Summarizing clinical pathways from event logs
Zhengxing Huang, Xudong Lu*, Huilong Duan, Wu Fan

College of Biomedical Engineering and Instrument Science, Zhejiang University, Zhou Yiqing Building 510, Zheda road 38#, Hangzhou, Zhejiang 310008, China

1. Introduction

Clinical pathways, guided by evidence-based medicine (EBM), are widely used by hospitals as tools for administrating, automating, and scheduling patient treatment activities [1–7]. It has been proven that clinical pathways can break functional boundaries and offer an explicit process-oriented view of health-care where the efficient collaboration and coordination of physicians become the crucial issue [3–5,8–13]. In China, clinical pathways are expected to play an important role in hospital reform, and may achieve the government, the medical insurance organization, the patient, the medical personnel and the medical service organization altogether wins [14].

Clinical pathway analysis, as a pivotal issue in ensuring high-quality care service delivery and medical cost controlling in clinical pathways, is receiving increasing attention in the field of healthcare management. The work of clinical pathway analysis tries to capture the actionable knowledge, which most commonly represents the best practice for most patients most of the time in their therapy and treatment processes. Many techniques and tools have been applied for clinical pathway analysis [4,6,7,15]. For instance, commercial business intelligence tools are introduced in health-care...
environments to look at aggregated data seen from an external perspective, e.g., length of stay (LOS), charge, bed utilization, medical service levels, and so on. However, many of these tools can only answer relatively simple questions like that how the average LOS or charge is in clinical pathways, etc. As a matter of fact, clinical pathways are evolving with the rapid development of medical technologies. While health-care organizations typically have an oversimplify and incorrect view of the actual situations in clinical pathways, efficient analysis should help them keep tracing essential/critical medical behaviors of clinical pathways, and extract potential information, which may substantially improve clinical pathways. In this regard, it requires to look inside the pathway (e.g., what are essential/critical medical behaviors of a pathway? In which time interval a specific medical behavior should be performed? Etc.) at a very refine level [15].

To this end, the methods using data mining and machine learning technologies to analyze clinical pathways based on associated event logs are receiving gradual attentions in medical informatics [3,15]. These techniques are also called process mining [16–19]. Process mining techniques have been widely studied in the domain of business process management, which attempt to extract non-trivial and useful information from event logs [15,19]. One important aspect of process mining is control-flow discovery, i.e., automatically constructing a process model (e.g., a BPMN model [20]) 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 [19].

In practice, many hospital information systems can monitor various medical behaviors in patient treatment journeys, which produce large event logs where each event has an associated clinical activity type, an occurrence time stamp as well as other attributes. It is, therefore, possible to apply process mining techniques to extract non-trivial knowledge from these event logs and exploit these for further analysis. However, the diversity of medical behaviors in clinical pathways is far higher than that of common business processes. The use of traditional process mining techniques may generate spaghetti-like pathway models that are difficult to be comprehended by clinical experts [15,21], such incomprehensible models are either not amenable or lack of assistance to efforts of analysis and improvement of clinical pathways. In addition, existing process mining algorithms often produce excessive volume of pathway models that may overwhelm the analysts [15]. In particular, the meanings or significance of the discovered pathway models sometimes goes untold. The last and most important point is that although many process mining techniques can tell us what medical behaviors are frequently performed and in what order, they seldom provide information about what critical medical behaviors occur in which specific stages in clinical pathways for further analysis. As indicated in [15,22], the use of traditional process mining techniques though successful in discovering clinical pathway models can prove inadequate in clinical pathway analysis.

In most cases, clinical analysts would like to get a good first impression of a clinical pathway before exploring the details of the pathway. In practice, to get a good initial impression, we usually first consult the summary [23,24]. For example, before deciding whether to read through a paper, a reader often wants to first look at a short summary of its main ideas. Similarly, a clinical pathway summary, as providing an indication of a global structure of the pathway, and describing the essential/critical medical behaviors in specific time intervals over the time period of the pathway, can quickly provide high-quality and high-level information about the pathway. Therefore, researchers have raised a new major challenge in clinical pathway analysis described as how to provide concise and comprehensive summaries to the best support the analysis and improvement of clinical pathways. To meet this challenge, we need to bring up a new aspect of clinical pathway analysis, namely how to concisely summarize clinical pathways from collected event logs.

1. Clinical pathway summary

Table 1 shows a bronchial lung cancer clinical pathway specification recommended by Ministry of Health of China. In our opinion, it is a summary of the bronchial lung cancer clinical pathway. Such a summary reveals the following features of clinical pathways [5,25]:

- A clinical pathway is typically arranged in a day-by-day format, enumerating the essential/critical medical behaviors (namely multidisciplinary clinical activities) that are spread along a series of time intervals (e.g., ‘Admission’, ‘Pre-OP Day’, and ‘Operation (OP) Day’, etc., as shown in Table 1) from a patient’s admission to discharge. The medical behaviors in a specific time interval are composed of a specific clinical pathway segment, and can be identified as a particular clinical pathway stage (e.g., the clinical stage ‘Pre-OP Day’ has a time interval ‘Days 2–6’, as shown in Table 1). And each pair of adjacent clinical pathway stages can be connected or overlapping with each other along the time-line of the pathway. For example, the time intervals of ‘Admission’ and ‘Pre-OP Day’ are connected with each other (i.e., ‘Days 1’ and ‘Days 2–6’), and the time intervals of ‘Pre-OP Day’ and ‘Operation (OP) Day’ are overlapping with each other (i.e., ‘Days 2–6’ and ‘Days 4–7’), as shown in Table 1;
- The designated medical behaviors of a clinical pathway represent specific clinical activities (for example, a medical order, a radiological examination test, and so on) and serve as checkpoints for the quality of care in various clinical stages of the pathway;
- Temporal constraint is a major category of control information in clinical pathways. Critical activities such as ‘admission’, ‘surgery’ and ‘discharge’ usually involve in additional constraints to other medical behaviors of a pathway. However, unlike an explicitly-defined medical procedure, the medical behaviors in a specific clinical stage are loosely-restricted in temporal aspect, and may be performed in any order.

With regard to the features listed above, we here propose a novel method for constructing comprehensive clinical pathway summaries from event logs. Our goal is to partition a clinical event log into several segments that span the observed time period of the log, within each segment identify one or a set of frequent patterns that represent essential/critical medical behaviors in the segment.

In this study, we formally define the problem of finding the best segmentations of event logs as an optimization problem. In particular, we develop a polynomial-time algorithm that optimally solves the summarization problem above. The computational complexity of our algorithms depends on the total length $T$ of the observed time period of the pathway from the associated event log. Our approach endeavors to provide insight into which frequent medical behaviors are essential/critical in which time intervals over the observed time period of a pathway. Such information gives us a better understanding of a clinical pathway, facilitates its analysis and improvement.

1.2. Structure of the paper

The rest of the paper is organized as follows. We review the related work in Section 2. Section 3 formalizes the problem. In
Section 4, we present the details of our clinical pathway summarization approach. Section 5 experimentally evaluates our approach. Section 6 provides an outlook on how the presented approach contributes to clinical pathway analysis and improvement. Finally, Section 7 concludes with a discussion of directions for future research.

### 2. Related work

The researches related to IT-support clinical pathways [3,8,26,27] mainly focus on modeling and enactment perspectives. Among them are works pertaining to clinical pathways supported by clinical practical guidelines (CPGs). CPGs are primarily designed to support doctors to make medical decisions (e.g., diagnosis, treatment) by providing them with deliberately summarized clinical principles [26,27]. Various clinical guideline modeling languages and methods (e.g., Asbru, EON, CLIF, etc.) have been developed, most of which are based on the formalism of task network models [28] considering the temporal property of guidelines. CPG-based clinical pathway modeling attempts to support for reasonable clinical decision-making in the paradigm of evidence-based medicine in hope for improvement of patient care and outcomes [3,8]. Nevertheless, CPGs are the product of expert consensus while clinical pathways emphasize the collaborations of a multidisciplinary ward team, including physicians, nurses, medical staff as well as other medical resources such as equipments, information systems, etc. [29]. The difference in scope of concerning issues in practical medical settings can consequently, partly serve as an interpretation of deviations of a certain clinical pathway from CPGs due to, for example, administrative reasons within a health-care organization.

In clinical practice, many hospital information systems use the electronic clones of paper-based clinical pathway specifications to represent clinical pathway models. For example, Wakamiya and Table 1

| Table 1 |
| A portion of the bronchial lung cancer clinical pathway summary recommended by Ministry of Health of China. |

<table>
<thead>
<tr>
<th>Time</th>
<th>Admission (Day 1)</th>
<th>Pre-OP Day (Days 2-3)</th>
<th>Operation/OPD Day (Days 4-6)</th>
<th>Post-OP 1 (Days 7-10)</th>
<th>Post-OP 11 (Days 11-14)</th>
<th>Discharge Day</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Diagnostics and Procedures</strong></td>
<td><strong>Medical history and physical examination</strong></td>
<td><strong>Chest X-ray</strong></td>
<td><strong>Laboratory tests</strong></td>
<td><strong>Physical examination</strong></td>
<td><strong>Discharge</strong></td>
<td><strong>Length of stay</strong></td>
</tr>
<tr>
<td><strong>Medical history and physical examination</strong></td>
<td><strong>Chest X-ray</strong></td>
<td><strong>Laboratory tests</strong></td>
<td><strong>Physical examination</strong></td>
<td><strong>Discharge</strong></td>
<td><strong>Length of stay</strong></td>
<td></td>
</tr>
</tbody>
</table>

---

Yamauchi present an electronic system of using checklist to represent clinical pathway models. This approach less considers the explicit modeling of medical behaviors, outcomes, and the relations between medical behaviors and outcomes, in clinical pathways. Recently, workflow technologies that have mainly been developed in business process management settings are also brought into the modeling of health-care processes, among which are business process modeling notation (BPMN), Petri Net based workflow language, etc. However, while workflow technologies are well-performant approaches in supporting simple and predictable business processes, they seldom lay emphasis on clinical workflow based on clinical pathways, and are incompetent in dealing with complex and dynamically-changing clinical workflow. In clinical practice, a patient trace may deviate from the clinical pathway (it should be followed) due to the uncertainty of patient statuses. In fact, a patient-care trace is an individual one, which may undergo changes from time to time, and in some cases changes are continuous. Thus, an important issue is to handle “variances” between the predefined clinical pathways and the actual patient traces.

The variance handling of clinical pathways needs to adjust the execution of patient-care traces. Many reasons may cause the deviations from a standardized pathway during patient-care traces, such as infection or poisoning. Thus, predefined clinical pathway models cannot provide necessary support for automatic recognition in exceptional situations. That is to say, there are cases that need adaptive clinical pathways. Many researchers have proposed different ways of addressing this problem, and some of these ideas have been implemented in commercial or academic systems. For example, Alexandrou et al., propose a rule-based clinical pathway adaptation method, which can reconfigure the clinical pathways in their executions by reasoning over the rules for the next steps of the treatment. Reichert et al., proposed a ADEPT system to offer the functionality of dynamic changes in the predefined health-care processes on execution time. Lu et al., proposed a hybrid approach of using rough set theory, and case based reasoning to recommend appropriate medical behaviors in clinical pathways by adopting a similarity measure to select appropriate medical behaviors executed on patients who presented similar features to the current one. Such medical behaviors are then applied to suggest which actions to perform next in clinical pathways. Note that these approaches assume that clinicians are independent and identically distributed. This assumption ignores the connections among physicians and is not consistent with clinical practice. In clinical practice, clinicians often rely on suggestions from others whom they trust and are experienced to solve a particular clinical problem. To this end, a collaborative service to support adaptive clinical pathways is presented in, which can guide clinicians in clinical pathways by providing recommendations on possible next steps based on the measurement of the target patient status and medical knowledge from past patient traces.

Although a well-defined body of knowledge for clinical pathway modeling and enactment has been developed, designing and comprehending clinical pathways are often difficult, time-consuming and error-prone tasks, which typically requires the intervention of a medical consultant armed with an appropriate level of experience and knowledge on the clinical pathway itself. This professional summarization of clinical principles involves in comprehensive understanding of medical knowledge, deep insight into coordination between departments, effective management of pertaining staff and, certainly, a formal representation of clinical pathways in order that the ultimate result can be widely understood and readyly executed. In addition, the design of clinical pathways is a continuous proceeding of revision and modification on the previous version and thus calls for the state-of-the-art knowledge and other feedbacks. In clinical environment, commercial Business Intelligence (BI) and Business Activity Monitoring (BAM) tools have been used to analyze clinical pathways, which typically look at aggregated data seen from the measures, e.g., LOS, mortality, and infection rate, etc. As valuable as these tools are, they restrict the attention to an external perspective of clinical pathway analysis.

As mentioned above, actual work can deviate from clinical pathway definitions due to many reasons, and it is very important for health-care organizations to discover these differences in order to improve the pathways. In this context, process mining, as a general method in business process analysis, is gaining increasing attention in analyzing clinical pathways and other kinds of health-care processes. The underlying idea of process mining is to discover process models in event logs that record their executions. Being transferred into medical settings, process mining methods may be applicable, for example, in retrieving frequent clinical pathway models from clinical event logs, which might be further utilized to refine the pathway itself. In fact, process mining has already been attempted in clinical environments by some researchers. Lin et al., reported a technique that was developed to discover the time dependency pattern of clinical pathways for managing brain stroke. The mining of time dependency pattern is to discover clinical pathway models and to identify the dependent relation between clinical activities in a majority of cases. Yang et al., propose a process mining algorithm to facilitate the automatic and systematic detection of health-care fraud and abuse. In a methodology of using process mining techniques to support health-care process analysis is 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, we have developed a new process mining algorithm to discover a set of medical behavior patterns given an input event log and a minimum support threshold value, such that it can find what critical medical behaviors are performed and in what order, and provide comprehensive knowledge about quantified temporal orders of medical behaviors in clinical pathways.

However, the diversity and complexity of medical behaviors in clinical pathways are far higher than that of common business processes. Traditional process mining techniques have several problems and challenges when used for mining clinical pathway patterns:

- Many existing process mining algorithms summarize the underlying process models in a structured manner. However, the assumption of the processes taking place in a structured fashion is not valid for clinical pathways. Clinical pathways, as typical human-centric processes, always take place in an unstructured fashion. In practice, many medical behaviors in a specific clinical stage can occur arbitrarily without a particular order. Bringing order in the chaos of clinical pathways probably requires a different mining strategy rather than existing process mining algorithms.
- In addition, although a patient trace is guided by a specific clinical pathway, it is possible that the events of that patient trace represents multiple underlying medical behavior patterns for specific treatment goals. For example, a bronchiolitis lung cancer patient with diabetes complication not only follows the bronchiolitis lung cancer pathway but also accepts treatment behaviors for his/her diabetes. In fact, even for the patients within the same clinical pathway, a slight dissimilarity of patient states may result in different medical behaviors. These variant behaviors occur commonly in clinical practice. As the obtained clinical event log contains patient traces that deal with a variety
of medical problems, it can be assumed that the log is actually generated by many underlying medical behavior patterns. Ferreira et al. [15,44,45] demonstrate that sequence clustering algorithms might be useful to detect clusters of frequently occurring sequences. However, due to the high number of underlying medical behavior patterns in clinical pathways, existing process mining algorithms did not lead to the identification of meaningful clusters of patient traces [42].

- To our best knowledge, existing process mining techniques cannot tell which time spans of clinical stages are comprised of the observed time period of a pathway, and what medical behaviors should be performed in which stages in the pathway. In many cases, clinical analysts may want to know essential/critical medical behaviors in a certain stage of a pathway before exploring the details of the pathway.

As a result, using existing process mining algorithms, it often either generate spaghetti-like clinical pathway models that are incomprehensible to health-care professionals, or usually discover excessive volume of clinical pathway models such that the meanings or significance of the discovered clinical pathway models sometimes goes untold. Such incomprehensible patterns are either incomprehensible to health-care professionals, or usually discover either generate spaghetti-like clinical pathway models that are not amenable or are lacking in assisting one in clinical pathway analysis and optimization efforts. Analyzing these models still requires significant effort. Therefore, it would be beneficial to give clinical analysts a first impression of a pathway by providing a concise and comprehensive summary. It is highly desirable to have such a summary for a pathway to suggest its potential meanings and to help clinical analysts decide whether and how to explore the pathway from the internal perspective.

Therefore, we argue that there is a critical need to develop new process mining techniques facilitating clinical pathway analysis by constructing summaries from associated event logs. Such summaries are short, accurate, and comprehensive for clinical analysts, and can be used as a feedback tool to assist with analysis and improvement of those already enacted clinical pathways.

3. Problem formalization

The starting point for the proposed approach is the concept of the clinical event log. We assume that it is possible to sequentially record all kinds of events in clinical workflow, and each clinical event refers to a specific case (i.e., a patient trace following a specific clinical pathway), is related to a particular activity type (i.e., a well-defined step in a specific clinical pathway) has an associated time-stamp, and other properties. Today, many hospital information systems record such information. Unfortunately, most systems lack or are not amenable in assisting one in clinical pathway analysis and optimization efforts. Analyzing these models still requires significant effort. Therefore, it would be beneficial to give clinical analysts a first impression of a pathway by providing a concise and comprehensive summary. It is highly desirable to have such a summary for a pathway to suggest its potential meanings and to help clinical analysts decide whether and how to explore the pathway from the internal perspective.

Therefore, we argue that there is a critical need to develop new process mining techniques facilitating clinical pathway analysis by constructing summaries from associated event logs. Such summaries are short, accurate, and comprehensive for clinical analysts, and can be used as a feedback tool to assist with analysis and improvement of those already enacted clinical pathways.

Definition 1 (Clinical event). Let \( E \) be the clinical event universe, \( A \) be the activity type domain, and \( T \) the time domain. We assume that it is possible to sequentially record all kinds of events in clinical workflow, and each clinical event refers to a specific case (i.e., a patient trace following a specific clinical pathway), is related to a particular activity type (i.e., a well-defined step in a specific clinical pathway) has an associated time-stamp, and other properties. Today, many hospital information systems record such information. Unfortunately, most systems lack or are not amenable in assisting one in clinical pathway analysis and optimization efforts. Analyzing these models still requires significant effort. Therefore, it would be beneficial to give clinical analysts a first impression of a pathway by providing a concise and comprehensive summary. It is highly desirable to have such a summary for a pathway to suggest its potential meanings and to help clinical analysts decide whether and how to explore the pathway from the internal perspective.

Therefore, we argue that there is a critical need to develop new process mining techniques facilitating clinical pathway analysis by constructing summaries from associated event logs. Such summaries are short, accurate, and comprehensive for clinical analysts, and can be used as a feedback tool to assist with analysis and improvement of those already enacted clinical pathways.

Definition 2 (Patient trace). Let \( E \) be the clinical event universe, \( A \) be the activity type domain and \( T \) the time domain. A patient trace is represented as a non-empty sequence of events associated with a particular patient, i.e., \( e = (e_1, e_2, \ldots, e_n) \), where \( \pi_{pid}(e_1) = \pi_{pid}(e_2) = \cdots = \pi_{pid}(e_n) \), and \( e_i \in E [1 \leq i \leq n] \).

Definition 3 (Clinical event log). Let \( \mathcal{C} \) be the set of all possible patient traces, and a clinical event log \( \mathcal{L} \) is a set of traces \( \mathcal{L} \) such that each event appears at most once in the entire log, i.e., for any \( \sigma_1, \sigma_2 \in \mathcal{C} : \forall e_1 \in \sigma_1 \cap \forall e_2 \in \sigma_2, e_1 \neq e_2 \) or \( \sigma_1 = \sigma_2 \).

As shown in Fig. 1, a log example of the bronchial lung cancer clinical pathway consists of six patient traces, i.e., \( \mathcal{L} = \{ \sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5, \sigma_6 \} \). The log contains 26 clinical activity types from a to z alphabetically, and spread along the observed time period of the log. Note that \( \mathcal{C} \) consists of clinical events that occur at discrete time points. Specifically, we consider the observed time period [1,7] in which time-stamps of clinical events are positive integers in [1,7], where \( T = \max \left\{ \sigma_{pid}(e_1) - \sigma_{pid}(e_1) + 1 \right\} \). That is, the observed time period consists of \( |T| \) different time-stamps at which clinical events of different activity types might occur. For example, given the log \( \mathcal{C} \) shown in Fig. 1, the observed time period is 33 days. Thus, the events of \( \mathcal{C} \) occur on \([1,33]\), which means that there are 33 time stamps at which these events may occur.

Definition 4 (Time interval, trace segment, activity set of a trace segment, and log segment). Given the observed time period [1,7] from the associated event log \( \mathcal{C} \), a time interval, denoted as \( t = [t_i, t_f] \) \((1 \leq t_i \leq t_f \leq 7)\), is a sub time period of \([1,7]\), \( 1 \leq t_i \leq t_f \leq 7 \), and \( t \) is the occurring time stamp of \( e \in T \). Taking the log shown in Fig. 1 as an example, and let \([1,3]\) be a specific time interval. The log segment associated with \([1,3]\) contains six trace segments \( \sigma_{\mathcal{L}[1,3]} \), \( \sigma_{\mathcal{L}[1,3]} \), \( \sigma_{\mathcal{L}[1,3]} \), \( \sigma_{\mathcal{L}[1,3]} \), \( \sigma_{\mathcal{L}[1,3]} \), \( \sigma_{\mathcal{L}[1,3]} \), where the activity sets of these trace segments are \( \Gamma(\sigma_{\mathcal{L}[1,3]}) = \{ a, b, c, d, e, f \} \), \( \Gamma(\sigma_{\mathcal{L}[1,3]}) = \{ a, d, b, c, e \} \), \( \Gamma(\sigma_{\mathcal{L}[1,3]}) = \{ a, d, b, c, o \} \), \( \Gamma(\sigma_{\mathcal{L}[1,3]}) = \{ a, b, c, o, q, m \} \) respectively.

Definition 5 (Clinical event log segmentation). Given a clinical event log \( \mathcal{L} \), and the observed time period \([1,3]\) of \( \mathcal{L} \). A segmentation of \( \mathcal{L} \) is a set of segments in \([1,3]\), i.e., \( \mathcal{L} = \{ L_1, L_2, \ldots, L_k \} \). Such a segmentation is defined by \( k \) contiguous and overlapping time intervals \( \{ t_{i_1}, t_{i_2}, \ldots, t_{i_k} \} \) over \([1,7]\), where \( t_{i_1} = 1, t_{i_k} = 7, \) and each \( t_i \) with \( 2 \leq i \leq k \), satisfying that \( 1 \leq t_{i-1} < t_i \leq t_{i+1} \), \( 1 \leq t_i \leq 7 \).
For example, a segmentation of the clinical event log shown in Fig. 1 is described in Fig. 2. The observed time period [1, 33] of the log is partitioned into four time intervals, and for each time interval there is a specific log segment.

Let us now introduce the concept of medical behavior pattern. Within the scope of behavior informatics [46], medical behavior refers to those clinical activities that present as patient-related therapy, treatment and care actions or events as well as activity sets/sequences conducted by clinicians, equipments, and medical services in health-care organizations. A medical behavior pattern is a normalized behavior structure that can effectively represent the common underlying activities in a specific time interval over the observed time period of a clinical pathway. Formally, we give the following definition.

**Definition 6 (Medical behavior pattern, pattern support).** Given a specific clinical event log $\mathcal{L}$, and the clinical activity domain $\mathcal{A}$, a medical behavior pattern $\phi$ is defined as a set of clinical activities such that $\phi \subseteq \mathcal{A}$. Furthermore, let $t_i$ be a specific time interval, and $\mathcal{L}_{t_i}$ be the log segment corresponding to $t_i$, the support of $\phi$ by $\mathcal{L}_{t_i}$, denoted $\text{supp}(\phi, \mathcal{L}_{t_i})$, is defined as follows:

$$\text{supp}(\phi, \mathcal{L}_{t_i}) = \frac{|\sigma(t_i) \cap \mathcal{L}_{t_i}|}{|\mathcal{L}_{t_i}|}$$

Taking the log shown in Fig. 1 as an example, let $\phi = \{a, b, c\}$ be a specific medical behavior pattern, which consists of three clinical activities, i.e., Admission, Color ultrasound examination, and ECG. If the time interval is $[1, 3]$, $\phi$ is supported by the partial traces $\sigma_a(1, 3], \sigma_b(1, 3], \sigma_c(1, 3], \sigma_d(1, 3]$, and $\sigma_e(1, 3]$ of the log shown in Fig. 1. Thus, the support of $\phi$ w.r.t $\mathcal{L}_{t_1}$ is 0.833. If the time interval is [3, 16], $\phi$ is not supported by the corresponding log segment $\mathcal{L}_{t_2}$, as a result, $\text{supp}(\phi, \mathcal{L}_{t_2}) = 0$.

A pattern $\phi$ is frequent for $\mathcal{L}_{t_i}$ if $\text{supp}(\phi, \mathcal{L}_{t_i})$ is no less than a predefined minimum support threshold value, denoted as $\text{minsupp}$. The objective of frequent medical behavior pattern mining is to find the set of patterns $\Phi$ in a specific log segment $\mathcal{L}_{t_i}$ which satisfy $\Phi = \{\phi | \text{supp}(\phi, \mathcal{L}_{t_i}) \geq \text{minsupp}\}$.

**Definition 7 (Representative error).** Let $\mathcal{L}$ be a clinical event log, [1, $T$] be the observed time period of $\mathcal{L}$, $t_i$ be a specific time interval ($t_i \subseteq [1, T]$), $\mathcal{L}_{t_i}$ be the log segment of $\mathcal{L}$ w.r.t $t_i$, and $\Phi_{t_i}$ be the set of frequent medical behavior patterns for $\mathcal{L}_{t_i}$. The representative error of $\Phi_{t_i}$ for $\mathcal{L}_{t_i}$, denoted as $\text{RE}(\Phi_{t_i}, \mathcal{L}_{t_i})$, is calculated as follows:

$$\text{RE}(\Phi_{t_i}, \mathcal{L}_{t_i}) = \sum_{\sigma(t_i) \in \mathcal{L}_{t_i}} \min_{\phi \in \Phi_{t_i}} (|\Gamma(\sigma(t_i))| - |\phi \cap \Gamma(\sigma(t_i))|)$$

where $|\phi \cap \Gamma(\sigma(t_i))|$ denotes the number of activities contained in both $\phi$ and $\Gamma(\sigma(t_i))$, and $\Gamma$ denotes the number of clinical events contained in $\mathcal{L}_{t_i}$. Note that each pattern in $\Phi_{t_i}$ represent essential/critical medical behaviors occurred in $t_i$. Such patterns are approximately similar to the occurred medical behaviors recorded in $\mathcal{L}_{t_i}$. Moreover, the union of the set medical behavior patterns derived from the specific log segments $\mathcal{L}_{t_{i1}}, \mathcal{L}_{t_{i2}}, \ldots, \mathcal{L}_{t_{ik}}$, denoted as $\Phi_{t_{i1}}, \Phi_{t_{i2}}, \ldots, \Phi_{t_{ik}}$, can effectively represent the common behaviors recorded in $\mathcal{L}$. The overall goal is to identify the set of time intervals over the observed time period [1, $T$] that partition $\mathcal{L}$ into a series of segments $\mathcal{L}_{t_{i1}}, \mathcal{L}_{t_{i2}}, \ldots, \mathcal{L}_{t_{ik}}$, and within each segment $\mathcal{L}_{t_{ik}}$ (1 $\leq i \leq k$) it identifies a set of frequent patterns that best represent the behaviors of $\mathcal{L}_{t_{ik}}$. The partitioning of $\mathcal{L}$ into a series of segments $\mathcal{L}_{t_{i1}}, \mathcal{L}_{t_{i2}}, \ldots, \mathcal{L}_{t_{ik}}$ and the corresponding sets of medical behavior patterns $\Phi_{t_{i1}}, \Phi_{t_{i2}}, \ldots, \Phi_{t_{ik}}$ construct a summary of a specific clinical pathway from $\mathcal{L}$.

We are now ready to give the formal definition of the clinical pathway summarization problem.

**Problem 1 (Clinical pathway summarization).** Given a clinical event log $\mathcal{L}$ and the observation time period [1, $T$] of a clinical pathway from $\mathcal{L}$, find a series of continuous and overlapping time intervals $t_{i1}, t_{i2}, \ldots, t_{ik}$ which partition $\mathcal{L}$ into a series of log segments $\{\mathcal{L}_{t_{i1}}, \mathcal{L}_{t_{i2}}, \ldots, \mathcal{L}_{t_{ik}}\}$, and discover a series of sets of frequent medical behavior patterns $\Phi = \Phi_{t_{i1}} \cup \Phi_{t_{i2}} \cup \cdots \cup \Phi_{t_{ik}}$ for the segments $\mathcal{L}_{t_{i1}}, \mathcal{L}_{t_{i2}}, \ldots, \mathcal{L}_{t_{ik}}$, respectively, such that the summarization error over the observed time period [1, $T$]

$$\text{SE}(1, T) = \sum_{i=1}^{k} \text{RE}(\mathcal{L}_{t_{i}}, \Phi_{t_{i}})$$

is minimized, where $\text{RE}(\mathcal{L}_{t_{i}}, \Phi_{t_{i}})$ is the representation error of the set of frequent medical behavior patterns $\Phi_{t_{i}}$ for the log segment $\mathcal{L}_{t_{i}}$. By this problem definition, we identify the optimal clinical pathway summarization through clinical event log segmentation. Note that we do not require the user to specify the set of time intervals over the observed time period [1, $T$] of $\mathcal{L}$. Essentially, this is done automatically, as the learned time intervals result in the minimum ratio of representative errors to total events.

**4. Method**

Now that we have defined how we can identify an optimal clinical pathway summary, we need a way to construct it. As indicated above, a clinical event log segmentation is defined as the partition of a log $\mathcal{L}$ into $k$ contiguous or overlapping segments covering all the events in the log such that each log segment is as homogeneous as possible. Despite the apparent interplay between the boundaries of the time intervals of the log segments, we can show that Problem 1 can be solved optimally in polynomial time.

Especially, we present our solution for the problem of clinical pathway summarization. As shown in Algorithm 1, given an input clinical event log $\mathcal{L}$ which records essential/critical medical behaviors of a pathway, the optimal $k$ log segments with their representative medical behavior pattern sets over the observed time period [1, $T$] of $\mathcal{L}$ are derived by using a specific clinical event log segmentation algorithm, e.g., the dynamic programming algorithm (DLog-Segmentation) [Line 15], which is introduced in Section 4.2.1. Note that the pseudo-code includes two major stages. The first stage is to generate the matrix of summarization error ratios, matrix, and to discover frequent medical behavior patterns from $\mathcal{L}$, for possible time intervals over the observed time period [1, $T$] (Lines 16–20) outputs a series of time intervals and the associated frequent medical behavior patterns, which bring up an optimal clinical pathway summarization.
Algorithm 1. Log-based clinical pathway summarization.

1: Procedure::CPSummarization($\mathcal{L}, \text{minsupp}$)
2: Input:
3: $\mathcal{L}$ is a clinical event log
4: $\text{minsupp}$ is a user-defined minimum support threshold value ($0 \leq \text{minsupp} < 1$)
5: Output:
6: $T = \varnothing$ is a series of time intervals for log segmentation
7: $\Phi(T) = \varnothing$ is the union of frequent medical behavior pattern sets associated with specific log segments
8: Steps:
9: Let $T = \varnothing$
10: Let $\Phi = \varnothing$
11: Let $[1,T]$ be the observed time period of $\mathcal{L}$
12: Let $\text{matrix}$ be the matrix of summarization error ratio
13: Let $TS = \varnothing$ be the optimal time-stamp quads
14: Using a specific log segmentation algorithm to partition $\mathcal{L}$, e.g.,
15: Let $\mathcal{S}$, $\Phi = \text{DPLogSegmentation}(\mathcal{L}, [1,T], \text{matrix}, \text{minsupp})$
16: Let $T \Leftrightarrow \text{OutputSegments}([1,T], TS)$ be a series of time intervals over $[1,T]$
17: For each $t \in T$
18: Extract frequent medical behavior patterns $\Phi_t$ from $\Phi$
19: Let $\Phi(T) = \Phi(T) \cup \Phi_t$
20: End For
21: Return $T$ and $\Phi(T)$
22: End Procedure

Procedure::OutputSegments ([t-, t+], TS)

Input:

23: $[t-, t+]$ is an observed time period
24: $TS$ is a set of time-stamp quads

Output:

25: $T$ is a set of time intervals for log segmentation

Steps:

26: Let $T = \varnothing$
27: Let $ts^- = t^-$ and $ts^+ = t^+$
28: Let $\text{flag} = \text{false}$ be a flag
29: For each element $(ts_1, ts_2, ts_3, ts_4) \in TS$
30: If $ts_1 = t^-$ and $ts_2 = t^+$
31: Break
32: End If
33: End For
34: If $\text{flag} = \text{true}$ and $t^- < ts^-$ and $t^+ > ts^+$
35: Let $T \Leftrightarrow T \cup \text{OutputSegments}([t-, t^+], TS)$
36: End If
37: Return $T$
38: End Procedure

Given the input log example shown in Fig. 1, the actual segmental results that our approach finds are presented in Fig. 2 (w.r.t a minimum support threshold value 0.33, which is introduced in Section 3). It illustrates what the output of the clinical pathway summarization approach conceptually looks like. The log is divided into four segments, identified by specific time intervals [1–3,16,17,19,20,33] over the time period [1,33], from which one or a set of frequent medical behavior patterns are discovered. For example, the underlying medical behavior patterns of the first log segment (identified by the time interval [1,2]) is $\{a, b, c, d, e\}$. Similarly, the underlying medical behavior pattern in the second log segment is $\{b, d, f, h, i, j, k, l, m, n, o, q\}$, in the third segment (j, k), and in the fourth segment $\{b, g, z\}$, $\{l, m, z\}$, and $\{l, r, z\}$. The details of log segmentation, and frequent pattern discovery are illustrated in Sections 4.2.1 and 4.2.2, respectively.

4.1. Dynamic programming for log segmentation

The clinical event log segmentation problem can be solved optimally using dynamic programming, which is introduced as follows.

Theorem 1. Given a clinical event log $\mathcal{L}$ and its observed time period $[1,T]$, for any time interval $t_i \subseteq [1,T]$, let $\mathcal{L}_t$ be the corresponding log segment, and $\Phi_t$ be the set of frequent medical behavior patterns discovered from $\mathcal{L}_t$.

Let $RE(\mathcal{L}_t, \Phi_t)$ be the representation error of $\Phi_t$ for $\mathcal{L}_t$. Then, Problem 1 can be solved optimally by evaluating the following dynamic programming recursion. For a pair of time stamps $i$ and $j$ ($1 \leq i < T$ and $1 \leq j < i + 1$), the optimal log segmentation can be defined as:

$$SE_{[i,j]} = \min_{1 \leq t \leq j - 1} \{SE_{[i,t]} + RE(\mathcal{L}_{i,T}^t, \Phi_{j}^{i,T})\}$$

where $SE_{[i,j]}$ indicates the optimal segmentation towards the observed time period $[i,j]$ of $\mathcal{L}$. 


Proof. Eq. (4) is a standard dynamic programming recursion that for every $i$ ($1 \leq i \leq T$) goes through all $j$'s ($1 \leq j \leq i + 1$) and evaluates the summarization error $SE_{1,i}$ and $RE(L_{j,T}, \phi_{j,T})$. There are most $i + 1$ ($1 \leq j \leq i + 1$) possible time stamps, which can be used for the last segment that, when having no overlap, would start at time stamp $i + 1$ and end at time stamp $T$. $SE_{1,i}$ is a lookup of an already computed value. Therefore, the only additional computation that needs to be done is the evaluation of $RE(L_{j,T}, \phi_{j,T})$, which can be optimal if a set of frequent medical behavior patterns are derived from $L_{j,T}$. Thus, the evaluation of Eq. (4) is also optimal. □

We developed the dynamic programming algorithm that implements Eq. (4), as shown in Algorithm 2. The inputs of Algorithm 2 are: a specific clinical event log $L$ describing the medical behaviors for a specific clinical pathway, the observed time period $[s, s']$, and a matrix of summarization error ratios matrix associated with possible segments over $[s, s']$. The outputs of Algorithm 2 are a set of time-stamp quads $TS$ to reserve the observed time period segmentation records, and a union of sets of frequent medical behavior patterns $\Phi$ for possible log segments over $[s, s']$. Note that each element $ts$ in $TS$ consists of four timestamps, i.e., $ts = (ts_1, ts_2, ts_3, ts_4) \ (ts_1 \leq ts_4 \leq ts_3 \leq ts_2)$. $ts_1$ and $ts_2$ are
composed of a specific time period \([t_s, t_s]_i\), \(t_s\) and \(t_s\) are composed of a specific time interval \([t_s, t_s]_j\) that could be segmented further iff \(t_s < t_s\) and \(t_s < t_s\). And \(t_s\) and \(t_s\) are composed of a specific time interval \([t_s, t_s]_k\) corresponding to a minimum segmentation error ratio and should not be segmented further. The pseudo-code is to generate the summarization error ratio for \(L\) over the time period \([t^-, t^+]\) (Lines 15–35). The time period \([t^-, t^+]\) is partitioned into two overlapping time intervals, i.e., \([t^-, i]\) and \([i, t^+]\) \((t^- < i < t^+\) and \(t^- < j < i + 1\)). The time interval \([t^-, i]\) is segmented optimally using dynamic programming (Lines 17–19) to calculate its summarization error ratio \(SER_{t^-i}\), while the set of frequent medical behavior patterns \(\phi_{[t^-, t^+]}\) are generated from \(L_{t^-i}\) (Line 23). Then, the summarization error ratio \(SER_{t^-t^+}\) is calculated by Eq. (4) (Line 26). Note that the process starts from the start time stamp \(t^-\), traces through the matrix step by step, and stops at the end time stamp \(t^+\).

### 4.1.2. Discovering frequent medical behavior patterns

Given a specific log segment \(L_{ts}\), we discover frequent medical behavior patterns from \(L_{ts}\). In clinical practice, various kinds of patients may accept different medical behaviors even in the same clinical stage (e.g., admission, pre-operation, post-operation, discharge, etc.) of a specific pathway. Therefore, it may discover multiple frequent behavior patterns from \(L_{ts}\). As mentioned above, a medical behavior pattern is a set of clinical activities that appear frequently together in \(L_{ts}\), which provides a higher-level abstraction on medical behaviors in a particular clinical stage of a pathway. In this sense, if multiple activities frequently appear together with each other in a specific clinical stage of a pathway, it would be a representative medical behavior pattern described by the co-occurred activity types. Such set of activity types, as a representative behavior pattern, can be discovered using existing frequent-pattern mining methods such as Apriori [47] and FP-growth [48].

#### Algorithm 2. A dynamical programming algorithm for the clinical event log segmentation.

1: **Procedure:** DPLogSegmentation\((L, [t^-, t^+], \text{matrix}, \text{minsupt})\)
2: **Input:**
3: \(L\) be a clinical event log
4: \([t^-, t^+]\) is the observed time period
5: \(\text{matrix}\) is the minimal summarization error matrix
6: \(\text{minsupt}\) is a user-defined medical behavior pattern support threshold \((0 \leq \text{minsupt} \leq 1)\).
7: **Output:**
8: \(TS\) is a set of time-stamp quads
9: \(\phi\) is a union of sets of frequent medical behavior patterns associated with \([t^-, t^+]\)
10: **Steps:**
11: Let \(\phi_{[t^-, t^+]} = \text{MBPGeneration}(L_{[t^-, t^+]}, \text{minsupt})\) be the set of frequent medical behavior patterns of \(L_{[t^-, t^+]}\)
12: Let \(\phi = \phi_{[t^-, t^+]}\)
13: Let \(\text{matrix}_{[t^-, t^+]} = \text{RE}(L_{[t^-, t^+]}, \phi_{[t^-, t^+]})\) be the representative errors of \(\phi_{[t^-, t^+]}\) for \(L_{[t^-, t^+]}\)
14: Let \(t_s_1 = t^-\) and \(t_s_2 = t^+\) be two specific time stamps
15: Let \(i = t^-\)
16: While \(i < t^+\)
17: If \(\text{matrix}_{[t^-, i]} \equiv \text{null}\)
18: Let \(\phi_{[t^-, i]} = \text{DPLogSegmentation}(L, [t^-, i], \text{matrix}, \text{minsupt})\)
19: End If
20: Let \(j = t^-\)
21: While \(j < i + 1\)
22: Let \(L_{j+t} = \text{a log segment of } L\text{ in }[j, t^+]\)
23: Let \(\phi_{[j+t]} = \text{MBPGeneration}(L_{j+t}, \text{minsupt})\) be the set of frequent medical behavior patterns discovered from \(L_{j+t}\)
24: Let \(\phi = \phi_{[j+t]} \cup \phi_{[j+t]}\)
25: Let \(\text{RE}(L_{j+t}, \phi_{[j+t]})\) be the representative errors of \(\phi_{[j+t]}\) for \(L_{j+t}\)
26: Let \(\text{SE}_{[t^-, t^+]} = \text{matrix}_{[t^-, i]} \times + \text{RE}(L_{j+t}, \phi_{[j+t]})\)
27: If \(\text{matrix}_{[t^-, t^+]} > \text{SE}_{[t^-, t^+]}\)
28: Let \(\text{matrix}_{[t^-, t^+]} = \text{SE}_{[t^-, t^+]}\)
29: Let \(t_s_1 = i\) and \(t_s_2 = j\)
30: Let \(\phi_i = \phi_{[j+t]}\)
31: End If
32: \(j = j + 1\)
33: End While
34: \(i = i + 1\)
35: End While
36: Let \(TS = TS_{[t^-, i]} \cup \{(t^-, t^+, t_s_1, t_s_2)\}\)
37: Return \(\phi\) and \(TS\)
38: **End Procedure**
frequent patterns, if a medical behavior pattern is frequent, then all subsets of that pattern are also frequent [47]. As a matter of fact, if all subsets of a certain pattern have the same support value, then they always represent the same medical behaviors in a specific clinical stage of a specific pathway. In this case, the specific pattern may be more valuable than each sub-pattern derived from the specific one. Thus, we prune such sub-patterns whose support values are the same as one another.

Algorithm 3. Frequent medical behavior pattern generation.

1: Procedure::MBPGeneration($\mathcal{L}_t$, minsupp)
2: Input:
3: $\mathcal{L}_t$ is a log segment of a specific clinical event log $\mathcal{L}$ in a specific time interval $ti$
4: minsupp is a user-defined medical behavior pattern support threshold ($0 \leq \text{minsupp} \leq 1$).
5: Output:
6: $\Phi_t$ is a set of frequent medical behavior patterns discovered from $\mathcal{L}_t$
7: Steps:
8: Let $\Phi_t = \emptyset$
9: Scan $\mathcal{L}_t$ to find $A[1]$, which is the set of frequent 1-patterns of clinical activity types
10: Let $\Phi_t = A[1]$
11: Let $PID[1] = \emptyset$ be the patient identifier list of each frequent 1-pattern.
12: For each $a \in A[1]$
13: Scan $\mathcal{L}_t$ to collect $PID_a$, which is the patient identifier list of frequent 1-pattern $a$
14: Let $PID[1] = PID[1] \cup \{PID_a\}$
15: End For
16: Let $k = 2$
17: Let $A[k] = \emptyset$ be the set of frequent $k$-patterns.
18: Let $PID[k] = \emptyset$ be the patient identifier list of each frequent $k$-pattern.
19: While $A[k-1] \neq \emptyset$
20: For each $p \in A[k-1]$ and $q \in A[k-1]$, where $p[1] = q[1], \ldots, p[k-2] = q[k-2], p[k-1] < q[k-1]$
21: Let $c = p[1], p[2], \ldots, p[k-1], q[k-1]$ be the candidate $k$-pattern
22: Let $PID_c = \text{intersection}(PID_p, PID_q)$ be the patient identifier list of $c$
23: If $\text{sup}(PID_c) \geq \text{minsup}$
24: Let $A[k] = A[k] \cup \{c\}$
25: Let $PID[k] = PID[k] \cup \{PID_c\}$
26: If $|PID_p| \equiv |PID_q|$
27: Let $\Phi_t = \Phi_t - \{p\}$
28: End If
29: If $|PID_p| \equiv |PID_q|$
30: Let $\Phi_t = \Phi_t - \{q\}$
31: End If
32: End If
33: End For
34: Let $\Phi_t = \Phi_t \cup A[k]$
35: Delete $PID[k-1]$
36: Let $k = k + 1$
37: End While
38: Return $\Phi_t$
39: End Procedure

As discussed above, frequent medical behavior pattern discovery is one classical frequent pattern mining problem. In this study, we present an Apriori-based algorithm to discover frequent medical behavior patterns from a specific log segment associated with a specific time interval. The detailed process is described in Algorithm 3. In the beginning, the algorithm scans the whole log segment to filter out all frequent 1-patterns of clinical activity types with their supports (Line 10). Then, the log segment is scanned again to record patient identifier list for each frequent 1-pattern (Lines 12–15). The patient identifier list of a particular pattern contains all patient identifiers of patient traces in which this pattern occurs, and the length of the patient identifier list is its support. So when calculate the support of a 2-pattern, we can simply intersect the two patient identifier lists of the corresponding 1-patterns composing that 2-pattern (Line 22). Because the intersection contains the patient identifiers of all traces in which both the two 1-patterns occur. It is the same way to get all other frequent k-patterns ($k \geq 2$) (Lines 19–37). Note that if a ($k - 1$)-pattern has the same support with its supper k-pattern, the ($k - 1$)-pattern will be removed from the return pattern list $\Phi_t$ (Lines 26–31).

Using the presented algorithm, we can discover a set of frequent medical behavior patterns form a specific log segment. Taking Fig. 1 as an example, suppose the time interval is $[1, 2]$, the set of frequent patterns can be derived from $\mathcal{L}_{1,2}$, i.e., $\{a, b, c, d, e\}$, as shown in Fig. 2.

4.2. Proof-of-concept prototype

We have implemented and tested the proposed approach using Microsoft C#. Fig. 3 depicts a screen-shot of our prototype. We used the clinical event log of the bronchial lung cancer pathway to describe the proposed approach. Based on the input log extracted from Zhejiang Huazhou Central hospital of China, we can
Fig. 3. A screen-shot of the system prototype.

Fig. 4. Running times for the collected logs.
construct a summary about the bronchial lung cancer pathway, which consists of a series of segments with frequent medical behavior patterns. The clinical pathway summary was generated automatically using our approach, as shown in Fig. 3.

The algorithms mentioned above have been implemented in the prototype. Some basic information about the log such as the trace numbers, the event numbers, the activity type numbers, and so on, are given on the left-bottom of Fig. 3, while the screen-shot of the derived summary shows on the right of Fig. 3. The generated summary indicates the actual medical behaviors being applied in clinical pathways.

5. 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.

5.1. Experimental design

The experimental data set was extracted from Zhejiang Huzhou Center 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 clinical processes in the hospital. In the experiments, we extracted four specific event logs about bronchial lung cancer, colon cancer, gastric cancer, 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 LOS) from the collected log. The details of reserved logs are shown in Table 2, including the patient trace number, clinical event number, clinical activity number, the average LOS, the minimum LOS, and the maximum LOS of each log. For example, the cerebral infarction event log consists of 445 patient traces. The average LOS of these traces is 17.2 days while some traces take a very short time, e.g., only one day in hospital, and other traces take much longer.

Table 2
Six diseases' event logs used in the experiments.

<table>
<thead>
<tr>
<th>Disease</th>
<th>Trace #</th>
<th>Event #</th>
<th>Activity #</th>
<th>Average LOS (days)</th>
<th>Min LOS (days)</th>
<th>Max LOS (days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bronchial lung cancer</td>
<td>48</td>
<td>3405</td>
<td>225</td>
<td>28.2</td>
<td>15</td>
<td>42</td>
</tr>
<tr>
<td>Colon cancer</td>
<td>52</td>
<td>4840</td>
<td>292</td>
<td>24.1</td>
<td>12</td>
<td>40</td>
</tr>
<tr>
<td>Gastric cancer</td>
<td>100</td>
<td>8024</td>
<td>274</td>
<td>24.6</td>
<td>1</td>
<td>66</td>
</tr>
<tr>
<td>Cerebral infarction</td>
<td>445</td>
<td>23106</td>
<td>513</td>
<td>17.2</td>
<td>1</td>
<td>173</td>
</tr>
</tbody>
</table>

Fig. 5. Summarization error for the collected logs.
longer, e.g., more than 170 days in the hospital, which implicitly indicates the diversity of medical behaviors in cerebral infarction clinical pathways.

All experiments were performed on an Intel Pentium IV-2.8G Windows 7 system with 4G byte main memory. The algorithms were implemented using Microsoft C#. All run-time in the figures are in seconds.

5.2. Results

In this subsection, we study the impact of the minimal support threshold value \( \text{minsupp} \) on the time-consuming for the computations of each event log. We vary the value of \( \text{minsupp} \) from 0.5 to 1.0 (for cerebral infarction log, the value of \( \text{minsupp} \) from 0.3 to 1.0). Fig. 4 shows a curve using the proposed approach (denoted by ‘proposed’). The general trends of running times are observed from all logs. For example, as shown in Fig. 4(A), the running times of the curve remain stable with the initial decrease of \( \text{minsupp} \). And when \( \text{minsupp} \) is less than a certain value, i.e., \( \theta \leq 0.65 \) for the bronchial lung cancer log, the running times of the curve quickly increase with the further decreases of \( \text{minsupp} \). In addition, the summarization error (SE) achieved for the collected logs is also reported in Fig. 5. The SE diseases linearly with the diseases of \( \text{minsupp} \) for the collected logs.

Next, we study how the proposed approach performs with the increasing size of a clinical event log. Fig. 6 shows how the presented approach scales up as the number of input-clinical pathway traces is increased, from 100 to 445. Note that the experiments were performed on the cerebral infarction clinical event log with the same minimum support threshold of 0.5. The running times and the SER are normalized with respect to the time for the 100 input-traces. It can be observed that the presented algorithm has a linear scalability in terms of the run-times against the increasing number of traces.

In addition, we observed from the Fig. 6 that the SE achieved by the presented algorithm linearly increases with the increases of the number of input traces. Thus, In order to investigate the impact of the number of input traces on the summarization quality. We measure the following summarization error ratio:

\[
\text{SER}_{1} = \frac{\text{SE}}{\sum_{i=1}^{\text{|L|}} \sum_{j=1}^{\text{|T|}} |\sigma(t_i)|} \quad (5)
\]

where \( \sum_{i \in L, j \in T} |\sigma(t_i)| \) denotes the number of clinical events contained in \( L \). Note that \( \sum_{i \in L} \sum_{j \in T} |\sigma(t_i)| \geq \sum_{i \in L} |\sigma| \) due to the overlap of the log segments with each other.

SER is the ratio between the achieved summarization errors and the total event numbers in the log. The smaller SER indicates that more medical behaviors are covered by the derived summary, and the better summarization quality is achieved. As we observed from the Fig. 6(C) that the SER achieved by the presented algorithm is linearly decreased with the increases of the number of input traces. Note that the collected logs cover all of patient traces of the example diseases, i.e., bronchial lung cancer, colon cancer, gastric cancer, and cerebral infarction, in hospital information systems in two years (2007.08–2009.09). And it could not expect to collect
a huge data-set for a particular disease from one hospital. However, the results shown in Fig. 6 indicates that the performance of the proposed approach could be improved if more data for a specific disease are provided.

5.3. Example

In this section, we look at the segments produced by the presented approach and judge their practical utility. Especially, we take the bronchial lung cancer log as an example. The log segments generated by our approach with \( \minsupp = 0.5 \), as depicted in Table 4 (which is an English translation version of Fig. 7). We noticed that some essential clinical activities, e.g., admission, lung excision operation, and discharge, are correctly partitioned into different segments. In addition, we noticed that the most frequent activity types occurring in the first segment correspond to "Stool routine test and OB", "Liver and kidney function, blood glucose and fat examination", etc. On the other hand, the most frequent activity types occurring in the last segment correspond to "Atomizing inhalation", "Postoperative drainage", etc. Thirdly, the segments appeared in the earlier observed time period have more examinational activities than the segments appeared in the later observed time period of a clinical pathway, which indicates that most examinational activities are performed before the operation. Furthermore, the constructed summary indicates the general occurring time of specific activities, e.g., the operation is routinely performed during 2–14 days. These results highlight the fact that the constructed clinical pathway summaries allow clinical analysts to explore the available data without getting overwhelmed by it.

Now we use a simple example to illustrate how these discovered summaries can contribute to clinical pathway analysis and improvement. As shown in Tables 1, 3 and 4, they are fragments of the bronchial lung cancer pathway recommended by the Chinese Ministry of Health, discovered by the presented approach from the collected event log and designed by clinical experts of Zhejiang Huzhou Central hospital, respectively. The three pathway fragments consist of a set of segments, of which the frequent medical behavior patterns are listed. Since our approach gathers information about the actual clinical workflow instead of the alleged pathway, some differences may be expected to exist between the pathway identified by the logs, and the formal description of the pathway provided by available documentation. We can see that the results reveal the following differences from the three pathway fragments:

- They have different segments indexed with specific time intervals over the time-line of the pathway. For example, The recommended clinical pathway by the Chinese Ministry of Health lists six segments, i.e., Admission (Day 1), Pre-OP Day (Days 2–6), Operation (OP) Day (Days 4–7), Post-OP I (Days 5–8), Post-OP II (Days 6–12) and Discharge (Days 13–21). In the clinical expert designed pathway, there are 9 segments, i.e., Admission (Day 1), Pre-OP (Days 2–4), Operation (OP) Day (Day 5), Post-OP I (Day 6), Post-OP II (Day 7), Post-OP III (Day 8), Post-OP IV (Day 9), Post-OP V (Day 10) and Discharge (Day 11). While in the discovered clinical pathway summary, there are three segments, i.e., Admission (Days 1), Operation (Days 2–14), and Discharge (Days 15–41).

- In addition, the segments of the three clinical pathways not only have different time intervals, but also contain specific medical behaviors. For instance, the medical behaviors of the first segment "Stool routine test and OB", "Liver and kidney function, blood glucose and fat examination", etc. On the other hand, the medical behaviors of the first segment in the constructed clinical pathway summary are activities "Stool routine test and OB", "Liver and kidney function and blood glucose examination", etc.
Furthermore, some essential medical behaviors might be performed in specific time intervals. For example, the activity "Surgery" is assumed to be performed in the time interval "days 4–7", "day 5", and "days 2–14", w.r.t. the recommended clinical pathway by the Chinese Ministry of Health, the clinical expert designed pathway, and the constructed summary, respectively. The results described above confirm the assumption that deviations often happen between actual behaviors and expected behaviors in clinical pathways. As a matter of fact, comparing actual behaviors with predefined behaviors of a clinical pathway brought out substantial differences and highlighted the potential gains of clinical pathway analysis and improvement. We must mention that the presented methodology in this paper, as a novel process mining technique, is strongly data dependent. Although the structure of the discovered clinical pathway summary (c.f. Table 4) is far away from either the recommended clinical pathway by Chinese Ministry of Health (c.f. Table 1) or the predefined pathway by Zhejiang Huzhou Central hospital (c.f. Table 3), they faithfully reflect the real information recorded in clinical workflow logs. We have discussed the discovered results with clinical managers of Zhejiang Huzhou Central hospital. They pointed out two possible reasons for these deviations.

- For one thing, there are many factors and uncertainties affecting the execution of clinical pathways in practice. The uncertainties can result from inter-observer variability, inaccurate evaluation of the patient and some deficiencies in grading scales. As a result, the patient care-flow may not go towards the expected direction, and different from the predefined clinical pathways. In such cases, the variations happen inevitably in clinical pathway execution. And discovered results, as a reflection of real situation, might be very valuable references for clinical managers in order to redesign clinical pathways.

- For the other thing, performed medical behaviors by physicians are not standardized and normalized currently in the hospital. Many predefined behaviors are changed, postponed, or canceled, arbitrarily in patient care-flow. Note that this is a main reason that the Chinese government asks to implement standardized clinical pathways in hospitals in order to normalize medical behaviors, control the budget for each medical behavior, and improve the quality of patient care-flow management, etc.

These discovered summaries, as a reflection of actual medical behaviors in clinical workflow, may reveal interesting information about patient therapy and treatment, and can be used as a feedback tool that supports auditing, analyzing and improving enacted clinical pathways.

### Table 4

A portion of the bronchial lung cancer clinical pathway summary discovered by the presented approach from the collected event log.

<table>
<thead>
<tr>
<th>Admission (Days 1)</th>
<th>Operation (Days 2-14)</th>
<th>Discharge (Days 15-41)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pattern 1</td>
<td>Pattern 1</td>
<td>Pattern 1</td>
</tr>
<tr>
<td>□ Admission</td>
<td>□ Atomising inhalation</td>
<td>□ Discharge</td>
</tr>
<tr>
<td>□ ECG</td>
<td>□ Postoperative drainage</td>
<td>□ Electrolyte</td>
</tr>
<tr>
<td>□ Stool routine test</td>
<td>□ Liver, kidney function, blood glucose and fat examination</td>
<td>□ Liver and kidney function, blood glucose and fat examination</td>
</tr>
<tr>
<td>and OB</td>
<td>□ Electrolyte</td>
<td>□ Blood routine and C-reactive protein test</td>
</tr>
<tr>
<td>□ Liver and kidney function, blood glucose and fat examination</td>
<td>□ Blood routine and C-reactive protein test</td>
<td></td>
</tr>
<tr>
<td>□ Urinary routine and sediment examination</td>
<td>□ Blood routine and C-reactive protein test</td>
<td></td>
</tr>
<tr>
<td>□ Coagulation function</td>
<td>□ Blood routine and C-reactive protein test</td>
<td></td>
</tr>
<tr>
<td>and D-Dimer detection</td>
<td>□ Blood routine and C-reactive protein test</td>
<td></td>
</tr>
<tr>
<td>□ Electrolyte</td>
<td>□ Blood routine and C-reactive protein test</td>
<td></td>
</tr>
<tr>
<td>□ Blood routine and C-reactive protein test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>□ Ten tumor examinations</td>
<td>□ Blood routine and C-reactive protein test</td>
<td></td>
</tr>
</tbody>
</table>

6. Discussion

We would like to mention that the constructed pathway summaries from event logs have been evaluated by hospital managers and clinical experts at the Zhejiang Huzhou Central Hospital of China, who understand the beneficial effects of discovered summaries. They indicate that the constructed summaries from event logs support clinical pathway (re)design and improvement. Note that in most cases, prior to the deployment of a clinical pathway in practice, a clinical workflow was already there. Redesigning an existing workflow is influenced by perceptions, e.g., the designed clinical pathway specifications are often normative in the sense that they state what should be done rather than describing the actual workflow. As a result, the designed pathway specifications tend to be rather subjective. A more objective way of designing is to use clinical events related to the actual medical behaviors that took place. Despite that the proposed clinical pathway summarization approach is not a tool for designing clinical pathways, it is evident that a good understanding of the existing workflow is vital for any design and improvement effort. Since event logs are reserved in most hospital information systems, the collected logs can be used to derive clinical pathway summaries explaining the events recorded. Besides, such summaries are not biased by perceptions, and are useful to confront with the man-made clinical pathway specifications. Thus, it is effective to be used in clinical pathway analysis and improvement.

Note that, in real-life situations, physicians or hospital managers are using specific clinical pathway specifications to treat patients more frequently [3]. These base specifications can automatically suggest the clinical workflow for individual patients. Note that it must be possible to react to urgent situations and physicians can deviate from these basic specifications when necessary. Clearly, it is desirable to keep these specifications aligned with the actual clinical workflow. Even if the actual workflow is fully compliant with a clinical pathway specification, it is still interesting to
see how frequently certain medical behaviors in the specification are actually used, and to remove obsolete behaviors without weakening its representativeness, which otherwise should be maintained. Since the frequent behaviors of treating individual patients can be discovered by the proposed approach, this suggests that it is increasingly more interesting to compare pre-designed clinical pathway specifications with the constructed summaries from event logs. The results of this analysis can assist physicians in continuously analyzing and improving clinical pathways. Although the main motivation for our work is the forensic analysis of clinical event logs, the techniques presented herein can also be applied to business process management domain, in which event logs are commonly recorded by various information systems such that useful business process summaries can be constructed using our methodology.

7. Conclusion and future work

In this paper, we present an approach for constructing clinical pathway summaries from the collected event logs which regularly record various kinds of medical behaviors by hospital information systems. The experimental results show the promises of our approach to facilitate clinical pathway analysis. It also initiates an alternative implementation of clinical pathways in the hospital (e.g. constructing specific clinical pathway summaries to be surrogates instead of starting modeling from scratch). Obviously, the performance varied by the threshold considered. This highlights the fact that not only efficient approaches are needed to perform clinical pathway summarization, but also a reasonable threshold should be determined from experiments on a practical basis.

We believe that our approach is highly appealing for the field of clinical pathway analysis, 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 medical behavior patterns, and also correlations between these patterns. We consider these properties, together with the ability to decompose clinical pathways into their low-level constituents, as a crucial advantage over traditional techniques for clinical pathway analysis and optimization. In addition, for our approach, deriving clinical pathway summaries from event logs is not limited to exploring the intrinsic property, i.e., clinical activities. For example, we can use alternatives such as the resources to perform clinical activities, and patient-specific information, etc. We will investigate this in the future work.

Last but not least, the role of information technology in the clinical pathway is expected to become more prominent in order to norm the standard medical behavior, improve the quality of medical treatment, and promote medical expense rationalization, etc. In China, the implementation of the clinical pathway, as complied with the medical service reform of the country, has been widely used in hospital management in recent years. It is expected that the identified and the related technique challenges will be addressed by other researchers and developers, and we will reach a step closer to realizing the vision of comprehensive healthcare management and professionalism.

Acknowledgment

This work was supported by the National Nature Science Foundation of China under Grant No. 81110126. The authors are especially thankful for the positive support received from the Zhejiang Huzhou Central Hospital of China as well as to all medical staff involved. In addition, the authors would like to thank the anonymous reviewers for their constructive comments on an earlier draft of this paper.

References

[22] Kiernan J, Terzi E. Constructing comprehensive summaries of large event sequences. ACM Trans Knowledge Discov Data 2012(0).


