Latent treatment pattern discovery for clinical processes

J Med Syst (2013) 37:9915
DOI 10.1007/s10916-012-9915-2

ORIGINAL PAPER

Latent Treatment Pattern Discovery for Clinical Processes

Zhengxing Huang · Xudong Lu · Huilong Duan

Received: 22 August 2012 / Accepted: 29 December 2012 / Published online: 8 February 2013

Abstract A clinical process is typically a mixture of various latent treatment patterns, implicitly indicating the likelihood of what clinical activities are essential/critical to the process. Discovering these hidden patterns is one of the most important components of clinical process analysis. What makes the pattern discovery problem complex is that these patterns are hidden in clinical processes, are composed of variable clinical activities, and often vary significantly between patient individuals. This paper employs Latent Dirichlet Allocation (LDA) to discover treatment patterns as a probabilistic combination of clinical activities. The probability distribution derived from LDA surmises the essential features of treatment patterns, and clinical processes can be accurately described by combining different classes of distributions. The presented approach has been implemented and evaluated via real-world data sets.

Keywords Clinical process analysis · Latent Dirichlet Allocation · Pattern discovery · Careflow log

Introduction

Clinical processes (CPs), as unique type of patient-linked processes (i.e., diagnostic and therapeutic procedures to be carried out for a particular patient), are becoming an important issue in the health-care domain [1–6]. Health-care organizations are constantly pushed to improve the quality of care services in an unfavorable economic scenario and under financial pressure by governments [3, 7, 8]. Improving clinical process efficiency is of utmost importance to care service delivery.

Clinical process analysis (CPA) has experienced increased attention over the years due to its importance to health-care management in general and to its usefulness for capturing the actionable knowledge to administrate, automate, and schedule the best practice for individual patients in their therapy and treatment processes [7–10]. This has resulted in various successful approaches that are capable of analyzing clinical activities in CPs. The majority of research has focused on CP pattern discovery that aims to recognize what clinical activities are essential/critical for CPs, and also where temporal orders of these activities are quantified with numerical bounds. Research in CP pattern discovery exploits the fact by automatically measure clinical activities to aid clinical experts analyze and improve CPs [7, 9, 10].

However, the task of CP pattern discovery is not easy. Patient careflow are typically executed according to a diagnostic-therapeutic cycle, comprising observation, reasoning and action [6, 11–14]. The diagnostic-therapeutic cycle heavily depends on medical knowledge to deal with case-specific decisions that are made by interpreting patient-specific information [7, 8, 11]. During CP execution, the patient state might be changed dynamically, such as complications, infections, or poisonings, which in turn leads to various clinical activities occurred in CPs, and urges the process to be a mixture of latent treatment patterns as well. What makes the discovery of such patterns more complex is that they are typically unknown in prior. In fact, these

Z. Huang (✉) · X. Lu · H. Duan
College of Biomedical Engineering and Instrument Science, Zhejiang University, 310008, Zhou Yiqing Building 510, Zheda road 38#, Hangzhou, Zhejiang, China
e-mail: zhengxing.h@gmail.com

X. Lu
e-mail: lxd@vico-lab.com

H. Duan
e-mail: dhl@vico-lab.com

patterns are composed of several clinical activities and that the composition of activities has a large variability depending on factors such as time, location and patient individual.

Although many excellent studies have been proposed for CP pattern discovery [7, 10, 15], they assume prior knowledge about CP [16, 17], using which treatment patterns are derived in a completely supervised manner. In addition, many approaches are based on the experiences and knowledge of clinical experts [18, 19]. The analysts interpret large amounts of collected data, and elaborate treatment patterns in patient traces of CPs, piece after piece, which can be very tedious. Furthermore, it appears that analysis results are somehow influenced by perceptions, e.g., treatment patterns in CPs are often normative in the sense that they state what should be done rather than describing the actual patterns in CPs. As a result, it tends to be rather subjective. The challenge, therefore, is how to discover latent patterns automatically and objectively without prior knowledge of CPs.

To this end, the methods using data mining and machine learning technologies to analyze CPs based on associated careflow logs are receiving gradual attentions in medical informatics [7, 10, 11]. These techniques are also called process mining [20–23]. Process mining techniques have been widely studied in the domain of business process management, which attempt to extract non-trivial and useful information from workflow logs [7, 23]. One important aspect of process mining is control-flow discovery, i.e., automatically constructing a process pattern (e.g., a BPMN model [24]) describing the causal dependencies between activities. Such discovered processes have proven to be very applicable to the understanding, redesign, and continuous improvement of business processes [23].

In clinical settings, many hospital information systems can monitor various clinical activities in CPs, which produce large careflow logs. It is, therefore, possible to apply process mining techniques to extract non-trivial knowledge from these logs and exploit these for further analysis. However, the diversity of clinical activities in CPs is far higher than that of common business processes. The use of traditional process mining techniques may generate spaghetti-like treatment patterns that are difficult to be comprehended by clinical experts [7, 10], such incomprehensible patterns are either not amenable or lack of assistance to efforts of analysis and improvement of CPs. In addition, existing process mining algorithms often produce excessive volume of patterns that may overwhelm the analysts [7]. In particular, the meanings or significance of the discovered patterns sometimes goes untold. As indicated in [7, 10, 25], the use of traditional process mining techniques can prove inadequate in CP pattern discovery.

In this regard, we introduce a novel approach to discover CP patterns from careflow logs. For this we propose to leverage the power of probabilistic topic models (1) to automatically extract latent treatment patterns from careflow logs and (2) to enable the recognition of CPs as a composition of such treatment patterns. Our approach is based on an assumption that we can discover latent treatment patterns by mining careflow logs which regularly record various clinical activities in CPs. As the heart of the assumption is the question, whether we can have appropriately descriptive yet robustly detectable careflow logs to record clinical activities in a variety of clinical settings. Since many hospital information systems regularly record a wide range of valuable data, such as which clinical activities are performed, and when, these data can be organized in such a way that they contain a history of what occurred during CP execution, in a manner that facilitates making useful higher-level inferences. The idea of discovering latent treatment patterns from careflow logs is therefore, essential to move us away from the traditional subjective approaches for CPA, and adopt a more objective perspective.

The rest of the article is organized as follows. Section "Related work" summarizes some related studies. Section "Method" describes steps for discovering latent treatment topics for CPs. Section "Case study" carefully presents our experimental results and the result analysis. Finally, some conclusions are given in section "Conclusion".

Related work

The application of information technologies for CPA is a relatively unexplored field, although it has already been attempted by some researchers from academia and industry. For example, commercial business intelligence and business activity monitoring tools have been used to analyze CPs, which typically look at aggregated data seen from the measures, e.g., length of stay, mortality, and infection rate, etc [9]. As valuable as these tools are, they restrict the attention to an external perspective of CPA. In clinical practice, actual work can deviate from the definitions of CPs due to many reasons, and it is very important for health-care organizations to discover and analyze these differences to improve CPs.

In this context, process mining [20, 21], as a general method in business process analysis, is gaining increasing attention in analyzing CPs and other kinds of health-care processes [7, 9, 10]. The underlying idea of process mining is to discover CP models from careflow logs that record their executions. Being transferred into medical settings, process mining methods may be applicable, for example, in retrieving frequent CP patterns from careflow logs, which might be further utilized to refine CP itself [10]. In fact, process mining has already been attempted in clinical environments

by some researchers. In [9], Lin et al. reported a technique that was developed to discover the time dependency pattern of CPs for managing brain stroke. In [22], Yang et al. propose a process mining algorithm to facilitate the automatic and systematic detection of health-care fraud and abuse for CPs. In [26], Mans et al. applied process mining to discover how stroke patients are treated in different hospitals. In [7], a methodology of using process mining techniques to support health-care process analysis is thoroughly investigated. Especially, a case study was conducted in the Hospital of São Sebastião in Portugal by gathering data from the hospital information system and analyzing the data set by utilizing a set of process mining techniques for the selected radiological examination processes. In our previous work [10], we have developed a new process mining algorithm to discover a set of treatment patterns given an input event log and a minimum support threshold value, such that it can find what critical clinical activities are performed and in what order, and provide comprehensive knowledge about quantified temporal orders of clinical activities in CPs.

However, the diversity of medical behaviors in clinical processes is far higher than that of common business processes. The use of traditional process mining techniques may generate spaghetti-like process models that are difficult to be comprehended by clinical experts [7], such incomprehensible models are either not amenable or lack of assistance to efforts of analysis and improvement of clinical processes. In addition, existing process mining algorithms often produce excessive volume of models that may overwhelm the analysts [7]. In particular, the meanings or significance of the discovered models sometimes goes untold. As indicated in [7, 25], the use of traditional process mining techniques though successful in discovering clinical process models can prove inadequate in CP analysis. Furthermore, although a patient trace is generally guided by a specific clinical process model, it is possible that clinical activities of that patient trace represents multiple underlying treatment topics. For example, a patient who follows the bronchiole lung cancer careflow may also be performed specific activities for his/her diabetes treatment. Even for the patients with the same disease, a slight dissimilarity of patient states may result in different patient careflow. As the obtained log contains patient traces that deal with a variety of medical problems, it can be assumed that the log is actually generated by multiple underlying treatment patterns [10, 27].

Method

In this section, we propose a Latent Dirichlet Allocation (LDA)-based method to discover underlying treatment patterns for CPs. We explain some notations and terminologies at first. Then we present our approach in detail.

Notation and terminology

The objective of this study is to discover latent treatment patterns for CPs. In particular, the proposed approach assumes that it is possible to record various clinical activities in CPs such that each activity refers to a well-defined step of CPs. In order to explain the proposed approach, we introduce the following notations and terminologies at first.

Let $\mathcal{A}$ be the set of clinical activities. A patient trace is represented as a non-empty set of clinical activities performed on a particular patient, i.e., $c = \langle a_1, a_2, \ldots, a_n \rangle$, where $a_i \in \mathcal{A}$ $(1 \leq i \leq n)$ is a particular clinical activity. A careflow log $\mathcal{L}$ is a set of tuples $\langle pid, c \rangle$, where pid is a patient identifier, and σ is a patient trace.

We depict a simple example of a careflow log $\mathcal{L}$ as shown in Fig. 1, which are correctly recorded in intracranial hemorrhage CP, by using letters of the alphabet. The meaning of the example alphabetic labels are described in Fig. 2. $\mathcal{L}$ consists of nine patient traces. Each trace consists of a set of clinical activities. For the particular trace (patient trace id is 411676), its first activity is *adm* (*admission*), and its last activity *dis* (*discharge*).

With respect to our LDA-based model, clinical activities are "words" in the model. A patient trace, which is a "document" in our model, is a bag of clinical activities. And a "corpus" is a collection of these "documents" (careflow logs).

We assume that clinical activities in CPs are regularly recorded into careflow logs by various kinds of hospital information systems, e.g., electronic medical record system (EMRs), radiology information system, picture archiving and communication system, laboratory information system, etc., which effectively reflects the real executing conditions in patient careflow.

Generative process

In this study, we assume that a patient trace is represented by a mixture of treatment patterns, with regard to specific categories of clinical activities in CPs. As shown in Fig. 3a, patient traces are coded as the bag of multinomial vectors. Each trace in the log is modeled as a finite mixture over an underlying set of K treatment patterns. The treatment pattern mixture is drawn from a Dirichlet prior to the entire careflow log. Figure 3b indicates the LAD-based model we employed for this study. Especially, we denote the pattern-trace distribution as θ, each being drawn independently from a symmetric Dirichlet prior α, and the activity-pattern distribution as ϕ, each being drawn from a symmetric Dirichlet prior β.

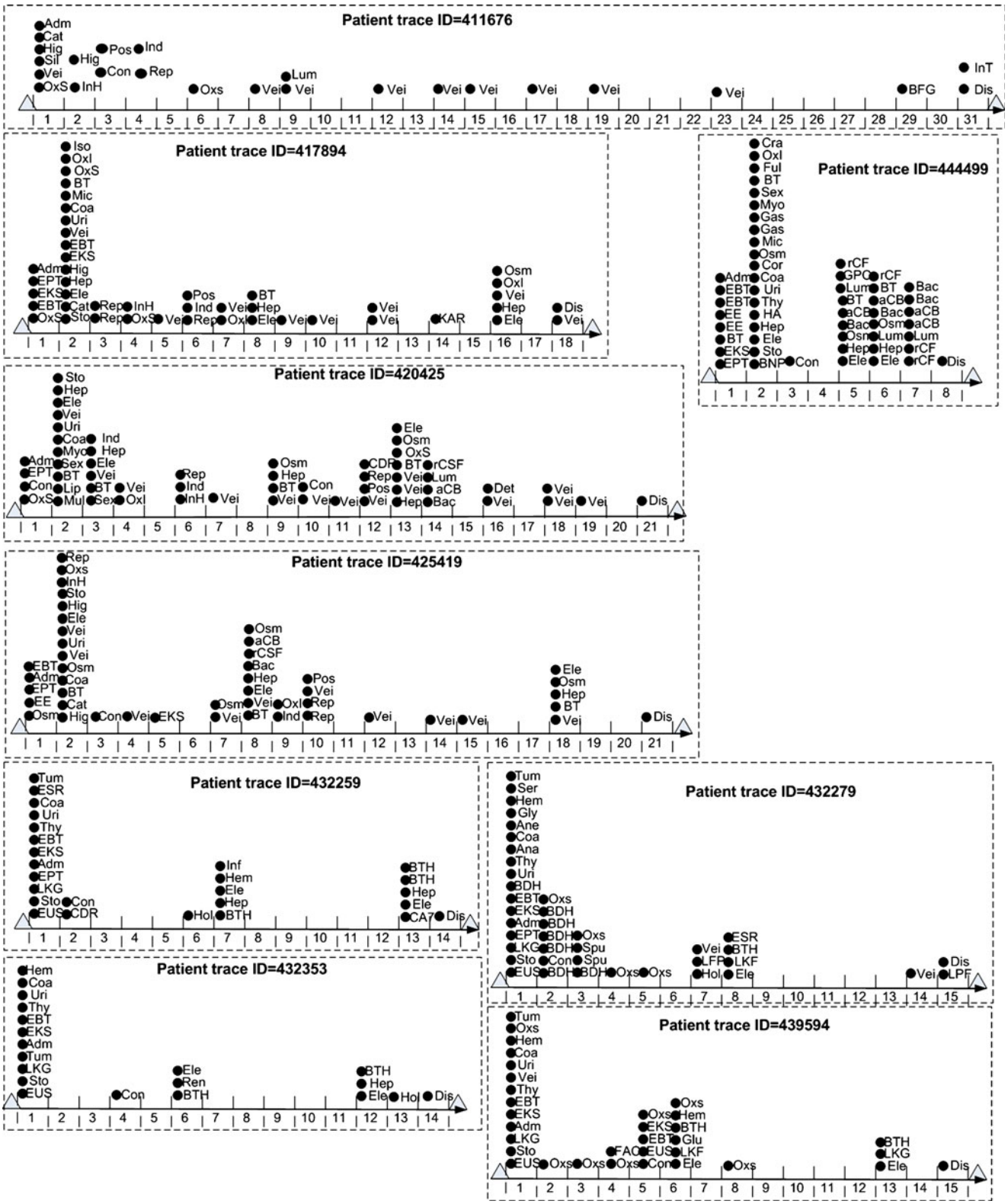

Fig. 1 An intracranial hemorrhage log example. The traces are simplified information extraction from patient records of Zhejiang Huzhou Central Hospital of China

aCB: Acute Cerebrospinal fluid biochemical
aCC: Acute CSFRT cryptococcal
Adm: Admission
Ana: Analysis of urine microalbumin
Ane: Anemia (3 items)
Ant: Anti-O rheumatoid
Bac: Bacteria and fungi were cultured and identified
BFG: (BFGF) topical bovine basic fibroblast growth
BT: Blood test
BTH: Blood test+Hypersensitive CRP
BNP: B-type natriuretic peptide
BDH: B-D Heparin cap
Cat: Catheterization
CA7: CA72-4
CDR: Color Doppler routine inspection
Coa: Coagulation + D-dimer
Con: Conventional ECG Exam
Cor: Cortisol
Cra: Craniotomy for intracranial decompression (including the brain,temporal
Pole)
CSF: CSF biochemical
CFA: Cerebrospinal fluid biochemical+ADA
Dis: Discharge
DLV: Determination of left ventricular function
DT3: Determination of tumor (3 items)
EBT: Emergency blood test
EE: Emergency Electrolyte
EKS: Emergency kidney, sugar
Ele: Electrolyte
EPT: Emergency PT
ERS: Emergency renal, sugar
ESR: ESR
EUS: Emergency ultra-sensitivity CRP
FAC: By the femoral artery catheter cerebral arteriography
Ful: Full set of Lipids (hospital)
Gas: Gastrointestinal high nutrition therapy
Glu: Glucose
Gly: Glycosylated hemoglobin
GPC: General physical cooling
Hem: Hemorheology
HA: Hepatitis A antibody
Hep: The hepatorenal sugar (hospitalization)
Hig: High-frequency oxygen/ hour
HLA: HLA-B27
Hol: 24-hour Holter
IDF: Intracranial Doppler flow imaging (TCD)
Imu: Immune (5 items)
In3: Inflammation (3 items)
Ind: Indwelling catheter
InT: Infrared treatment
InH: Intracranial hematoma(including simple epidural)
Iso: Isoflurane (live Ning)/1ml/ml
KAR: Kidneys and renal vascular color Doppler ultrasound
LFP: Low-frequency pulse power treatment
LKF: Liver and kidney function (hospitalization)
LKG: Liver, kidney, the glycolipid heart enzyme (hospitalization)
Lip: Lipids (7 items, hospitalization)
Lum: Lumbar puncture
Mic: Micro-jet atomization mask
Mul: Multiple intracranial hematoma
Myo: Myocardial enzymes
OxS: Oxygen saturation monitoring
OxI: Oxygen inhalation
Osm: Osmotic pressure
Pos: Postoperative drainage
rCF: CSF routine
Ren: Renal function (hospitalization)
Rep: Replacement of drainage
Ser: Serum troponin T assay
SE+: Stool examination+OB
Sex: Sex hormones
Sil: Silicone suction drainage
Spu: Sputum culture
Sto: Stool examination
Th7: Thyroid function (7 items)
Thy: Thyroid (five items)
Tum: Tumors (10 items)
Uri: Urine + sediment test
Vei: Vein catheterization

Fig. 2 The meaning of the example alphabetic labels of the example log

The generative process is as follows:

1. For each patient trace c in a clinical workflow log $\mathcal{L}$, a multinomial parameter θ_c over K treatment patterns is sampled from Dirichlet prior $\theta_c \sim Dir(\alpha)$.
2. For each clinical activity a in c,
   (a) A treatment pattern t is sampled from multinomial distribution $z_{t,c,a} \sim Mult(\theta_{z,c})$.
   (b) The value $w_{t,c,a}$ is sampled from multinomial distribution $w_{t,c,a} \sim Mult(\phi_{t,z_{t,c,a}})$.

From the generative graphical model depicted in Fig. 3b, we can write the joint distribution of all known and hidden variables given the Dirichlet parameters as follows.

$$p(c, t, \theta|\alpha, \beta) = p(\theta|\alpha) \prod_{a \in c} p(t|\theta)p(a|t, \beta) \quad (1)$$

And the likelihood of a patient trace c is obtained by integrating over θ and summing over t as follows.

$$p(c|\alpha, \beta) = \int p(\theta|\alpha) \prod_{a \in c} p(a|\theta, \beta)d\theta \quad (2)$$

Finally, the likelihood of the careflow log $\mathcal{L}$ is product of the likelihoods of all patient traces in $\mathcal{L}$:

$$p(\mathcal{L}|\alpha, \beta) = \prod_{c \in \mathcal{L}} p(c|\alpha, \beta) \quad (3)$$

Parameter estimation for LDA by directly and exactly maximizing the likelihood of $\mathcal{L}$ in Eq. 1 is intractable. One solution is to use approximate estimation methods such as Variational Methods and Gibbs Sampling [28]. Gibbs Sampling is a special case of Markov-chain Monte Carlo (MCMC) and often yields relatively simple algorithms for approximate inference in high-dimensional models such as LDA [29]. In this study, we use Gibbs sampling to estimate treatment

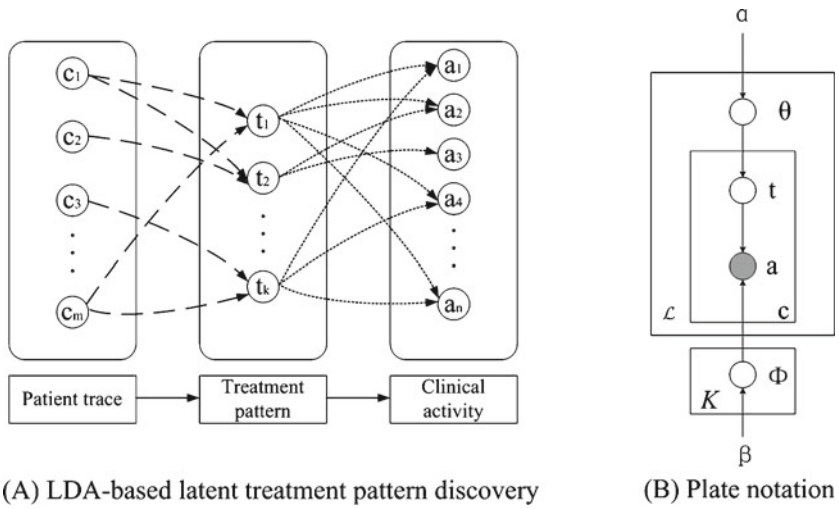

Fig. 3 Graphical representation of LDA-based model

patterns from the collected log as well as to estimate the activity-pattern and pattern-trace probability distributions.

Using this generative model, the pattern assignment for a particular clinical activity can be calculated based on the current pattern assignment of all the other activity positions. More specifically, the pattern assignment is sampled from:

$$p(t_i = k|t_{\neg i}, c) = \frac{n^a_{k,\neg i} + \beta}{\sum_{b\in A} n^b_k + \beta|A|} \frac{n^k_{c,\neg i} + \alpha}{\sum_{j\in T} n^j_\sigma + \alpha K} \quad (4)$$

where $t_i = k$ represents the assignment of the ith occurrence to pattern k, $t_{\neg i}$ represents all treatment pattern assignments not including the ith occurrence, $n^a_{k,\neg i}$ is the number of times the activity a is assigned to pattern k, not including the current instance, and $n^k_{c,\neg i}$ is the number of times pattern k is assigned to the patient trace c, not including the current instance.

After finishing Gibbs Sampling, two matrices θ and ϕ are computed as follows.

$$\theta_{k,a} = \frac{n^a_k + \beta}{\sum_{b\in A} n^b_k + \beta|A|} \quad (5)$$

$$\phi_{c,k} = \frac{n^k_c + \alpha}{\sum_{k\in T} n^k_c + \alpha K} \quad (6)$$

where $\theta_{k,a}$ is the probability of containing activity a in the treatment pattern k, and $\phi_{c,k}$ is the probability of the trace c has the treatment pattern k. The algorithm randomly assigns a pattern to each activity, updates the pattern to each activity using Gibbs sampling, and then repeats the Gibbs sampling process to update pattern assignment for several iterations [28]. Suppose there are K treatment patterns, $|A|$ clinical activity types, the total computational complexity of running l Gibbs sampling iterations for a clinical careflow log $\mathcal{L}$ is $O(|A| \cdot |\mathcal{L}| \cdot K \cdot l)$.

Before doing Gibbs sampling to discover latent treatment patterns from careflow logs, we need to identify the number of patterns contained in the collected log. Perplexity is a common measure of the ability of a model to generalize to unseen data. It is defined as the reciprocal geometric mean of the likelihood of a test corpus given a model. The perplexity score has been widely used in LDA to determine the number of topics, which is a standard measure to evaluate the prediction power of a probabilistic model [28]. In this study, we calculate the perplexity score for a particular collected log to determine the number of underlying treatment patterns in the log.

$$Perplexity = \exp\left[-\frac{\sum_{c\in\mathcal{L}} \log p(c|\mathcal{L})}{\sum_{c\in\mathcal{L}} |c|}\right] \quad (7)$$

where $|c|$ is the number of clinical activities in c, $\mathcal{L}$ is the careflow log. The perplexity of a set of activities from a specific patient trace, is defined as the exponential of the negative normalized predictive log-likelihood under the training model. As indicated in [28], the perplexity is monotonically decreasing in the likelihood of the test data. Therefore, a lower perplexity score over a held-out log indicates a better generalization performance. The value of K, which results in the smallest perplexity score over a randomly selected test data-set, is selected as the number of treatment patterns.

Taking the log shown in Fig. 1 as an example, it achieves the smallest perplexity score when $K = 2$. The typical activity labels for the derived patterns are ($p(a|t) \geq 0.01$) shown in Fig. 4. Clinical experts from the cooperated hospital have indicated that the two derived patterns have specific clinical intentions, i.e., cerebral hemorrhage treatment (ICD-10: I61), and subdural hematoma treatment (ICD-10: I62.006), respectively. If we look at the associated activities, most of them are related with the particular pattern, for example, the first pattern does not include

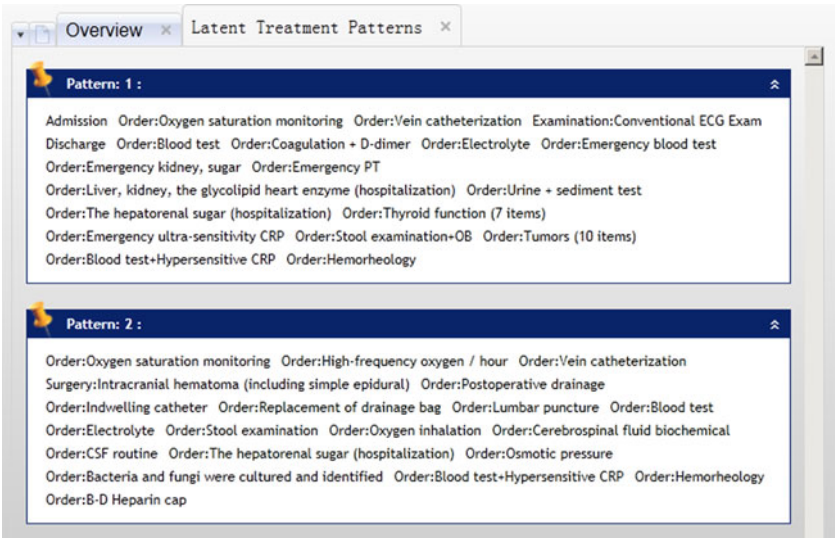

Fig. 4 The typical activity labels for the derived patterns from the log example shown in Fig. 1

Table 1 Pattern-trace distribution of the example log

Patient trace no.	Pattern no.	Probability	Pattern no.	Probability
411676	1	0.308	2	0.692
419894	1	0.451	2	0.549
420425	1	0.463	2	0.537
425419	1	0.463	2	0.537
444499	1	0.589	2	0.411
432259	1	0.599	2	0.401
432279	1	0.538	2	0.462
432353	1	0.604	2	0.396
439594	1	0.517	2	0.483

the surgery, while the second does. In addition, the discovered patterns from the example log demonstrates the implicit relationships between activities, for example, clinical activities "Surgery: Intracranial hematoma (including simple epidural)" and "Order: Postoperative drainage" are correlated with each other, and they have the same value of activity-pattern distribution. The relationships between activities via patterns can be used to provide good classification of patient traces.

In addition, the pattern-trace distribution $p(t|c)$ measures the connection (or relatedness) of a specific patient trace with a specific treatment pattern (i.e., the conditional probability of a treatment pattern in a given patient trace). We used this statistical probability to group patient traces by associating them with patterns. Taking the log shown in Fig. 1 as an example, the pattern-trace distribution values are listed in Table 1. The traces are grouped into specific clusters based on their pattern-trace distribution values. For example, the traces "420425" and "425419" have the same probability to belong to pattern 2 (with surgical treatment). Thus, both traces will be grouped into the same cluster.

Case study

To test the feasibility of the proposed approach, experiments on data-sets collected from Zhejiang Huzhou Central Hospital of China were performed. The explanation of the experimental setups and obtained results are presented in the following (Table 2).

Experimental design

The experimental data set was extracted from Zhejiang Huzhou Central hospital of China. The application of information technology in this hospital is at a relatively high level, and the EMRs has been gradually used since 2004. The system regularly records all kinds of information of CPs in the hospital. In the experiments, we extracted two specific careflow logs about intracranial hemorrhage and cerebral infarction from the system. The collected data is from 2007/08 to 2009/09. In addition, we removed those unclosed or incomplete patient traces, e.g., the trace of which the patient died or was transferred during his or her length of stay (LOS) from the collected log. The details of reserved logs are shown in Table 1, including the patient trace number, clinical activity number, activity type number, the average LOS, the minimum LOS, and the maximum LOS of each log. For example, the intracranial hemorrhage log consists of 259 patient traces. The average LOS of these traces is 22.95 days while some traces take a very short time, e.g., only two days in hospital, and other traces take much longer, e.g., 100 days in the hospital, which implicitly indicates the diversity of treatment behaviors in intracranial hemorrhage CP.

LDA-based model construction

Constructing LDA-based model is to fit latent treatment patterns to careflow logs. In this study, we evaluate the constructed model by using measures like likelihood on the collected log. Our goal is to derive clinical activity distribution density estimation, and a high likelihood on the collected log is expected.

The Dirichlet prior α and β of LDA are set to 0.2 and 0.1, which are common settings in literature. The number of iterations of the Markov chain for Gibbs sampling is set to 10000. Note that Gibbs sampling usually converges before 10000 iterations for the collected logs. In addition, to expand the number of trials when we construct LDA-based model, we adopt a fivefold cross-validation strategy. For each collected log, we split it randomly into five mutually exclusive subsets of equal size. We then designate each subset as the testing data set are used to compute the perplexity score while the others serve as the training data set. To minimize potential biases that may result from the randomized folding process, we perform this fivefold

Table 2 Careflow logs used in the experiments

Disease	Trace #	Activity #	Activity type #	Average LOS (days)	Min LOS (days)	Max LOS (days)
Intracranial hemorrhage	259	14194	274	22.95	2	100
Cerebral infarction	419	12038	269	15.2	4	100

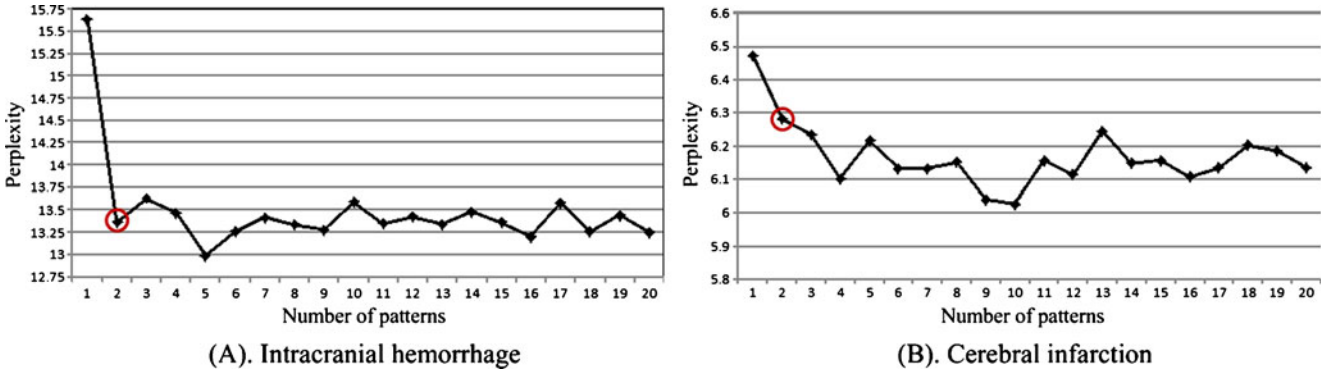

Fig. 5 Perplexity on the collected logs

cross-validation process five times and estimate the overall performance by averaging the performance estimates obtained from the 250 individual trials.

While performing LDA estimation, the number of patterns K varies from 1 to 20. As mentioned above, perplexity is algebraically equivalent to the inverse of the geometric mean per-activity likelihood. As mentioned above, a lower perplexity score indicates better generalization performance of the likelihood, which is in form of perplexity can be computed on the collected logs. Thus, for each log,

Table 3 Significant activities associated with the derived patterns from the collected logs

Pattern id	Significant activities	Support
Intracranial hemorrhage log		
1–1	Admission, Conventional ECG, Venipuncture catheter, Oxygen saturation monitoring, Catheterization, Color Doppler routine inspection, 24-hour Holter, Physical cooling, Low-frequency pulse electric treatment, Determination of left ventricular function, Serum troponin T determination, Electrolyte, Liver and kidney function, Blood test, CSF examination, Blood test + ultra-sensitivity CRP, Sputum culture, Emergent blood test, The hepatorenal sugar, Urine test + sediment, Urine culture, Emergency ultra-sensitivity CRP, Emergent electrolyte, Hepatorenal sugar lipase, Coagulation + D-dimer, Hemorheology full set, Inflammation 3 items, Emergent kidney, sugar, Emergent PT, Stool test + OB, Thyroid function 7 items, Analysis of urine microalbumin, Tumor 10 items, Anemia 3 items, ECG, Physical therapy, Osmotic pressure, Discharge	0.584
1–2	Admission, Conventional ECG, Venipuncture catheter, Oxygen saturation monitoring, Catheterization, Replace the drainage bag, High-frequency oxygen, Intracranial hematoma surgery, Drainage after surgery, Oxygen mask, Micro-jet atomization mask, Oxygen inhalation, Catheterization, Lumbar puncture, Physical cooling, Gastrointestinal high nutritional treatment, Electrolyte, Blood test, CSF examination, Cerebrospinal fluid biochemical, Bacteria and fungi culture and identification, Sputum culture, Emergent blood test, The hepatorenal sugar, Urine test+sediment, Coagulation + D-dimer, Myocardial enzymes, Stool test + OB, Lipids 7 items, Sex hormone, Emergent kidney and sugar, Physical therapy, Osmotic pressure, Discharge	0.416
Cerebral infarction log		
2–1	Admission, Glucose determination, Venipuncture catheter, Electrolyte, Emergent blood test, Emergent calcium determination, HCT, Emergent potassium determination, Emergent Sodium determination, Whole blood lactic acid, Blood gas analysis, Emergent serum bicarbonate determination, Hemoglobin, The hepatorenal sugar, High-frequency oxygen, Color Doppler routine inspection, Discharge	0.377
2–2	Admission, Conventional ECG, Glucose determination, Venipuncture catheter, Laser therapy, 24-hour Holter, Low-frequency pulse electric treatment, Stool test + OB, Electrolyte, Hepatorenal sugar lipase, Emergent PT, Emergency ultra-sensitivity CRP, Emergent blood test, Thyroid function 7 items, Urine test + sediment, Analysis of urine microalbumin, Coagulation + D-dimer, Blood test + ultra-sensitivity CRP, Hemorheology full set, Inflammation 3 items, Chronic brain stimulation, Color Doppler routine inspection, Intracranial Doppler flow imaging, Determination of left ventricular function, Emergent kidney and sugar, Anemia 3 items, Tumor 10 items, Discharge	0.623

the number of latent treatment patterns is chosen so that it balances model complexity and fitness according to perplexity. As shown in Fig. 5, the perplexity decreases quickly before stabilizing. Note that the greater the value of K, the more likely the model over fits the data and more sampling computation and storage are required. The general rule of thumb is to choose a balance between simplicity of the model and the degree of fitness. As shown in Fig. 5, when $K = 2$, the perplexity achieves a relatively stabilized value for both the intracranial hemorrhage log and the cerebral infarction log.

Pattern extraction

Let's now turn to the contents of the patterns, i.e. the learned activity labels that have a high probability of being part of a particular pattern. We examine clinical activities associated with each treatment pattern to evaluate the quality of discovered patterns. For each treatment pattern t, we list all activity labels a with $p(a|t) \geq 0.01$. As shown in Table 3, the content often represents a meaningful set of activity labels.

Pattern support is predicated on the learned model. The underlying assumption is that if clinical activities belonging to a certain treatment pattern occur more frequently, then this pattern has high support. For example, in Table 3, the support value of pattern 1–1 is 0.584. It indicates that about 58.4 % patient traces in the intracranial hemorrhage log support that treatment pattern.

In addition, the aforementioned analysis procedure was conducted to investigate whether our approach can group and classify various activities according to their clinical intention. Taking the collected intracranial hemorrhage log as an example, both derived patterns covered various clinical activities in the cerebral infarction CPs. The discovered patterns were related to activities used for cerebral hemorrhage medical treatment (patterns 1–1), and intracranial hematoma surgical treatment (pattern 1–2), respectively.

Note that each treatment pattern represents certain common properties, which reflects the pattern in CPs. The patterns populated with more activities do not necessarily correlate with the degree of shared commonality among activities. Finding out the exact meanings of the discovered patterns require additional information and domain-specific knowledge. For example, pattern 1–2 was significant in both subdural hematoma treatment purpose and epidural hematoma treatment purpose. This result also indicates that a pattern is not necessary associated with only one concept, and it could be related to several commonalities shared by clinical intentions. However, in order to discern the hidden meanings of a pattern, careful analysis and domain-specific knowledge are required, indicating that the presented method can be a powerful hypothesis generation tool to guide systemic investigation on the relationship between the discovered treatment patterns and clinical intentions.

Conclusion

In this paper, we have introduced a novel approach for discovering latent CP patterns from careflow logs. The main idea is to collect careflow logs and then estimate latent patterns for the collected logs based on Latent Dirichlet Allocation. In an evaluation using real-world careflow logs, we showed that our method can discover the gist and hidden treatment patterns for CPs. In addition, discovered patterns can be successfully used for grouping and identifying clinical activities within the same therapy and treatment intention. Last but not least, the findings provide a foundation for future research using our approach.

We believe that our approach is highly appealing for the field of CPA, and that so far we have only exploited some of its potential, e.g., the probabilistic nature of the approach allows for handling of concurrent and overlapping treatment patterns, and also correlations between these patterns. We consider these properties, together with the ability to decompose CPs into their low-level constituents, as a crucial advantage over traditional techniques for CP analysis and optimization. In addition, for our approach, discovering latent treatment patterns is not limited to exploring the intrinsic property, i.e., activity labels. For example, we can use alternatives such as the occurring time-stamps of activities, resources to perform clinical activities, and patient-specific information, etc. We will investigate this in the future work.

Discovered patterns can profitably be exploited as a basis for further CPA tasks, e.g., to measure and understand the similarities among patient traces, to rank and retrieval similar patient traces as references for a specific patient in his or her careflow, to obtain patient trace clusters in which a patient trace mixed with manifold treatment patterns should be put in multiple clusters, and to refine/redesign CPs based on the ascertained treatment patterns, etc. We will address these tasks in our future studies.

Acknowledgments This work was supported by the National Nature Science Foundation of China under Grant No 81101126. The authors would like to give special thanks to all experts who cooperated in the evaluation of the proposed method.

References

1. Lee, K. H., and Anderson, Y., The association between clinical pathways and hospital length of stay: A case study. *J. Med. Syst.* 31:79–83, 2007.

2. Wakamiya, S., and Yamauchi, K., What are the standard functions of electronic clinical pathways? *Int. J. Med. Inform.* 78(8):543–550, 2009.
3. Lenz, R., Blaser, R., Beyer, M., Heger, O., Biber, C., Aumlein, M., and Schnabe, M., IT support for clinical pathways-lessons learned. *Int. J. Med. Inform.* 76(3):S397–S402, 2007.
4. Schuld, J., Schaer, T., Nickel, S., Jacob, P., Schilling, M. K., and Richter, S., Impact of IT-supported clinical pathways on medical staff satisfaction. A prospective longitudinal cohort study. *Int. J. Med. Inform.* 80(3):151–156, 2011.
5. Lu, X., Huang, Z., and Duan, H., Supporting adaptive clinical treatment processes through recommendations. *Comput. Methods Prog. Biomed.* 107(3):413–424, 2012.
6. Tello-Leal, E., Chiotti, O., and Villarreal, P., Process-oriented integration and coordination of healthcare services across organizational boundaries. *J. Med. Syst.* 36(6):3713–3724, 2012.
7. Rebuge, A., and Ferreira, D. R., Business process analysis in healthcare environments: A methodology based on process mining. *Inf. Syst.* 37(2):99–116, 2012.
8. Huang, B., Zhu, P., and Wu, C., Customer-centered careflow modeling based on guidelines. *J. Med. Syst.* 36(5):3307–3719, 2012.
9. Lin, F., Chen, S., Pan, S., and Chen, Y., Mining time dependency patterns in clinical pathways. *Int. J. Med. Inform.* 62(1):11–25, 2001.
10. Huang, Z., Lu, X., and Duan, H., On mining clinical pathway patterns from medical behaviors. *Artif. Intell. Med.* 56(1):35–50, 2012.
11. Lenz, R., and Reichert, M., IT support for healthcare processes-premises, challenges, perspectives. *Data Knowl. Eng.* 61(1):39–58, 2007.
12. de Luc, K., Care pathways: an evaluation of their effectiveness. *J. Adv. Nurs.* 32(2):485–496, 2000.
13. Chen, C. (Cliff)., Chen, K., Hsu, C. Y., and Li, Y. C. (Jack)., Developing guideline-based decision support systems using protègè and jess. *Comput. Methods Prog. Biomed*. 102(3):288–294, 2011.
14. Isern, D., Sanchez, D., and Moreno, A., Ontology-driven execution of clinical guidelines. *Comput. Methods Prog. Biomed.* 107(2):122–139, 2012.
15. Reichert, M., Rinderle, S., and Dadam, P., Adept workflow management system: flexible support for enterprise-wide business processes. In: *The Third International Conference on Business Process Management.* pp. 370–379, 2003.
16. Dykes, P. C., Currie, L. M., and Cimino, J. J., Adequacy of evolving national standardized terminologies for interdisciplinary coded concepts in an automated clinical pathway. *J. Biomed. Inform.* 36(4–5):313–325, 2003.
17. Huang, Z., Lu, X., and Duan, H., Using recommendation to support adaptive clinical pathways. *J. Med. Syst.* 36(3):1849–1860, 2012.
18. Hunter, B., and Segrott, J., Re-mappling client journeys and professional identities: A review of the literature on clinical pathways. *Int. J. Nurs. Stud.* 45:608–625, 2008.
19. Gurzick, M., and Kesten, K. S., The impact of clinical nurse specialists on clinical pathways in the application of evidence-based practice. *J. Prof. Nurs.* 26:42–48, 2010.
20. Agrawal, R., Gunopulos, D., and Leymann, F., Mining process models from workflow logs, In: Schek, H. J., Saltor, F., Ramos, I., Alonso, G. (Eds.) *Sixth International Conference on Extending Database Technology.* pp. 469–483. London: Springer-Verlag, 1998.
21. Cook, J. E., and Wolf, A. L., Discovering models of software processes from event-based data. *ACM Trans. Softw. Eng. Methodol.* 7(3):215–249, 1998.
22. Yang, W., and Hwang, S., A process-mining framework for the detection of healthcare fraud and abuse. *Expert Syst. Appl.* 31(1):56–68, 2006.
23. van der Aalst, W. M. P., Weijters, A. J. M. M., and Maruster, L., Workflow Mining: Discovering process models from event logs. *IEEE Trans. Knowl. Data Eng.* 16(9):1128–1142, 2004.
24. BPMN. http://www.bpmn.org/, Last accessed on 2012-2-14.
25. Lang, M., Urkle, T. B., Laumann, S., and Prokosch, H.-U., Process mining for clinical workflows: challenges and current limitations, In: Andersen, S. K., Klein, G. O., Schulz, S., Aarts, J. (Eds.) *Proceedings of MIE2008 The XXIst International Congress of the European Federation for Medical Informatics*, pp. 229–234, 2008.
26. Mans, R., Schonenberg, H., Leonardi, G., Panzarasa, S., Cavallini, A., Quaglini, S., and vander Aalst, W., Process mining techniques: an application to stroke care. *Stud. Health Technol. Inform.* 136:573–C578, 2008.
27. Goedertier, S., De Weerdt, J., Martens, D., Vanthienen, J., and Baesens, B., Process discovery in event logs: An application in the telecom industry. *Appl. Soft Comput.* 11(2):1697–1710, 2011.
28. Blei, D.M., Ng, A.Y., and Jordan, M.I., Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022, 2003.
29. Newman, D., Asuncion, A., Smyth, P., and Welling, M., Distributed algorithms for topic models. *J. Mach. Learn. Res.* 10:1801–1828, 2009.