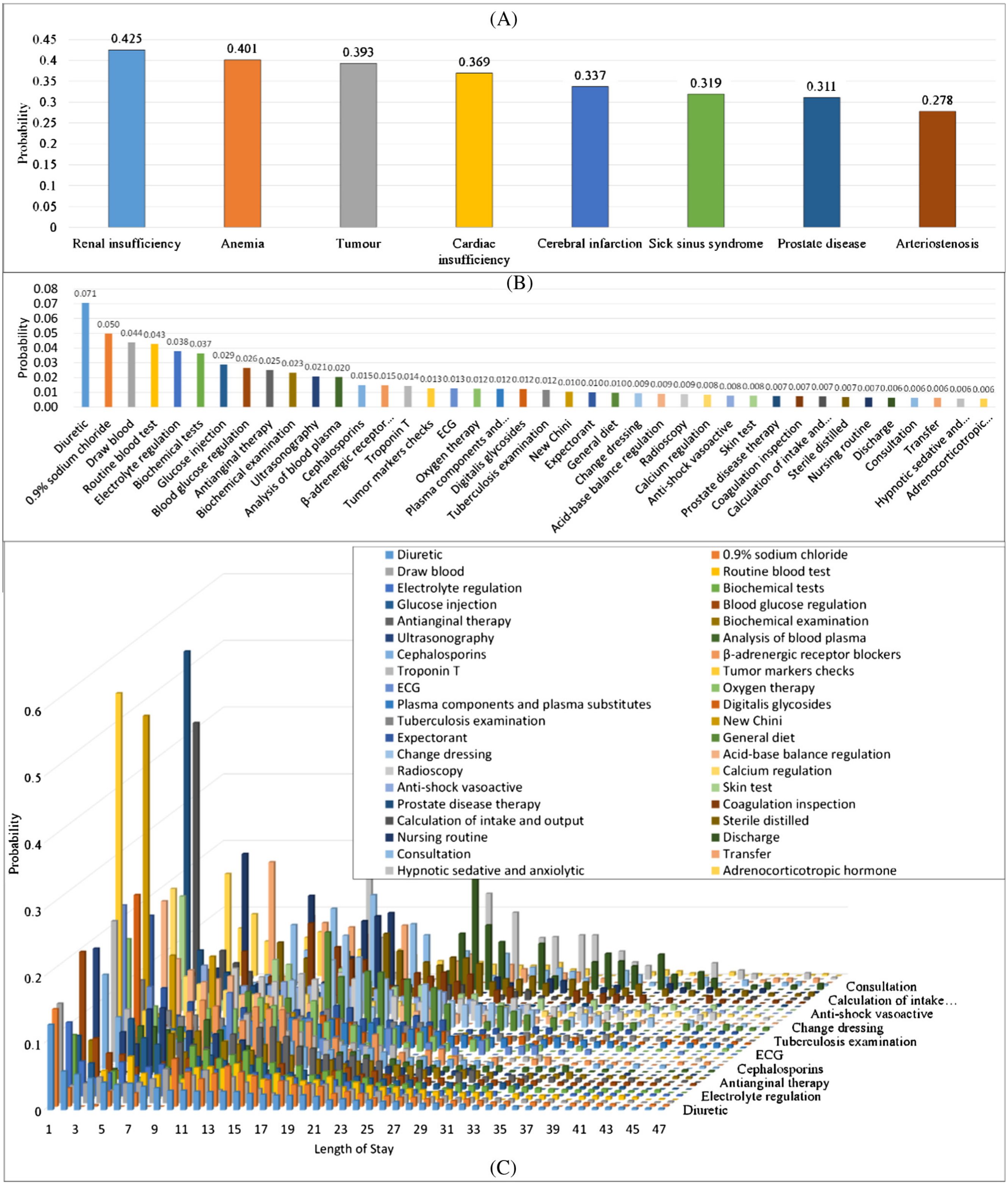

Fig. 9. The generated comorbidity-incorporated treatment Pattern-3. (A) Top 8 most related comorbidities and their probabilities, (B) top 40 most related treatment activities and their probabilities, and (C) the probabilities of occurring time stamps of top 40 most related treatment activities.

4.3. The generated diagnosis-incorporated treatment patterns

As stated earlier, the proposed DTM can jointly model the latent treatment patterns and patient's comorbidities. Fig. 7 offers a deeper understanding on treatment pattern discovery by showing the top 8 comorbidities (ranked by $\theta_{c,z}$) and top 50 treatment activities (ranked by $\phi_{z,a}$) and their occurring time stamps for the discovered patterns. Clearly, therefore, different treatment behaviors exist for patients with different comorbidities. The samples indicate that the proposed DTM can discover meaningful latent treatment patterns with trigger first-diagnosis and typical comorbidities, where both explicit and implicit treatment events are identified. Pattern 1

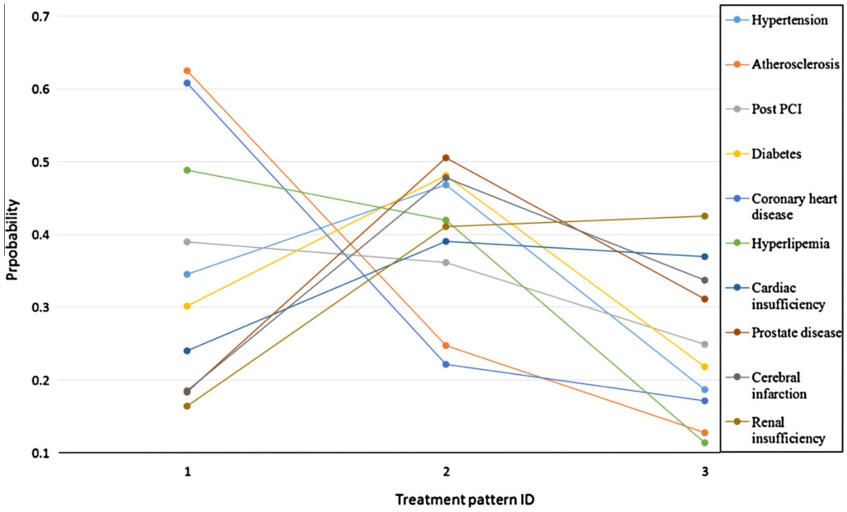

Fig. 10. Comorbidity distribution over treatment patterns.

shown in Fig. 7(B) contains typical treatment activities (e.g., "*Coronary angiography*" and "*Stenting*", etc.) for unstable angina. There is little variation occurred and common treatment activities are carried out smoothly. The activity "*Discharge*" occurs in the 3rd to 12th day after "*Admission*", indicating that patients who follow this pattern have shorter LOS (less than 12 days) than others, and almost all physical examinations (e.g., "*Ultrasonography*", "*X-ray*", "*CT Examination*", etc.) are performed on the first day after "*Admission*". It indicates that Pattern 1 is a background treatment pattern for unstable angina CP.

In comparison with Pattern 1, Pattern 2 shown in Fig. 8 contains typical conservative treatments of unstable angina, e.g., "*Anti-angina therapy*", "*Antiplatelet therapy*", etc. As shown in Fig. 8, Pattern 2 also contains a typical treatment, i.e., "*β-adrenergic receptor blockers*", which is not included in the representative treatments of Pattern 1. Note that this is a particular treatment activity for a typical comorbidity "*Sick sinus syndrome*" of unstable angina. Other explicit interventions for several typical types of comorbidities of unstable angina, e.g., "*Hyperlipidemia*", "*Hypertension*", "*Diabetes*", etc., can also be found in Pattern 2, such as "*Lipid regulation*" and "*Anti-diabetic therapy*". In general, patients who follow treatment Pattern 2 is larger than two weeks.

Moreover, it is interesting to see that "Pattern 3", as shown in Fig. 9, has captured typical treatments of unstable angina patients who have more complex conditions than others such that many serious comorbidities can be found in this pattern, e.g., "*Renal insufficiency*", "*Tumor*", "*Cardiac insufficiency*", etc. In general, patients who follow Pattern 3 prefer to a conservative treatment strategy, and essential treatments for the typical comorbidities are disclosed in Pattern 3, such as "*Diuretics*" for "*Cardiac insufficiency*" and "*Tumor markers checks*" for "*Tumor*".

Furthermore, the proposed DTM can discover the changes of treatment behaviors of the first diagnosis due to the occurrence of specific comorbidities. For example, Pattern 1 includes the typical treatment intervention "*Coronary angiography*" for the first-diagnosis unstable angina, on the other side, Pattern 3 lists treatment interventions for unstable angina patients with the comorbidity "*Renal insufficiency*", in which "*Coronary angiography*" is not included. It clearly indicates the influence of the typical comorbidity "*Renal insufficiency*" on the treatments of unstable angina. It indicates that the generated pattern can reflect not only explicit treatment behaviors w.r.t the therapy of typical comorbidities of the first diagnosis, but also implicit changes of treatments of the first-diagnosis due to the incorporation of typical comorbidities potentially. Although the baseline Naïve Bayes based model can also associate diagnosis labels and treatment events, the influence of comorbidities on the therapy and treatment of the first-diagnosis cannot be disclosed like the ones shown in Fig. 9.

As we illustrated above, the Naïve Bayes based model cannot utilize the co-occurrence information of treatment activities within particular patient traces and thus cannot distinguish the background treatment inventions for the first-diagnosis from the specific treatment interventions for particular comorbidities. On the other side, traditional probabilistic statistical models (e.g., LDA, CPM, etc.) can only discover hidden treatment patterns underlying clinical event log and cannot derive the correlations between comorbidities and treatment interventions. Our DTM utilizes the complementary advantages of both approaches. In particular, the proposed DTM has the ability to uncover latent treatment patterns that exhibit strong comorbidity-focus. Fig. 10 shows the distribution of top 10 comorbidities (ranked by frequency) over the three discovered treatment patterns. It is easy to see that there are no single-comorbidity-focused treatment patterns and all three discovered patterns are multi-comorbidity-focused, which share a variety of comorbidities at the same time. In addition, Fig. 10 shows some correlations between comorbidities. For example, Pattern 1 is dominated by the comorbidities "*Atherosclerosis*" and "*Coronary heart disease*", which also has similar distributions over all three discovered patterns. In addition, similar trends can be found for comorbidities "*Cardiac insufficiency*" and "*Renal insufficiency*". In clinical practice, patients who have "*Renal insufficiency*" may also have a great probability to have "*Cardiac insufficiency*".

4.4. Treatment recommendation

To evaluate the performance of the proposed model, we propose one possible application, i.e., treatment recommendation, As illustrated above, given a set of diagnosis labels c_σ of a particular patient trace, the clinical activity probability $P(a|c_\sigma)$ represents the most likely medications for σ, which can be measured as follows:

$$P(a|c_\sigma) = \sum_{z\in Z} P(a|z)P(z|c_\sigma) = \sum_{z\in Z}\sum_{c\in c_\sigma} P(a|z)P(z|c) = \sum_{z\in Z}\sum_{c\in c_\sigma}\theta_{c,z}\phi_{z,a}. \tag{12}$$

Table 3
Treatment recommendations using both the proposed DTM and the Naïve Bayes based model.

	MP@10	*MAP*
DTM	0.887	0.576
Naïve Bayes based model	0.853	0.550
TPM	0.885	0.572
CPM	0.865	0.557
LDA	0.871	0.562

In order to test our model quantitatively, we make an assumption that the diagnosis and treatments given by clinicians for particular patients are the ground-truth.[1] Then we test the consistency of the possible treatments suggested by our model with the ground truth. To this end, we split the log into a training set and a testing set, and evaluate the performance by 10-fold cross-validation. In particular, we recommended both the top 10 types of treatment activities for each patient trace, and the same number of treatment activities of each patient trace. Then, we checked if the recommended activities are actually included in the ground truth. In this sense, we utilize two measurements, i.e., "mean precision at top 10" (*MP*@10), and "mean average precision" (*MAP*).

We compare the results with both the Naïve Bayes based model, the standard LDA, and the other two extensions of LDA proposed in our previous work, i.e., CPM [24] and TPM [19]. Note that CPM does not take the comorbidities into account during the generation of latent treatment patterns from EMRs. With respect to TPM, although it incorporates patient-specific information into the model generation, it processes comorbidities as normal patient variables and thus does not model the correlations between treatments and comorbidities explicitly.

Table 3 shows *MP*@10 and *MAP* for the different models. The LDA-based models outperform the baseline approach in terms of treatment recommendation. Among the four LDA-based models, the proposed DTM achieves the best performance although TPM can obtain the comparative performance with DTM. The experimental results indicate that the explicitly modeling of the correlations between treatment behaviors and comorbidities can improve the quality of treatment recommendations in CPs.

5. Conclusions

In this paper, we explore a new problem called comorbidity-incorporated latent treatment pattern mining from the execution data of CPs, which aims to discover and model the associations between comorbidities and treatment behaviors in CPs. In detail, we propose a probabilistic statistical model to link clinician's diagnosis labels and consequent treatment events together so as to unveil the latent associations between diagnosis labels (including both the first-diagnosis and comorbidities) and treatments, and compute the contribution of comorbidities on treatments' adoption in CPs. The experimental results on a collection of 12,120 unstable angina patient traces from Chinese PLA General Hospital reveal practical importance over the proposed approach for the task of comorbidity-incorporated latent treatment pattern mining. The discovered patterns can help clinicians better understand their specialty and learn previous experiences from real healthcare data. In particular, clinicians can utilize the discovered treatment patterns to redesign composite clinical pathways with specific comorbidities into considerations.

The model proposed in this paper offers many avenues for further expansions and applications. In the future work, we aim to study the merits of our proposal in further clinical collaborations to investigate on additional sources of information (e.g., patient features, medical resources, etc.), explore on the experiments results, and empirical study the usage of our approach in real clinical environments. As well, several extensions of our approach will be investigated, such as treatment grouping and identification within the same therapy and treatment intention and anomaly detection for ongoing patient traces. These extensions will be explored and evaluated on a larger scale of clinical data collections as for our future work.

Conflicts of interest

The authors declare that they have no competing interests.

Acknowledgments

This work was supported by the National Nature Science Foundation of China under Grant Nos. 81101126 and 61562088, and the Fundamental Research Funds for the Central Universities under Grant No. 2014QNA5014. The authors are especially thankful for the positive support received from the cooperative hospitals as well as to all medical staff involved.

References

[1] H. Campbell, R. Hotchkiss, N. Bradshaw, M. Porteous, Integrated care pathways, BMJ 316 (7125) (1998) 133–137.
[2] B. Hunter, J. Segrott, Re-mapping client journeys and professional identities: a review of the literature on clinical pathways, Int. J. Nurs. Stud. 45 (4) (2008) 608–625.
[3] J. Schuld, T. SchSfer, S. Nickel, P. Jacob, M.K. Schilling, S. Richter, Impact of IT supported clinical pathways on medical staff satisfaction. A prospective longitudinal cohort study, Int. J. Med. Inform. 80 (3) (2011) 151–156.
[4] Z. Huang, X. Lu, H. Duan, On mining clinical pathway patterns from medical behaviors, Artif. Intell. Med. 56 (1) (2012) 35–50.
[5] Z. Huang, X. Lu, H. Duan, W. Fan, Summarizing clinical pathways from event logs, J. Biomed. Infor. 46 (1) (2013) 111–127.
[6] http://www.moh.gov.cn/zwgkzt/lclj/list.shtml (last access on 2015-2-28).
[7] H. Li, Informationazation is the foundation to achieve the standard management of clinical pathway, China Digit. Med. 5 (10) (2010) 1.
[8] M. Panella, S. Marchisio, F. Di Stanislao, Reducing clinical variations with clinical pathways: do pathways work?, Int J. Qual. Health Care 15 (6) (2003) 509–521.
[9] Z. Huang, X. Lu, H. Duan, Anomaly detection in clinical processes, AMIA Annu. Symp. Proc. 370–9 (2012).
[10] D. Verhelst, M. Nachtergaele, C. Hindryckx, V. Vandevyvere, S. Seghers, K. Smessaert, S. Vanderschueren, Can a care pathway help streamline the care process for patients with chronic fatigue syndrome?, Int J. Care Pathways 15 (4) (2011) 115–118.
[11] J. Choo, Critical success factors in implementing clinical pathways/case management, Ann. Acad. Med. Singapore 30 (4) (2001) 17–21.
[12] J.L. Aderson, C.D. Adams, E.M. Antman, et al., 2012 ACCF/AHA focused up-date incorporated into the ACCF/AHA 2007 guidelines for the management of patients with unstable angina/non-ST-elevation myocardial infarction, J. Am. Coll. Cardiol. 61 (23) (2013) e179–e347.
[13] J.A. Finegold, P. Asaria, D.P. Francis, Mortality from ischaemic heart disease by country, region, and age: statistics from World Health Organisation and United Nations, Int. J. Cardiol. 168 (2) (2012) 934–945.
[14] S. Evans-Lacko, M. Jarrett, P. McCrone, G. Thornicroft, Facilitators and barriers to implementing clinical care pathways, BMC Health Serv. Res. 10 (2010) 182.
[15] T. Rotter, L. Kinsman, E.L. James, A. Machotta, H. Gothe, J. Willis, P. Snow, J. Kugler, Clinical Pathways: Effects on Professional Practice, Patient Outcomes, Length of Stay and Hospital Costs (Review), John Wiley & Sons, Ltd, 2010.
[16] B. Emmerson, A. Frost, L. Fawcett, et al., Do clinical pathways really improve clinical performance in mental health settings?, Austral Psychiat. 14 (4) (2006) 295–398.
[17] J. Shi, Q. Su, Z. Zhao, Critical factors for the effectiveness of clinical pathway in improving care outcomes, in: 2008 International Conference on Service Systems and Service Management, June 30, 2008–July 2, 2008, pp. 1–6.
[18] Z. Huang, X. Lu, C. Gan, H. Duan, Variation prediction in clinical processes, in: M. Peleg, N. Lavrac, C. Combi (Eds.), Artificial Intelligence in Medicine. Lecture Notes in Computer Science, vol. 6747, Springer, Berlin/Heidelberg, 2001, pp. 286–295.

[1] In clinical settings, the given diagnosis and treatments are biased. Even for the same patient, different clinicians may have different opinions on patient conditions so as to give different diagnosis and treatments.

[19] Z. Huang, W. Dong, L. Ji, P. Bath, H. Duan, On mining latent treatment patterns from electronic medical records, Data Min. Knowl. Discov. 29 (4) (2015) 914–949.

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

[21] M. Peleg, Computer-interpretable clinical guidelines: a methodological review, J. Biomed. Infor. 46 (4) (2013) 744–763.

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

[23] D.M. Blei, Andrew Y. Ng, Michael I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (4–5) (2003) 993–1022.

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

[25] Z. Huang, X. Lu, H. Duan, Latent treatment pattern discovery for clinical processes, J. Med. Syst. 37 (2) (2013).

[26] S. Montani, G. Leonardi, Retrieval and clustering for supporting business process adjustment and analysis, Inform. Syst. 40 (2014) 128–141.

[27] D.R. Ferreira, Applied sequence clustering techniques for process mining, in: J. Cardoso, W. van der Aalst (Eds.), Handbook of Research on Business Process Modeling, Information Science Reference, IGI Global, 2009, pp. 492–513.

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

[29] R. Mans, H. Schonenberg, G. Leonardi, S. Panzarasa, A. Cavallini, S. Quaglini, W. van der Aalst, Process mining techniques: an application to stroke care, Stud. Health Technol. Infor. 136 (2008) 573–578.

[30] S. Gupta, Workflow and Process Mining in Healthcare, Master's Thesis, Technische Universiteit Eindhoven, 2007.

[31] A. Partington, M. Wynn, S. Suriadi, C. Ouyang, J. Karnon, Process mining for clinical processes: a comparative analysis of four Australian hospitals, ACM Trans. Manage. Inf. Syst. 5 (4) (2015) 18 (Article 19, January 2015).

[32] J. van de Klundert, P. Gorissen, S. Zeemering, Measuring clinical pathway adherence, J. Biomed. Infor. 43 (6) (2010) 861–872.

[33] G.T. Lakshmanan, S. Rozsnyai, F. Wang, Investigating clinical care pathways correlated with outcomes, in: F. Daniel, J. Wang, B. Weber (Eds.), Business Process Management, Lecture Notes in Computer Science, vol. 8094, Springer, Berlin, 2013, pp. 323–338.

[34] T.L. Griffiths, Finding scientific topics, Proc. Natl. Acad. Sci. USA 101 (2004) 5228–5235.

[35] D. Phung, B. Adams, S. Venkatesh, M. Kumar, Unsupervised context detection using wireless signals, Pervasive Mobile Comput. 5 (6) (2009) 714–733.

[36] Adam Perer, Fei Wang, Frequence: interactive mining and visualization of temporal frequent event sequences, in: ACM Intelligent User Interfaces (IUI), Haifa, Israel, 2014, pp. 153–162.

[37] Adam Perer, Fei Wang, Hu Jianying, Mining and exploring care pathways from electronic medical records with visual analytics, J. Biomed. Infor. 56 (2015) 369–378.

[38] Zhengxing Huang, Wei Dong, Lei Ji, Liangyin Ying, Huilong Duan, On local anomaly detection and analysis for clinical pathways, Artif. Intell. Med. 65 (3) (2015) 167–177.

[39] Wei Dong, Zhengxing Huang, A method to evaluate critical factors for successful implementation of clinical pathways, Appl. Clin. Infor. 6 (4) (2015) 650–668.