Artificial Intelligence in Medicine 56 (2012) 35–50

Contents lists available at SciVerse ScienceDirect

Artificial Intelligence in Medicine

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

On mining clinical pathway patterns from medical behaviors

Zhengxing Huang, Xudong Lu*, Huilong Duan

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

ARTICLE INFO

Article history:
Received 7 June 2011
Received in revised form 21 May 2012
Accepted 10 June 2012

Keywords:
Clinical pathway analysis
Pattern mining
Process mining
Clinical workflow log

ABSTRACT

Objective: Clinical pathway analysis, as a pivotal issue in ensuring specialized, standardized, normalized and sophisticated therapy procedures, is receiving increasing attention in the field of medical informatics. Clinical pathway pattern mining is one of the most important components of clinical pathway analysis and aims to discover which medical behaviors are essential/critical for clinical pathways, and also where temporal orders of these medical behaviors are quantified with numerical bounds. Even though existing clinical pathway pattern mining techniques can tell us which medical behaviors are frequently performed and in which order, they seldom precisely provide quantified temporal order information of critical medical behaviors in clinical pathways.

Methods: This study adopts process mining to analyze clinical pathways. The key contribution of the paper is to develop a new process mining approach to find a set of clinical pathway patterns given a specific clinical workflow log and minimum support threshold. The proposed approach not only discovers which critical medical behaviors are performed and in which order, but also provides comprehensive knowledge about quantified temporal orders of medical behaviors in clinical pathways.

Results: The proposed approach is evaluated via real-world data-sets, which are extracted from Zhejiang Huzhou Central hospital of China with regard to six specific diseases, i.e., bronchial lung cancer, gastric cancer, cerebral hemorrhage, breast cancer, infarction, and colon cancer, in two years (2007.08–2009.09). As compared to the general sequence pattern mining algorithm, the proposed approach consumes less processing time, generates quite a smaller number of clinical pathway patterns, and has a linear scalability in terms of execution time against the increasing size of data sets.

Conclusion: The experimental results indicate the applicability of the proposed approach, based on which it is possible to discover clinical pathway patterns that can cover most frequent medical behaviors that are most regularly encountered in clinical practice. Therefore, it holds significant promise in research efforts related to the analysis of clinical pathways.

1. Introduction

A clinical pathway, guided by evidence-based medicine (EBM) and clinical practice guidelines (CPGs), is a standardized and normalized therapy pattern and procedure constructed for a specific disease that follows the contemporary medical empirical data and clinical experts' experiences [1–4]. It has been proven that implementations of clinical pathways improve the quality of patient-care, provide an opportunity to identify good practice, remove bad practice, identify and apply evidence, identify education and training needs, and appreciate the skills and contributions of all professionals and care sectors [2,3,5–12].

Researchers in medical informatics and other relevant scientific areas have paid much attention to the areas of research, experiment, application and dissemination of clinical pathways, and continuously launch extensive research and project cooperation in this field. In particular, clinical pathway analysis, which is seen as a pivotal issue in ensuring specialized, standardized, normalized and sophisticated patient therapy procedures [2,3,6,12–16], is receiving increasing attention in the field of medical informatics.

Clinical pathway analysis is the process of (1) discovering knowledge about how clinical activities impact on patients in their care journeys, and (2) using the discovered knowledge for various applications, including clinical pathway (re)design, clinical pathway optimization, clinical decision support, medical deviation detection and business management, and so on [3,12,17,18]. In this study, we represent medical behaviors as flexible, transparent, and re-usable pieces of functionality that consist of one or several clinical activities required to set up a clinical solution. It would be interesting to know which medical behaviors are essential/critical for clinical pathways and where temporal orders of these medical behaviors are quantified with numerical bounds. For example, we would like to know if the radical resection of colon cancer surgery is a critical medical behavior with respect to the colon cancer

* Corresponding author. Tel: +86 571 87951792; fax: +86 571 87951960.
E-mail address: lxd@vico-lab.com (X. Lu).

0933-3657/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.artmed.2012.06.002

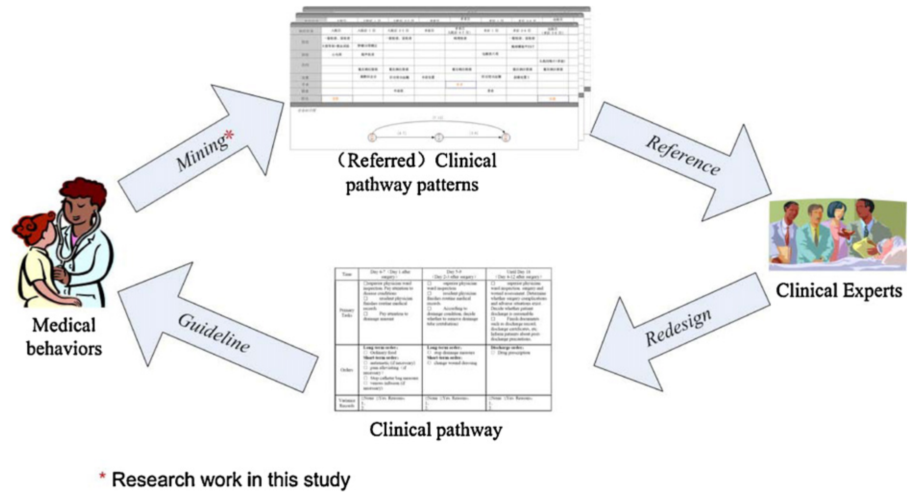

Fig. 1. Process mining based clinical pathway analysis and optimization.

clinical pathway, or if patients normally undergo their radical resection of colon cancer surgeries in 3–7 days after admission, and are typically discharged 7–10 days after surgery. We note that temporal relationships among medical behaviors are also called *chronicles* [19–21], which not only allow the researcher to set a relative order of medical behavior occurrences in clinical pathways, but also allow one to quantify the time gaps between behaviors. In the process of clinical pathway analysis, this quantification is very useful as it allows the researcher to make differences between different situations in clinical pathways that have same medical behaviors, but different time spreadings. Indeed, different time spreadings of the same set of medical behaviors may indicate, for example, that the same patient therapy behaviors are realized in different contexts. The essential medical behaviors and chronicle information form the backbone of clinical pathway patterns and should be conserved. This task, called *clinical pathway pattern mining*, is one of the most important aspects of clinical pathway analysis.

Many techniques have been proposed for clinical pathway pattern mining, to extract knowledge and information in clinical pathways, and help analysts to redesign/optimize clinical pathways. Most of these techniques are based on the experiences and knowledge of clinical experts, or are oriented to clinical data statistical analysis such as the statistics of pathway coincidence rate and abort rate, etc [6,22]. In such techniques, the analysts interpret large amounts of collected medical behaviors, and elaborate clinical pathway patterns, piece after piece, which can be a very tedious process. In addition, it appears that analysis of the results are somehow influenced by perceptions, e.g., medical behaviors in clinical pathways are often normative in the sense that they state what should be done rather than describing the actual medical behaviors in clinical pathways. As a result, it tends to be a rather subjective process.

Another possible approach uses data mining and machine learning technologies to measure medical behavior from clinical workflow logs. This is also called process mining [23–26]. Process mining, as a valuable set of techniques, has been widely studied in the business process management domain. It uses workflow logs to record business process execution information, to mine the actual behaviors in business processes, and discover business process patterns. Based on process execution data, with its logic and reasoning ability, process mining guarantees integrity, objectivity and universality of the discovered process patterns [26,27].

Process mining can be an objective way of analyzing clinical pathways as it is not biased by perceptions or normative behaviors. Note that the medical behaviors in patient-care journeys can be recorded into clinical workflow logs through various kinds of hospital information systems. This can be used to verify and analyze medical services. In addition, it effectively reflects the real executing conditions in clinical pathways. Consequently, as Fig. 1 shows, process mining can be applied to analyze all kinds of medical behaviors, mine frequent clinical pathway patterns, which can often indicate critical medical behaviors in patient-care journeys, and can also provide a valuable reference for clinical experts to help them redesign and continuously optimize clinical pathways.

Taking into account these reasons and considering the fact that clinical workflow logs are often recorded by hospital information systems and are easy to collect, we adopt process mining to analyze clinical pathways. In particular, the *key objective* of our paper is to mine comprehensive clinical pathway patterns given a specific clinical workflow log and a minimum support threshold, i.e., the complete discovery consists in discovering all clinical pathway patterns with respect to the clinical workflow log such that the support degree of each pattern is larger than a minimum support threshold. The support degree of a clinical pathway pattern with respect to a clinical workflow log is the number of clinical pathway traces in the log which contain medical behaviors of the pattern, and satisfy the temporal constraints (i.e., chronicles) of the pattern.

However, the diversity of medical behaviors and the complexity of chronicle information among medical behaviors in clinical pathways is far higher than that of common business processes. Traditional process mining techniques have many problems and challenges when used for mining clinical pathway patterns [27,28]. Although many process mining techniques can tell us which medical behaviors are frequently performed and in which order, they seldom provide precise chronicled information about the critical medical behaviors in clinical pathways for further decision support. In addition, applying traditional process mining techniques may generate spaghetti-like pathway patterns that are difficult to comprehend for clinical experts [27,29]. Such incomprehensible patterns are either not amenable or are lacking in assisting one in the clinical pathway redesign and optimization efforts.

Therefore, it is necessary to develop a new process mining technique to effectively mine clinical pathway patterns. To this end, our *main contribution*, in this paper, is to develop a novel process mining approach to mine clinical pathway patterns from medical behaviors. In comparison to the traditional process mining techniques, the proposed approach can discover closed clinical pathway patterns. It can also answer questions related to the most common (likely) behavior, the pathway traces that share/capture a desired behavior, the time spans between desired behaviors in clinical pathways, etc. Our approach is evaluated via real-world data sets from Zhejiang Huzhou Central Hospital of China.

This paper is organized as follows. Section 2 introduces the concept of the clinical pathway and process mining and provides a brief overview of process mining in health-care. Section 3 formulates a clinical pathway pattern mining problem. In Section 4, we introduce our approach for mining clinical pathway patterns from medical behaviors, which are regularly recorded in clinical workflow logs. Evaluation of the proposed approach is shown in Section 5. Section 6 summarizes our conclusions and considers future research directions.

2. Background

In this section, we provide some background on clinical pathways, and then review related work on process mining and its various applications in health-care.

2.1. Understanding clinical pathway

Clinical pathways aim to coordinate the patient-care process by a team of health-care professionals for a specific diagnosis or procedure [18]. In essence, it describes the functional knowledge pertaining to an institution clinical practices in terms of time-sensitive and outcome-driven processes, represented as a combination of plans, tasks, decisions, resources and care providers that essentially resembles a workflow [1,30].

Hunter and Segrott point out that a clinical pathway is a specific trajectory of the sequencing and timing of practitioners' care [6]. Bryan et al. [31] describe a clinical pathway as "a map of the process involved in managing a common clinical condition or situation". In fact, such a trajectory or map defines a set of medical behaviors, and the time that these behaviors take to occur. As shown in Table 1,[1] a clinical pathway consists of different categories of medical behaviors, namely multidisciplinary clinical activities, and their dependencies with the following characteristics [6,32]: (1) clinical activities are spread along a predefined time-line, which is the expected length of stay (LOS) in the hospital, and each activity represents a specific clinical task (e.g., a medical order, a radiological examination test, etc.). For instance, a clinical activity such as the surgery of lung cancer, is preferably performed in the time span between the fourth day and the seventh day of LOS. In addition, there are specific starting/ending activity types in clinical pathways. For example, the starting/ending activity types of clinical pathways published by Ministry of Health of China are *admission* and *discharge*. Moreover, certain temporal relations exist between clinical activities. For example, a color ultrasound examination should be performed on the first day after *admission*, and another color ultrasound examination should be performed on the day before *discharge*. (2) A clinical pathway normally enumerates regular medical behaviors that are expected to occur in patient-care journeys and serves as checkpoints for the performance of the pathway. We note that medical behaviors in this study are represented as flexible, transparent, and re-usable pieces of functionality that consist of one or several clinical activities required to set up a clinical solution. (3) These medical behaviors, as basic alternatives, can be applied in considering specific patient states in clinical pathways. A clinical pathway may generate a set of regular medical behaviors with occasional variants. Note that respective model adaptations result in large collections of process model variants that are derived from the same process model, but differ slightly in structure [33]. In particular, variants are common in clinical settings because the changes in the patient state make the applied medical behaviors inappropriate, such that patient-care has to be adjusted. As a matter of fact, a clinical pathway frequently entails improvisation or ad hoc combinations. Often unexpected delays and iterations might occur if new medical behaviors are not substituted or short-cuts are not taken. Over time, these improvisations may become accepted as formal procedures, but in the short run, they would appear as variants on the standardized pathway.

2.2. Process mining and its applications in health-care

The goal of process mining is to extract information (e.g., process or organizational models) from workflow logs, i.e., process mining describes a family of a-posteriori analysis techniques exploiting the information recorded in workflow logs [26,34]. Typically, these approaches assume that it is possible to record events sequentially such that each event refers to an activity (i.e., a well-defined step in the process) and is related to a particular case (i.e., a process case).

Process mining addresses the problem that most "process/system owners" have limited information about what is actually happening [23,24,35]. In practice, there is often a significant gap between what is prescribed or supposed to happen, and what happens in reality [36]. Only a concise assessment of reality, which process mining strives to deliver, can help to verify process patterns, and ultimately be used in process redesign efforts [35,36]. Discovering frequently occurring temporal patterns in process cases facilitates intelligent and automatic extraction of useful knowledge to support business decision-making. Similarly, data mining techniques are exploited in workflow management contexts to mine frequent workflow execution patterns [37]. The sequence of activities within a process, the execution cost and the reliability of the process can be predicted by using the process path mining technique [38]. Based on the process patterns and process paths, unexpected but useful knowledge about the process is extracted to help the user make appropriate decisions. For more details on process mining, please refer to [26,36].

The application of process mining in health-care is a relatively unexplored field, although it has already been attempted by some authors [27], who have devised a methodology based on process mining in order to support business process analysis in health-care. Their methodology includes process mining techniques that are especially useful in health-care environments, given the characteristics of health-care processes. 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.

To the best of our knowledge, the approaches that are most similar to the ones presented in this paper, are highlighted in [39,40]. In [39], Klundert et al., presents a model to measure clinical pathway adherence, which can cope with variations in pathways and deviations from pathways. They evaluated their method by using real-life data from the years 2001-2005 at the Maastricht University Medical Centre (MUMC). Lin et al. [40] reported a data mining technique that was developed to discover the time dependency pattern of clinical pathways for managing brain stroke. The mining of time dependency patterns allows us to discover patterns of process execution sequences and to identify the dependent relation between activities in a majority of cases. By obtaining the time dependency patterns, it is possible to predict the paths for new patients who are admitted into the hospital.

Note that the work mentioned above is only confined to one or several well-structured fragments of patient-linked treatment processes, such as radiological workflow. To the best of our knowledge,

[1] This is a translation of a bronchial lung cancer clinical pathway published by Ministry of Health of China. For the original version, please refer to: http://www.moh.gov.cn/publicfiles/business/cmsresources/mohyzs/cmsrsdocument/doc4905.doc.

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

Suitable for: Patient with **First Diagnosis** of bronchial lung cancer (ICD-10:C34;D02.2) for local excision of pulmonary/lobectomy/pneumonectomy+systematic lymph node dissection/Thoracotomy surgery (ICD-9-CM-3:32.29/32.3-32.5)

Patient Name: ____________ Gender: ____________ Age: ____________

Outpatient Service NO: ____________ Hospitalization NO: ____________ Admission Date: ____________ Discharge Date: ____________ Length Of Stay: 14 -21 Day

Time	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)	Discharge (Days 13-21)
Diagnosis and Treatment	☐Medical history inquiry and physical examination ☐Write patient record ☐Issue laboratory orders and check request form ☐Attending rounds ☐Set treatment plan	☐Higher authority physician rounds ☐Preoperative preparation ☐Preoperative evaluation ☐Preoperative discussion and surgical planning ☐Preoperative consultation ☐Resident completes medical records including progress and preoperative log summary, superior physician records ☐Sign the informed consent procedure, expense agreement, blood transfusion consent, consent authorization	☐Indwelling catheter before surgery ☐Surgery ☐Surgeon completes operation record ☐Resident completes postoperative course ☐Higher authority physician rounds ☐Observation of vital signs ☐Account of illness and postoperative precautions to patient and family	☐Higher authority physician rounds ☐Resident completes progress note ☐Observe chest drainage ☐Note vital signs and breath sounds in the lungs ☐Encourage and assist patients with expectoration ☐Bronchoscopy sputum	☐Higher authority physician rounds ☐Resident completes progress note ☐Review blood routine examination,biochemistry and chest X-ray ☐Remove chest drain and dress incision according to patient condition ☐Bronchoscopy sputum ☐Discontinue/adjust antibiotic	☐Remove incision suture ☐Higher authority physician rounds and determine discharge ☐Resident completes discharge summary,medical record homepage, etc. ☐Inform patient and family of issues after discharge ☐Determine treatment planning according to postoperative pathology
Medical Order	**Long term order:** ☐Thoracic surgery Secondary care ☐Normal diet **Temporary Order:** ☐Blood, urine, stool routine examination ☐Coagulation, blood type, liver and kidney function examination, electrolytes, infectious disease screening,tumor markers check ☐Lung function, arterial blood gas analysis, ECG, echocardiography ☐Sputum cytology, bronchoscopy+biopsy ☐Imaging: lateral chest X-ray, chest CT, abdominal ultrasound or CT, whole body bone scan, brain MRI or CT ☐When necessary: PET-CT or SPECT, mediastinoscopy, 24-hour ambulatory ECG,percutaneous lung biopsy, etc.	**Long term order:** ☐Atomizing inhalation **Temporary Order:** ☐Schedule Local excision of pulmonary/Lobectomy/ Pneumonectomy/Thoracotomy surgery under general anesthesia ☐No food or water intake 6 hours before surgery ☐Enema the night before surgery ☐Preoperative skin preparation ☐Preparation of blood transfusion ☐Sedative drugs ☐Preparation of antibacterial drugs in surgery ☐Other special advices	**Long term order:** ☐General thoracic surgery postoperative care ☐Premium or first level nursing ☐Liquid food intake 6 hour after clear-headed ☐Oxygen inhalation ☐Body temperature, ECG, blood pressure,respiration, pulse, blood oxygen saturation monitoring ☐Record the amount of chest drainage ☐Continued catheterization, record 24-hour intake and output ☐Atomizing inhalation ☐Prophylactic antibiotics ☐Analgesic **Temporary Order:** ☐Other special advices	**Long term order:** ☐Thoracic surgery level I care ☐Normal diet **Temporary Order:** ☐Blood routine, liver and kidney function, electrolytes examination ☐Chest X-ray ☐Other special advices	**Long term order:** ☐Thoracic surgery level II care ☐Stop measurement of closed chest drainage ☐Stop urine record,oxygen, ECG monitoring ☐Stop atomization inhalation ☐Stop antimicrobial **Temporary Order:** ☐Remove closed chest drainage tube ☐Remove catheter ☐Dress the incision ☐Review blood routine, liver and kidney function, electrolytes examination according to patient condition ☐Other special advices	**Temporary Order:** ☐Remove suture ☐Dress the incision ☐Inform of discharge ☐Discharge with medication ☐Regular return visit
Nursing Care	☐Introduce the ward environment, facilities and equipment ☐Admission nursing assessment ☐Aid smoking cessation	☐Education, preoperative ☐skin preparation ☐Inform of no water and food intake ☐Respiratory exercises	☐Observe changes in condition ☐Postoperative care of psychological and life ☐Maintain patency of airway	☐Observe patient condition ☐Care of psychological and life ☐Aid patient expectoration	☐Observe patient condition ☐Care of psychological and life ☐Aid patient expectoration	☐Observe patient condition ☐Care of psychological and life ☐Recovery instruction
Variance Record	☐No ☐Yes, caused by: 1. 2.	☐No ☐Yes, caused by: 1. 2.	☐No ☐Yes, caused by: 1. 2.	☐No ☐Yes, caused by: 1. 2.	☐No ☐Yes, caused by: 1. 2.	☐No ☐Yes, caused by: 1. 2.
Nurse Signature						
Physician Signature						

previous work is not yet involved in mining the core behaviors of health-care processes such as clinical pathways. Because of complex medical behaviors generated during clinical pathway execution, traditional process mining techniques have many problems and challenges when applied to clinical practice. They often generate spaghetti-like clinical pathway patterns that are incomprehensive to health-care professionals.

Our approach is different from the traditional process mining techniques, which typically documents the start/end of each activity execution and therefore, reflects the behavior of the implemented processes. Our approach is specific to mining clinical pathway patterns. Thus, given clinical workflow logs, this approach can discover understandable clinical pathway patterns that provide not only the sequential order of activities, but also information about the time span between different pairs of activities precisely. These discovered patterns allow medical staff to realize/study which medical behaviors can be performed and the time periods during which these behaviors can be performed in clinical pathways.

3. Problem definition

The goal of mining clinical pathway patterns is to extract knowledge about target clinical pathways from medical behaviors. In order to analyze any clinical activity, and thus to discover interesting behavioral knowledge about this activity, it is necessary to collect observational data about the activity. In this study, we assume that it is possible to record medical behaviors represented as clinical events in patient-care journeys in clinical workflow logs. We also assume that the occurrence times of these clinical events are also recorded in clinical workflow logs. In fact, many electronic medical record (EMR) systems record such information. In order to explain the kind of input needed for our approach, we first define the following concepts.

Definition 1 *(Clinical event).* Let A be a set of clinical activities, and T the time domain. A clinical event e is represented as $e = (a, t)$, where a is the activity type of e ($a \in A$), and t is the occurring time of event e ($t \in T$). A clinical event is a clinical activity occurring at a particular time stamp.

For example, we let $(a, 1)$ be a particular clinical event, where a is the activity type, i.e., *admission*, of the event, and 1 is occurring time of the event.

In this study, we assume that clinical events are point-based events, which is the common assumption adopted by most pattern mining studies [41–43]. A point-based event is viewed as something that occurs at a certain point in time. In clinical pathways, however, events cannot always be represented as points. For instance, in patient-care journeys, medical behaviors may be represented as interval-based events, if we record when the medical behaviors are performed and how long the behaviors last. When clinical events are represented as intervals, an event can be described with three major characteristics: activity name, event starting time, and event ending time. However, an interval-based event can be represented by two point-based events. For example, as shown in Fig. 2,[2] an interval-based event "Post operation drain" can be represented as two point-based events, i.e., "Post operation drain begin", and "Post operation drain end". Thus, in this study, we simply represent each interval-based event as two corresponding point-based events, and develop an approach to discover clinical pathway patterns from point-based events.

[2] These traces are patient-care cases from Zhejiang Huzhou Central Hospital of China. We have simplified these cases by keeping several critical clinical events in each clinical pathway trace.

Definition 2 *(Clinical pathway trace).* A clinical pathway trace is represented by $\sigma = \langle tid, \langle e_1, e_2, \ldots, e_n\rangle\rangle$, where tid is the identifier of this trace and $\langle e_1, e_2, \ldots, e_n\rangle$ is a finite non-empty sequence of clinical events such that each event appears only once, and time is non-decreasing, i.e., for $1 \le i \le j \le n : e_i \neq e_j$ and $e_i . t \le e_j . t$.

For example, as shown in Fig. 2, there are four clinical pathway traces. Each trace consists of a set of clinical events. These traces are represented as $\sigma_1, \sigma_2, \sigma_3$ and σ_4, respectively, in Table 2. When checking if a clinical pathway trace appears in a sequence, we usually have to determine the relation between the two events.

Definition 3 *(Arrangement of events).* In a clinical pathway trace, event e_i must be placed before event e_j based on the following conditions:

1. $e_i . t < e_j . t$,
2. If $e_i . t = e_j . t$, but e_i's activity type $e_i . a$ alphabetically precedes that of e_j.

Definition 4 *(Clinical workflow log).* Let *Trace* be the set of all possible clinical pathway traces, and a clinical workflow log $\mathcal{L}$ is a set of traces $\mathcal{L} \subseteq Trace$ such that each event appears at most once in the entire log, i.e., for any $\sigma_1, \sigma_2 \in \mathcal{L} : \forall e_1 \in \sigma_1 \forall e_2 \in \sigma_2, e_1 \neq e_2$ or $\sigma_1 = \sigma_2$.

Fig. 2 shows an example of a clinical workflow log of bronchial lung cancer clinical pathway, which consists of four clinical pathway traces, i.e., $\mathcal{L} = \{\sigma_1, \sigma_2, \sigma_3, \sigma_4\}$. Table 2 shows the details of clinical event sequence of these traces.

Definition 5 *(Temporal constraint).* Let A be a set of clinical activities and let T be the time domain. A clinical pathway temporal constraint is a 4-tuple $\theta = (a_1, a_2, t^-, t^+)$, denoted $(a_1, [t^-, t^+], a_2)$, where $a_1, a_2 \in A$ are clinical activity types, and $t^-, t^+ \in T$ are the lower bound and the upper bound of the temporal constraint, such that $t^- \le t^+$. Two events e_1 and e_2 of a particular clinical pathway trace σ are said to satisfy the temporal constraint $(a_1, [t^-, t^+], a_2)$, if $e_1 . a = a_1, e_2 . a = a_2, e_2 . t - e_1 . t \in [t^-, t^+]$.

For example, let $(a, [3, 4], g)$ be a temporal constraint. It enforces that g appears between 3 and 4 time units after a. The clinical pathway traces σ_1, σ_3, and σ_4 as shown in Table 2, satisfy the temporal constraint $(a, [3, 4], g)$. For convenience, let $\theta . t^-, \theta . t^+$ be the lower bound and upper bound of the temporal constraint θ, respectively.

In addition, we define the frequency of a temporal constraint θ in a clinical workflow log $\mathcal{L}$ as its number of occurrence in $\mathcal{L}$ with respect to the total count of traces in $\mathcal{L}$. A simple way of collecting the occurrences of θ in $\mathcal{L}$ is to decide that each combination of events of each trace of $\mathcal{L}$ that satisfies the temporal constraint of θ is considered as an occurrence of θ. For example, there are four occurrences of the temporal constraint $\theta = (a, [3, 9], g)$ in clinical workflog log shown in Table 2. Thus, the frequency of θ is 100%. While for the temporal constraint $\theta' = (a, [3, 4], g)$, its frequency in the clinical workflow log shown in Table 2 is 75%.

Definition 6 *(Clinical activity sequence).* Let A be a set of clinical activities. Let $\mathcal{A} = \langle a_1, a_2, \ldots, a_k\rangle$ be an ordered clinical activity sequence, where $a_i \in A$ for $1 \le i \le k$.

For example, we let $\mathcal{A} = \langle a, g, v\rangle$ be a particular clinical activity sequence, which consists of three activities, i.e., *Admission*, *Radical surgery*, and *Discharge*. These three activities are performed sequentially in patients' clinical pathway traces.

Definition 7 *(Chronicle).* Let $\mathcal{A}$ be an ordered clinical activity sequence, and T the time domain. A chronicle is a set of temporal constraints $\mathcal{C}_{\mathcal{A}} = \{\theta_{a_i a_j}\}$ on $\mathcal{A}$.

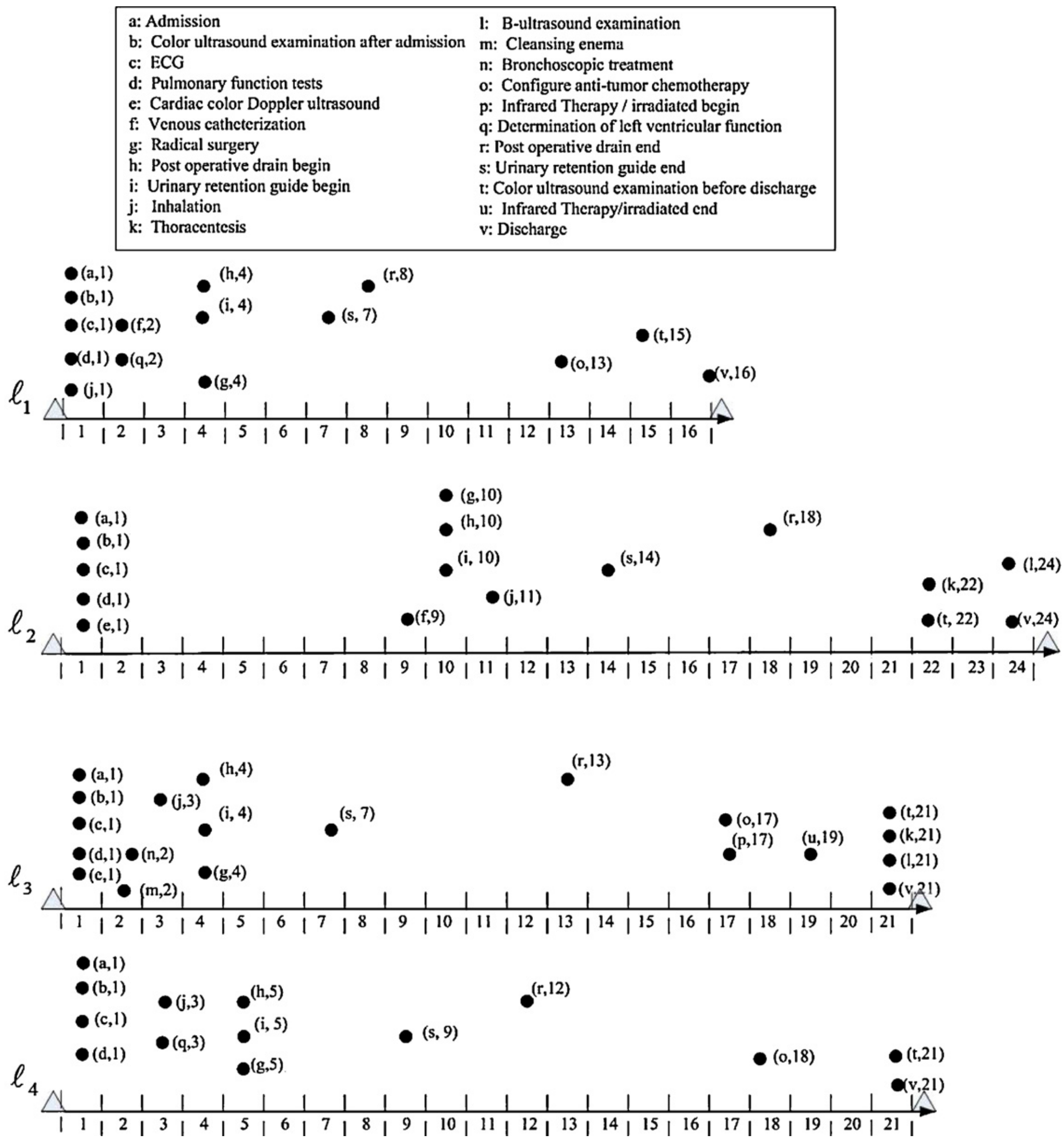

Fig. 2. A clinical workflow log example of bronchial lung cancer clinical pathway.

Table 2
A clinical workflow log example of bronchial lung cancer clinical pathway.

id	Sequence
σ_1	$\langle(a, 1), (b, 1), (c, 1), (d, 1), (j, 1), (f, 2), (q, 2), (g, 4), (h, 4), (i, 4), (s, 7), (r, 8), (o, 13), (t, 15), (v, 16)\rangle$
σ_2	$\langle(a, 1), (b, 1), (c, 1), (d, 1), (e, 1), (f, 9), (g, 10), (h, 10), (i, 10), (j, 11), (s, 14), (r, 18), (k, 22), (t, 22), (l, 24), (v, 24)\rangle$
σ_3	$\langle(a, 1), (b, 1), (c, 1), (d, 1), (e, 1), (m, 2), (n, 2), (j, 3), (g, 4), (h, 4), (i, 4), (s, 7), (r, 13), (o, 17), (p, 17), (u, 19), (k, 21), (l, 21), (t, 21), (v, 21)\rangle$
σ_4	$\langle(a, 1), (b, 1), (c, 1), (d, 1), (j, 3), (q, 3), (g, 5), (h, 5), (i, 5), (s, 9), (r, 12), (o, 18), (t, 21), (v, 21)\rangle$

For example, we let $\mathcal{A} = \langle a, g, v\rangle$ be a particular clinical activity sequence, and $\{(a, [3, 9], g), (a, [15, 23], v), (g, [12, 17], v)\}$ be a particular chronicle on $\mathcal{A}$.

Note that a chronicle is a set of temporal constraints on a particular ordered activity sequence. It apparently suggests some sequential behavior between these activities. In particular, it satisfies the following property:

Property 1. *Let $\mathcal{A} = \langle a_1, a_2, \ldots, a_k\rangle$ be an ordered clinical activity sequence, and $\mathcal{C}_\mathcal{A}$ be a chronicle on $\mathcal{A}$. The temporal constraints in a particular chronicle $\mathcal{C}_\mathcal{A}$ must satisfy $t^-_{a_{i_1}a_{i_2}} + t^-_{a_{i_2}a_{i_3}} + \cdots + t^-_{a_{i_{m-1}}a_{i_m}} \leq t^-_{a_{i_1}a_{i_m}}$ and $t^+_{a_{i_1}a_{i_2}} + t^+_{a_{i_2}a_{i_3}} + \cdots + t^+_{a_{i_{m-1}}a_{i_m}} \geq t^+_{a_{i_1}a_{i_m}}$ for $1 \leq i_1 < i_2 < \cdots < i_m \leq k$.*

The property above guarantees that each temporal constraint is consistent with the other temporal constraints in a particular chronicle. In the example above, the chronicle on the activity sequence $\langle a, g, v\rangle$ satisfies $t^-_{ag} + t^-_{gv} \leq t^-_{av}$, and $t^+_{ag} + t^+_{gv} \geq t^+_{av}$.

Definition 8 *(Clinical pathway pattern).* A clinical pathway pattern is a pair $\phi = (\mathcal{A}, C)$, such that:

1. $\mathcal{A} = \langle a_1, a_2 \ldots, a_k, \rangle$ is an ordered clinical activity sequence; and,
2. C is a chronicle on $\mathcal{A}$ such that for all pairs (a_i, a_j) of $\mathcal{A}$ satisfying $i < j$, there exists a temporal constraint $\theta_{a_ia_j} \in \mathcal{C}$, where $\theta_{a_ia_j}$ is denoted by $(a_i, [t^-_{a_ia_j}, t^+_{a_ia_j}], a_j)$.

$\mathcal{A}$ is called the sequence of ϕ, in the sense of frequent sequence pattern discovery [44]. In addition, we call

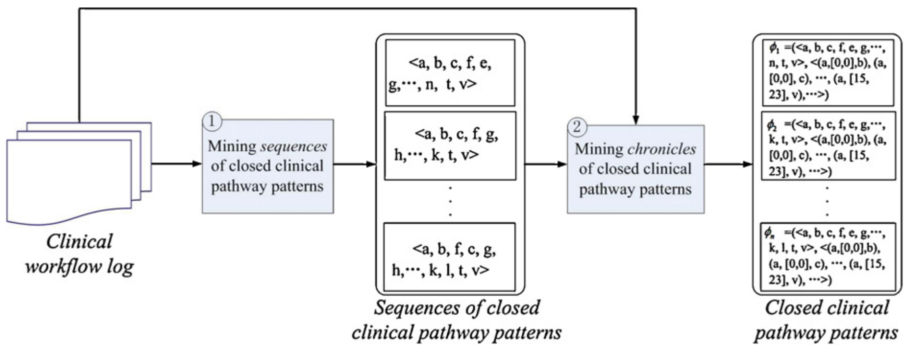

Fig. 3. The methodology of the proposed approach.

$\phi = (\mathcal{A}, \mathcal{C})$ a k-pattern, if $\mathcal{A} = \langle a_1, a_2, \ldots, a_k\rangle$. For example, $\langle a\rangle$ is a 1-pattern, and $(\langle a, g, v\rangle, \{(a, [3, 9], g), (a, [15, 23], v), (g, [12, 17], v)\})$ is a 3-pattern.

Definition 9 *(Pattern support).* Let $\sigma = \langle e_1, e_2, \ldots, e_n\rangle$ be a clinical pathway trace, and $\phi = (\langle a_1, a_2, \ldots, a_k\rangle, \{\theta_{a_i a_j} | 1 \le i < j \le k\})$ be a clinical pathway pattern. We say that ϕ is supported by σ, denoted as *supp*(ϕ, σ), if there exists a strictly increasing function f on the indexes of σ satisfying the following:

1. $a_1 = e_{f(1)}.a$, $a_2 = e_{f(2)}.a, \ldots, a_k = e_{f(k)}.a$; and,
2. $e_{f(j)}.t - e_{f(i)}.t$ is in the temporal constraint of $\theta_{a_i a_j}$, i.e., $t^-_{a_i a_j} \le e_{f(j)}.t - e_{f(i)}.t \le t^+_{a_i a_j}$ for $1 \le i \le j - 1$ and $2 \le j \le k$.

For example, we let $\phi = (\langle a, g, v\rangle, \{(a, [3, 9], g), (a, [15, 23], v), (g, [12, 17], v)\})$ be a clinical pathway pattern. Clearly, ϕ is supported by all four clinical pathway traces in Table 2.

Definition 10 *(Sub clinical pathway pattern).* Let $\phi' = (\langle b_1, b_2, \ldots, b_m\rangle, \{\theta_{b_l b_s} | 1 \le l < s \le m\})$ be a sub-pattern of another clinical pathway pattern $\phi = (\langle a_1, a_2, \ldots, a_k\rangle, \{\theta_{a_i a_j} | 1 \le i < j \le k\})$, if the following conditions are satisfied:

1. There exists a strictly increasing function f on the indexes of ϕ, such that $b_1 = a_{f(1)}$, $b_2 = a_{f(2)}, \ldots, b_m = a_{f(m)}$; and,
2. $\forall l, s$ such that $1 \le l < s \le m < k$, $[t^-_{f(l)f(s)}, t^+_{f(l)f(s)}] \subseteq [t^-_{ls}, t^+_{ls}]$, where $\theta_{b_l b_s}.t^- \le t^-_{ls} \le t^+_{ls} \le \theta_{b_l b_s}.t^+$ and $\theta_{a_{f(l)} a_{f(s)}}.t^- \le t^-_{f(l)f(s)} \le t^+_{f(l)f(s)} \le \theta_{a_{f(l)} a_{f(s)}}.t^+$.

For example, a clinical pathway pattern $\phi' = (\langle a, g\rangle, \langle(a, [3, 9], g)\rangle)$ is a sub-pattern of $\phi = (\langle a, g, v\rangle, \langle(a, [3, 4], g), (a, [15, 23], v), (g, [12, 17], v)\rangle)$. Note that we also call ϕ a super-pattern of ϕ', and ϕ contains ϕ'.

In order to efficiently mine clinical pathway patterns, it is necessary to discard non-typical behaviors according to the user's view-point (i.e., to avoid capturing temporal patterns that occur too infrequently for it to be worth attempting to learn lessons from such particular traces). Performing such a task requires providing any patterns with a support value, which provides the number of occurrences in the clinical workflow log. Note that parts of the log may be incorrect, incomplete, or refer to exceptions. Obviously, these exceptions, which are recorded only once, should not automatically become part of the regular clinical pathway patterns. However, it may be the case that a particular patient condition or clinical situation does require a deviation from the normal clinical pathways, and that infrequent behavior is duly justified. As a result, these variants should be inspected carefully [27].

Definition 11 *(Support).* Let $\mathcal{L}$ be a clinical workflow log, and ϕ be a clinical pathway temporal pattern. The support of ϕ in $\mathcal{L}$, denoted *supp*(ϕ, $\mathcal{L}$), is defined as:

$$supp(\phi, \mathcal{L}) = \frac{|\{\sigma | \sigma \in \mathcal{L} \wedge supp(\phi, \sigma)\}|}{|\mathcal{L}|} \tag{1}$$

For example, a clinical pathway temporal pattern $\phi = (\langle a, g, v\rangle, \{(a, [3, 4], g), (a, [15, 23], v), (g, [12, 17], v)\})$ is supported by clinical pathway traces σ_1, σ_3 and σ_4 of log $\mathcal{L}$, as shown in Table 2. Thus, the support of ϕ in $\mathcal{L}$ is $(|\{\sigma_1, \sigma_3, \sigma_4\}|)/|\mathcal{L}| = 0.75$.

Given a user-defined minimal support threshold, denoted as *minsupp*, the problem of clinical pathway pattern mining is the extraction of a clinical pathway pattern from a clinical workflow log that *supp*(ϕ, $\mathcal{L}$) ≥ *minsupp*. Such a clinical pathway pattern is defined as being 'frequent'.

Property 2. *If a clinical pathway pattern is frequent, so are all of its sub patterns. Accordingly, if a clinical pathway pattern is not frequent, then its super pattern will not be either.*

Definition 12 *(Closed clinical pathway pattern).* Let $\mathcal{L}$ be a clinical workflow log. A clinical pathway pattern $\phi = (\langle a_1, a_2, \ldots, a_k\rangle, \{\theta_{a_i a_j} | 1 \le i < j \le k\})$ is a closed pattern if it satisfies the following conditions:

1. *supp*(ϕ, $\mathcal{L}$) ≥ *minsupp*; and,
2. $\nexists \phi'$ such that ϕ is a sub-pattern of ϕ', and *supp*(ϕ, $\mathcal{L}$) = *supp*(ϕ', $\mathcal{L}$).

In this definition, the first criteria ensures that a closed clinical pathway pattern is frequent, and the second criteria ensures that there is no super-pattern with the same support for a closed clinical pathway pattern. The objective of this study is to find the set of closed clinical pathway patterns given a particular clinical workflow log, i.e., the complete discovery consists in discovering all closed clinical pathway patterns ϕ in clinical workflow log $\mathcal{L}$ such that *supp*(ϕ, $\mathcal{L}$) ≥ *minsupp*, where *minsupp* is a minimum support threshold.

4. Method

In this section, we present a novel approach of mining closed clinical pathway patterns from clinical workflow logs, that regularly record medical behaviors in patient-care journeys. Note that a closed clinical pathway pattern consists of a particular clinical activity sequence, and a particular chronicle on the activity sequence. Therefore, as shown in Fig. 3, given the input element, i.e., a clinical workflow log, the proposed approach (1) mines clinical activity sequences at first, and then (2) mines chronicles on the sequences to generate closed clinical pathway patterns.

4.1. Mining sequences of closed clinical pathway patterns

In this study, we propose a closed clinical pathway pattern's sequence mining algorithm, SCP-Miner, based on the ideas of classical sequence pattern mining algorithms. Before introducing the proposed SCP-Miner algorithm, the definitions of prefix, projection, and projected clinical workflow log are given as follows.

Definition 13 *(Prefix).* Let $\sigma = \langle e_1, e_2, \ldots, e_n\rangle$ be a clinical pathway trace, and $\mathcal{A} = \langle a_1, a_2, \ldots, a_k\rangle$ be a clinical activity sequence. We say that $\mathcal{A}$ is a prefix of σ if and only if $a_i = e_i . a$ for $1 \le i \le k \le n$.

For example, a clinical activity sequence $\mathcal{A} = \langle a, b, c\rangle$ is a prefix of the clinical pathway trace σ_1, as shown in Table 2.

Definition 14 *(Projection).* Let $\sigma = \langle e_1, e_2, \ldots, e_n\rangle$ be a clinical pathway trace, and $\mathcal{A} = \langle a_1, a_2, \ldots, a_k\rangle$ be a clinical activity sequence. A sub clinical pathway trace $\beta = \langle b_1, b_2, \ldots, b_m\rangle$ is a projection of σ with respect to $\mathcal{A}$ if and only if:

1. There exists a strictly increasing function f on the indexes of σ satisfying $e_{f(1)} . a = a_1, e_{f(2)} . a = a_2, \ldots, e_{f(k)} . a = a_k$ where $e_{f(1)}, e_{f(2)}, \ldots, e_{f(k)} \in \sigma$ and $a_1, a_2, \ldots, a_k \in \mathcal{A}$;
2. $\mathcal{A}$ is a prefix of β; and,
3. the last $m-k$ elements of β are the same as the last $m-k$ elements of σ.

For example, if the clinical trace σ_1, as shown in Table 2, is projected by a sequence $\mathcal{A} = \langle a, g\rangle$, a projection is obtained, i.e., $\langle (a, 1), (g, 4), (h, 4), (i, 4), (s, 7), (r, 8), (o, 13), (t, 15), (v, 16)\rangle$.

Definition 15 *(Projected clinical workflow log).* The projected clinical workflow log with respect to a clinical activity sequence $\mathcal{A}$ contains all the projections of $\mathcal{A}$ in the clinical workflow log $\mathcal{L}$.

When we generate a clinical activity sequence, we need to do some closure checking to determine whether or not the generated sequence is closed. Note that clinical pathways may have been specified through starting/ending activity types, which can be used in closure checking. For example, the starting/ending activity types of clinical pathways published by Ministry of Health of China are *admission* and *discharge*. Thus, the sequence can efficiently grow from a particular frequent 1-pattern, i.e., *admission*. This feature of clinical pathway patterns allows us to use a forward checking instead of bi-directional checking to determine if the generated sequence is closed.

Definition 16 *(Forward checking).* A clinical pathway pattern ϕ is not closed if the last activity of ϕ's sequence, $\mathcal{A}$, is not *discharge*.[3]

The proposed SCP-Miner algorithm, outlined in Algorithm 1, consists of two phases. First, we scan a clinical workflow log $\mathcal{L}$ from the pre-known frequent 1-activity sequence, i.e, $\langle admission\rangle$, and then build a projected clinical workflow log $\mathcal{L}|_{admission}$. Then, we recursively use a frequent k-activity sequence and its projected clinical workflow log to generate its frequent super-patterns at the next level in the frequent sequence tree, where $k \ge 1$. For each frequent k-activity sequence, we build its projected clinical workflow log and find all frequent 1-activity sequences in the projected clinical workflow log. During this phase, we use a forward checking to determine if the frequent sequences generated are closed. If $\mathcal{A}'$ is closed and the support of $\mathcal{A}'$ is not less than *minsupp*, $\mathcal{A}'$ is added into *SCP*.

[3] Note that for different clinical pathways, there may are different starting/ending activity types. Thus, the activity type used in forward checking may be different. As a matter of fact, it could be more general to allow the user to specify a set of starting/ending activity types according to different clinical pathways.

Algorithm 1 *(The clinical activity sequence mining algorithm).*

Procedure::SCP-Miner($\mathcal{L}$, *minsupp*)
Input:
  $\mathcal{L}$ is a clinical workflow log
  minsupp is a minimum support threshold value
Output:
  SCP is the set of sequences of closed clinical pathway patterns
Steps:
  Let $SCP = \emptyset$ be a set of sequences of closed clinical pathway patterns
  Let a be the clinical activity *admission* and $\mathcal{L}|_a$ be a projected clinical workflow log of a
  Call *SCPMiner*(a, $\mathcal{L}|_a$, *minsupp*, *SCP*)
  Output *SCP*
End Procedure
Procedure::*SCPMiner*(ϕ, $\mathcal{L}|_\phi$, *minsupp*, *SCP*)
Input:
  ϕ is a temporal pattern
  $\mathcal{L}|_\phi$: a projected workflow log of ϕ
  minsupp is a minimum support threshold value
  SCP is the set of sequences of closed clinical pathway patterns
Output:
  SCP is the set of sequences of closed clinical pathway patterns
Steps:
  Scan $\mathcal{L}|_\mathcal{A}$ and find all frequent clinical activities X_{k+1}
  If ($\mathcal{A}$ passes the forward checking) then
    If ($\mathcal{A}$ is closed with respect to *SCP*) then
      $SCP = SCP \cup \{\mathcal{A}\}$
    End If
  End If
  For each a_{k+1} in X_{k+1}
    Append a_{k+1} to $\mathcal{A}$ as $\mathcal{A}'$
    Let $\mathcal{L}|_{\mathcal{A}'}$ be the projected clinical workflow log of $\mathcal{A}'$
    Checking if $\mathcal{A}'$ is contained by any sibling pattern and both share the same projections
    If not, call *SCPMiner*($\mathcal{A}'$, $\mathcal{L}|_{\mathcal{A}'}$, *minsupp*, *SCP*)
  End For
  return *SCP*
End Procedure

Let us take the clinical workflow log, as shown in Table 2, as an example. We assume that $minsupp = 0.5$. First, we scan the projected clinical workflow log $\mathcal{L}|_{\langle a\rangle}$ from the activity a, i.e., *admission*. Next, we grow the frequent 1-activity sequence $\langle a\rangle$ to find its frequent super-patterns in the projected clinical workflow log. For example, if $\langle b\rangle$ is a frequent 1-activity pattern in $\langle a\rangle$'s projected clinical workflow log, then we can grow the frequent 1-activity sequence $\langle a\rangle$ by appending b to it, and thus obtain a frequent 2-activity sequence $\langle a, b\rangle$. Obviously, $\langle a, b\rangle$ is not closed. Therefore, we continue to grow the sequence by appending a frequent 1-activity sequence in its projected clinical workflow log. The work is recursively performed until we get final sequences as follows, $\mathcal{A}_1 = \langle a, b, c, d, e, g, h, i, j, s, r, k, l, t, v\rangle$, and $\mathcal{A}_2 = \langle a, b, c, d, j, q, g, h, i, s, r, o, t, v\rangle$.

4.2. Mining chronicles on generated clinical activity sequences

Based on the set of clinical activity sequences generated by the algorithm SCP-Miner, we can mine chronicles on each sequence to generate closed clinical pathway patterns. Before we present our method of mining chronicles on each clinical activity sequence, we introduce the following concepts.

Definition 17 *(Stricter chronicle).* A chronicle $\mathcal{C}_\mathcal{A}$ is stricter than another chronicle $\mathcal{C}_\mathcal{A}'$, denoted $\mathcal{C}_\mathcal{A} \prec \mathcal{C}_\mathcal{A}'$, if $\forall a_i, a_j \in \mathcal{A}$, $[t^-_{a_ia_j}, t^+_{a_ia_j}] \subset [t^{-'}_{a_ia_j}, t^{+'}_{a_ia_j}]$.

For example, we let $\mathcal{C}_{\{a,g,v\}} = \{(a, [3, 4], g), (a, [15, 23], v), (g, [12, 17], v)\}$ and $\mathcal{C}'_{\{a,g,v\}} = \{(a, [3, 9], g), (a, [15, 23], v), (g, [12, 17], v)\}$ be two chronicles, $\mathcal{C}_{\{a,g,v\}} \prec \mathcal{C}'_{\{a,g,v\}}$ since $[3, 4] \subset [3, 9]$. The relation $\prec$ is a partial relation of order over any chronicles.

Definition 18 *(Chronicle relation).* Given a particular clinical activity sequence $\mathcal{A}$, we say a chronicle $\mathcal{C}_\mathcal{A}$ "is child of" another chronicle

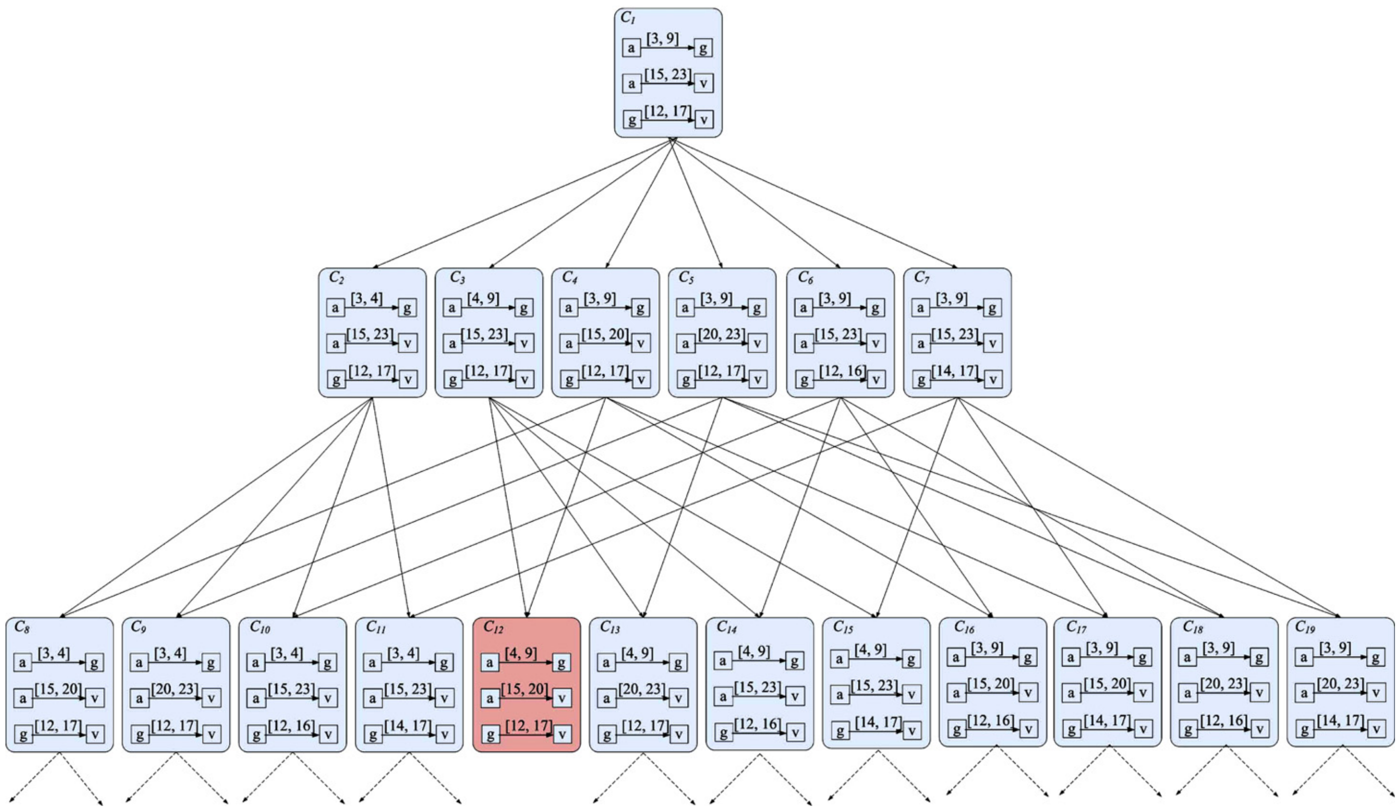

Fig. 4. A set of derived chronicles given the particular clinical activity sequence, $\mathcal{A} = \{a, g, v\}$, and the particular clinical workflow log shown in Table 2.

$\mathcal{C}'_\mathcal{A}$ if and only if $\mathcal{C}_\mathcal{A} \prec \mathcal{C}'_\mathcal{A}$ and there is no other chronicle $\mathcal{C}''_\mathcal{A}$ satisfying the condition such that $\mathcal{C}_\mathcal{A} \prec \mathcal{C}''_\mathcal{A} \prec \mathcal{C}'_\mathcal{A}$.

Note that given a particular clinical workflog log $\mathcal{L}$ and a clinical activity sequence $\mathcal{A}$, we can derive a set of chronicles form $\mathcal{L}$. For example, as shown in Fig. 4, a set of chronicles are generated on a particular clinical activity sequence $\mathcal{A} = \{a, g, v\}$ given the clinical workflow log shown in Table 2, named $\mathcal{G}|_{\{a,g,v\}}$. There is an arrow from the chronicle $\mathcal{C}_1 = \{(a, [3, 9], g), (a, [15, 23], v), (g, [12, 17], v)\}$ to the chronicle $\mathcal{C}_2 = \{(a, [3, 4], g), (a, [15, 23], v), (g, [12, 17], v)\}$ because $\mathcal{C}_1$ is the parent of $\mathcal{C}_2$, and an arrow from the chronicle $\mathcal{C}_2$ to the chronicle $\mathcal{C}_8 = \{(a, [3, 4], g), (a, [15, 20], v), (g, [12, 17], v)\}$ because $\mathcal{C}_2$ is the parent of $\mathcal{C}_8$. However, there is no arrow from the chronicle $\mathcal{C}_1$ to the chronicle $\mathcal{C}_8$ because $\mathcal{C}_1$ is not the parent of $\mathcal{C}_8$. Note that these chronicles are organized in an acyclic directed graph, where nodes are chronicles and arrows represent "is child of" relations (respectively, "is parent of" relation).

Property 3. *Let $\mathcal{A}$ be a particular clinical activity sequence. Given a particular clinical workflow log $\mathcal{L}$, there is the one and only one top chronicle derived from $\mathcal{L}$, denoted as $\mathcal{C}^{TOP}_\mathcal{A}$, satisfying that there is no other derived chronicle $\mathcal{C}'_\mathcal{A}$ such that $\mathcal{C}^{TOP}_\mathcal{A} \prec \mathcal{C}'_\mathcal{A}$.*

For example, we deduce from Fig. 4 that $\mathcal{C}_1$ is the top chronicle with respect to the given clinical workflow log shown in Table 2. In order to derive the top chronicles from particular clinical workflow logs, we propose an algorithm, i.e., TC-Miner, as shown in Algorithm 2.

In the algorithm TC-Miner, for each activity pair (a_i, a_j) of clinical activity sequence $\mathcal{A}$, the occurrences and the set of occurrence distances are calculated (Line 10 and Line 11) based on the input clinical workflow log $\mathcal{L}$. Following this, the maximum occurrence distance of (a_i, a_j) is picked up to generate a particular temporal constraint θ on (a_i, a_j) (Line 12). All other temporal constraints on (a_i, a_j) are stricter than θ. Note that the frequency of θ is 100% with respect to $\mathcal{L}$. At last, all possible temporal constraints are grouped together to generate to a particular top chronicle.

Property 4. *Let $\mathcal{A}$ be a particular clinical activity sequence. Given a particular clinical workflow log $\mathcal{L}$, the frequency of the derived top chronicle $\mathcal{C}^{TOP}_\mathcal{A}$ is 100%.*

Note that since the frequency of each temporal constraint that is contained in the top chronicle is 100%, the frequency of the top chronicle is also 100%. For example, as shown in Fig. 4, the frequency of the top chronicle $\mathcal{C}_1$ derived from the clinical workflow log depicted in Table 2 is 100%.

Algorithm 2 *(The top chronicle generating algorithm).*

1: **Procedure**::TC-Miner($\mathcal{A}$, $\mathcal{L}$)
2: Input:
3: $\mathcal{A}$ is a particular clinical activities sequence
4: $\mathcal{L}$ is a clinical workflow log
5: Output:
6: $\mathcal{C}^{TOP}_\mathcal{A}$ is the top chronicle with the limitation of B
7: Steps:
8: Let $\mathcal{C}^{TOP}_\mathcal{A} = \emptyset$
9: For each activity pair (a_i, a_j) in $\mathcal{A}$, do
10: $\mathcal{O}(a_i, a_j) \leftarrow \{\langle(a_i, t_i)(a_j, t_j)\rangle | (a_i, t_i) \in \sigma \wedge (a_j, t_j) \in \sigma \wedge \sigma \in \mathcal{L}\}$
11: $\Omega(a_i, a_j) \leftarrow$ sort $(\{(t_j - t_i) | \langle(a_i, t_i)(a_j, t_j)\rangle \in \mathcal{O}(a_i, a_j)\})$
12: Let $\theta(a_i, a_j) = (a_i, [\Omega(a_i, a_j)[0], \Omega(a_i, a_j)[|\Omega(a_i, a_j)|]], a_j)$
13: $\mathcal{C}^{TOP}_\mathcal{A} \leftarrow \{\theta(a_i, a_j)\}$
14: End For
15: Return $\mathcal{C}^{TOP}_\mathcal{A}$
16: **End Procedure**

The top chronicle is the "seed" to mine stricter chronicles on particular clinical activity sequences with respect to a particular clinical workflow log. In this study, an algorithm, CCP-Miner, is presented to mine chronicles on clinical activity sequences so that the closed clinical pathway patterns can be generated. The algorithm CCP-Miner, outlined in Algorithm 3, implements chronicle discovery on particular clinical activity sequences generated by the algorithm SCP-Miner. As shown in Line 9 of the algorithm, Ψ stores a set of possible frequent chronicles for a particular sequence $\mathcal{A}$ given a particular clinical workflow log $\mathcal{L}$. Each element in Ψ would be combined with $\mathcal{A}$ to generate a closed clinical pathway pattern. χ is the set of candidates of chronicles. Initially, there is a top chronicle $\mathcal{C}^{TOP}$ in χ (Line 11). Then, CCP-Miner works as follows: it takes one candidate chronicle $\mathcal{C}$ from χ, calculates the support of $(\mathcal{A}, \mathcal{C})$, and adds its children to χ if $(\mathcal{A}, \mathcal{C})$ is frequent. The algorithm ends when χ is empty (Line 27).

In each iteration, a chronicle $\mathcal{C}$ in χ is chosen and removed from χ. Then, the support of $(\mathcal{A}, \mathcal{C})$ is calculated. If $(\mathcal{A}, \mathcal{C})$ $(\mathcal{C} \in \Psi)$ has a support greater than *minsupp*, $\mathcal{C}$ is added into Ψ and Ψ is updated, which removes from Ψ any chronicle $\mathcal{C}'$ such that $\mathcal{C} \prec \mathcal{C}'$. Then, the procedure *GetChildren* generates all children of $\mathcal{C}$.

Note that in the procedure *GetChildren*, each temporal constraint contained in the particular chronicle $\mathcal{C}$ gets stricter in order to generate a set of possible children of $\mathcal{C}$ (Lines 42–48). In detail, if a particular temporal constraint θ contained in $\mathcal{C}$ has stricter temporal constraints Ξ learned by the procedure *Strict*, each element θ' in Ξ will take the place of θ in $\mathcal{C}$ to generate a new child chronicle $\mathcal{C}'$ of $\mathcal{C}$ (Line 45). The generated child $\mathcal{C}'$ is added to χ if $\mathcal{C}'$ has never been added to χ before (Lines 22–26). By checking that $\mathcal{C}'$ has never been added into χ, we ensure that the iteration will never process the same chronicle twice. This ensures that CCP-Miner will always terminate.

Algorithm 3 *(Chronicles mining algorithm).*

1: **Procedure**::CCP-Miner($\mathcal{A}$, $\mathcal{L}$, *minsupp*)
2: Input:
3: $\mathcal{A}$ is a sequence of a particular closed clinical pathway pattern
4: $\mathcal{L}$ is a clinical workflow log
5: *minsupp* is a minimum support threshold value
6: Output:
7: Φ is a set of closed clinical pathway patterns
8: Steps:
9: Let $\Psi = \emptyset$
10: Let $\mathcal{C}^{top}$ = TC-Miner($\mathcal{A}$, $\mathcal{L}$) is the top chronicle with respect to $\mathcal{L}$
11: Let $\chi = \{\mathcal{C}^{top}\}$
12: Repeat
13: Let $\mathcal{C}$ be the first element of χ
14: $\chi \leftarrow \chi - \{\mathcal{C}\}$
15: Let $\phi = (\mathcal{A}, \mathcal{C})$
16: If *Supp*$(\phi, \mathcal{L}) \geq$ *minsupp* then
17: *Update*$(\Psi, \mathcal{C})$
18: Else
19: Go to Line 12
20: End If
21: Let *Children* = GetChildren($\mathcal{C}$, $\mathcal{L}$)
22: For each $\mathcal{C}' \in$ *Children* do
23: If $\mathcal{C}'$ has never been added into χ before
24: $\chi \leftarrow \chi \bigcup \{\mathcal{C}'\}$
25: End If
26: End For
27: Until $\chi = \emptyset$
28: For each $\mathcal{C}$ in Ψ
29: $\Phi \leftarrow \Phi \bigcup (\mathcal{A}, \mathcal{C})$
30: End For
31: return ϕ
32: **End Procedure**
33: **Procedure**::GetChildren($\mathcal{L}$, $\mathcal{C}$)
34: Input:
35: $\mathcal{L}$ is a clinical workflow log
36: $\mathcal{C}$ is a chronicle with respect to $\mathcal{L}$
37: Output:
38: *Children* is a set of chronicles whose parent is $\mathcal{C}$
39: Steps:
40: Let *Children* = $\emptyset$
41: Let Θ be the set of temporal constraints contained in $\mathcal{C}$
42: For each $\theta \in \Theta$
43: Let Ξ = Strict($\mathcal{L}$, θ, *minsupp*) be a set of stricter temporal constraints given $\mathcal{L}$ and θ
44: For each θ' in Ξ
45: Let $\mathcal{C}' = \mathcal{C} - \{\theta\} + \{\theta'\}$
46: *Children* $\leftarrow$ *Children* $\bigcup \{\mathcal{C}'\}$
47: End For
48: End For
49: return *Children*
50: **End Procedure**
51: **Procedure**::Strict($\mathcal{L}$, θ, *minsupp*)
52: Input:
53: $\mathcal{L}$ is a clinical workflow log
54: θ is a temporal constraint with respect to $\mathcal{L}$
55: *minsupp* is a minimum support threshold value
56: Output:
57: Ξ is a set of stricter temporal constraints given $\mathcal{L}$ and θ
58: Steps:
59: Let $\Xi = \emptyset$
60: Let (a_i, a_j) be the activity pair of θ
61: $\mathcal{O}(a_i, a_j) \leftarrow \{\langle(a_i, t_i)(a_j, t_j)\rangle | \theta.t^- \leq t_i \leq t_j \leq \theta.t^+ \wedge (a_i, t_i) \in \sigma \wedge (a_j, t_j) \in \sigma \wedge \sigma \in \mathcal{L}\}$
62: $\Omega(a_i, a_j) \leftarrow$ sort $(\{(t_j - t_i) | \langle(a_i, t_i)(a_j, t_j)\rangle \in \mathcal{O}(a_i, a_j)\})$
63: Let $k = |\Omega(a_i, a_j)| - 1$
64: If $\frac{k}{|\mathcal{L}|} \geq$ *minsupp*
65: $\Xi \leftarrow \{(a_i, [\Omega(a_i, a_j)[l], \Omega(a_i, a_j)[l+k-1]], a_j) | 0 \leq l \leq |\Omega(a_i, a_j)| - k + 1 \wedge \Omega(a_i, a_j)[l] \neq \Omega(a_i, a_j)[l-1] \wedge \Omega(a_i, a_j)[l+k-1] \neq \Omega(a_i, a_j)[l+k]\}$
66: End If
67: Return Ξ
68: **End Procedure**

The procedure *Strict* is used to tighten a particular temporal constraint θ to generate a set of stricter temporal constraints Ξ such that each one in Ξ is frequent with respect to $\mathcal{L}$. At first, we build a complete set of all occurrences of the temporal constraint θ with respect to the particular clinical workflow log $\mathcal{L}$, denoted $\mathcal{O}(a_i, a_j)$ (Line 61), where (a_i, a_j) is the activity pair of θ. Formally, $\mathcal{O}(a_i, a_j) = \{(a_i, t_i)(a_j, t_j) | (a_i, t_i) \in \sigma \wedge (a_j, t_j) \in \sigma \wedge t_i \leq t_j \wedge \sigma \in \mathcal{L}\}$. And then we build and sort the set of occurrence distances, denoted $\Omega(a_i, a_j)$ (Line 62). Formally, $\Omega(a_i, a_j) = \{t_j - t_i | (a_i, t_i)(a_j, t_j) \in \mathcal{O}(a_i, a_j)\}$. Taking clinical workflow log $\mathcal{L}$ in Table 2 as an example, $\mathcal{O}(a, g) = \{(a, 1)(g, 4), (a, 1)(g, 9), (a, 1)(g, 4), (a, 1)(g, 3)\}$, and $\Omega(a, g) = \{3, 9, 3, 4\}$, i.e., $\Omega(a, g) = \{3, 3, 4, 9\}$. Furthermore, for the activity pair (a_i, a_j), we build a set of candidate temporal constraints, by applying a minimum support threshold. In particular, we adopt Cram's approach [21] to slide a window of width $k = |\Omega(a_i, a_j)| - 1$ from the first occurrence of an element in $\Omega(a_i, a_j)$ to the last occurrence of an element in $\Omega(a_i, a_j)$ (Line 65) to generate a set of stricter temporal constraints w.r.t θ, provided that the frequency of the generated temporal constraints is greater than the minimum support threshold, i.e., $(|\Omega(a_i, a_j)| - 1)/|\mathcal{L}| \geq$ *minsupp* (Line 64). For example, for the pair (a, g) with *minsupp* = 0.5, the window width k is $|\Omega(a, g)| - 1 = 3$, after which we slide a window with 3 over $\Omega(a, g) = \{3, 3, 4, 9\} : (a, [3, 4], g)$. As a result, a stricter temporal constraint $\theta' = (a, [3, 4], g)$ is generated. Note that the window is from the first occurrence of an element in $\Omega(a_i, a_j)$ to the last occurrence of an element in $\Omega(a_i, a_j)$, i.e., $\Omega(a_i, a_j)[l] \neq \Omega(a_i, a_j)[l-1]$ and $\Omega(a_i, a_j)[l+k-1] \neq \Omega(a_i, a_j)[l+k]$ since some occurrences in $\Omega(a_i, a_j)$ may be identical with each other.

Note that the derived chronicles may not be frequent for a particular clinical activity sequence given a particular clinical workflow log $\mathcal{L}$, even though each temporal constraint of those chronicles may be frequent. For example, as shown in Fig. 4, chronicle $\mathcal{C}_{12}$ is not frequent on the clinical workflow log shown in Table 2, although the temporal constraints in $\mathcal{C}_{12}$ are frequent. To this end, we check the frequency of both temporal constraints and chronicles with respect to a particular workflow log, respectively (Line 16, and Line 64).

The time complexity of the algorithm CCP-Miner has a strong dependence on the generated chronicle number and temporal constraint number. If there are many chronicles and each chronicle has many temporal constraints, the overall complexity of the proposed approach will grow exponentially. Supposing that there is a k-pattern ϕ, such that there are $C_k^2/2$ pairs of clinical activities.

This implies that $C_k^2/2$ temporal constraints in each chronicle can be built on ϕ. For each pair of clinical activities, we assume the number of occurrences in a clinical workflow log is x. Then, at most, there are altogether $x^{C_k^2/2} = O(x^{k^2})$ different chronicles that can be built on ϕ. Since this new factor x^{k^2} is very high, we were forced to find strategies that limit its impact on the discovery time. This included allowing the user to define some milestone clinical activities in advance so that only the pairs between these milestone activities and other interesting activities are considered in building chronicle graph; and then mining chronicle information on the interested activities. For example, we assume that clinical activities *admission*, *surgery* and *discharge* are three milestone activities, and users may only be interested in mining the temporal constraints between *admission*, *surgery*, *discharge* and other clinical activities, respectively, regardless of the temporal constraints among other clinical activities except for *admission*, *surgery*, and *discharge*. Thus, based on the proposed algorithm TC-Miner, we present a top chronicle generate algorithm based on milestone activities, i.e., MATC-Miner, outlined in Algorithm 4.

Algorithm 4 *(The top chronicle generating algorithm based on milestone activities).*

Procedure::MATC-Miner($\mathcal{A}$, B, $\mathcal{L}$)
Input:
  $\mathcal{A}$ is a particular clinical activities sequence
  B is the set of milestone clinical activities that users select
  $\mathcal{L}$ is a clinical workflow log
Output:
  $\mathcal{C}_{\mathcal{A}}^{TOP}$ is the top chronicle with the limitation of B
Steps:
  Let $\mathcal{C}_{\mathcal{A}}^{TOP} = \emptyset$
  For each $(a_i, a_j) \subset \mathcal{A}$ where $a_i \in B$ or $a_j \in B$ do
    $\mathcal{O}(a_i, a_j) \leftarrow \{\langle(a_i, t_i)(a_j, t_j)\rangle | (a_i, t_i) \in \sigma \wedge (a_j, t_j) \in \sigma \wedge \sigma \in \mathcal{L}\}$
    $\Omega(a_i, a_j) \leftarrow$ sort $(\{(t_j - t_i) | \langle(a_i, t_i)(a_j, t_j)\rangle \in \mathcal{O}(a_i, a_j)\})$
    Let $\theta(a_i, a_j) = (a_i, [\Omega(a_i, a_j)[0], \Omega(a_i, a_j)[|\Omega(a_i, a_j)|], a_j)$
    $\mathcal{C}_{\mathcal{A}}^{TOP} \leftarrow \{\theta(a_i, a_j)\}$
  End For
  Return $\mathcal{C}_{\mathcal{A}}^{TOP}$
End Procedure

Combining algorithms SCP-Miner, MATC-Miner, and CCP-Miner, a clinical pathway pattern mining approach is presented. Note that users can run the algorithms several times until they are satisfied by the results. It makes the platform more proactive in pattern elaboration, and thus less tedious for the human analyst. This is the reason why milestone activities are selected by users in advance. The advantage of this strategy is that the analyst can modify and refine his mining request and run the mining process again with the new request, and continue on iteratively. Another advantage is that the mining process can be practical and thus, users can search clinical pathway traces for very complex pattern structures.

5. Experiment

In this section, we compare the proposed approach with sequential pattern mining algorithms on multiple clinical workflow logs, discuss our empirical evaluation, and illustrate how our approach can contribute to clinical pathway redesign.

5.1. Comparison with sequential pattern mining algorithms

The proposed approach consists of two steps: frequent closed clinical activity sequence mining and frequent chronicles mining on discovered clinical activity sequences, provided particular clinical workflow logs. In the experiments, we firstly evaluated the performance of the proposed algorithm SCP-Miner to mine frequent closed clinical activity sequences. To our best knowledge, two efficient frequent closed sequence pattern mining algorithms have been proposed, i.e., CloSpan [45], and BIDE [46]. The algorithm CloSpan follows a candidate maintenance-and-test paradigm over the set of already mined closed sequence candidates. Using CloSpan for mining long sequences or for mining with very low support thresholds tends to be prohibitively expensive [46]. The algorithm BIDE adopts a closure checking scheme, called BI-Directional Extension, which mines closed sequences without candidate maintenance. Performance studies [46] have shown that BIDE is more efficient than CloSpan.

Table 3
Six diseases' clinical workflow logs used in the experiments.

Diseases	# of clinical pathway traces	# of clinical events	# of clinical activities
Bronchial lung cancer	48	3405	225
Gastric cancer	100	8024	274
Cerebral hemorrhage	262	27,949	520
Breast cancer	157	4539	46
Infarction	445	23,106	513
Colon cancer	52	4840	292

However, it may be not efficient to adopt BIDE in mining frequent closed clinical activity sequences directly. As we have mentioned above, clinical pathways have their specific characteristics (e.g., specific starting/ending activity types, etc.). It is, therefore, necessary to design a specific closure checking scheme instead of BI-Directional Extension of BIDE in mining clinical activity sequences. To this end, we present a specific clinical activity sequence mining algorithm, i.e., SCP-Miner, in Section 4.1. In this study, we compare the performance of the proposed algorithm SCP-Miner with the algorithm BIDE [46] using a set of real-life clinical workflow logs of clinical pathways of six diseases recorded by the EMR system in Zhejiang Huzhou Central Hospital of China. The system was brought on-line in August 2007, and the collected data are from 2007/08 to 2009/09. The details of the experimental data set are shown in Table 3. All experiments were performed on a Lenevo Compatible PC with an Intel Pentium IV CPU 2.8 GHz, 4G byte main memory running on Microsoft Windows 7. The algorithms were implemented using Microsoft C#. All run-times in the figures are in seconds.

We must mention that algorithm BIDE can mine frequent closed clinical activity sequences without chronicles information on the sequences. However, it is possible to compare the proposed algorithm SCP-Miner with BIDE, and algorithms SCP-Miner + MATC-Miner + CCP-Miner with BIDE, respectively, in order to investigate the performances of the proposed approach. In particular, for SCP-Miner + MATC-Miner + CCP-Miner, we let *admission*, *surgery* and *discharge* be the milestone activities in order to generate the top chronicle. This means that we consider the temporal constraints between *admission*, *surgery* and *discharge* and other activities in the sequences of closed clinical pathway patterns, regardless of the temporal constraints between those activities except for *admission*, *surgery* and *discharge*.

In addition, we note that for algorithms SCP-Miner and BIDE, we applied the pseudo projection technique in order to save both time and memory space. The main idea of the pseudo projection technique is that instead of generating numerous physical projections in main memory, one can register the index of the projected position with its sequence identifier in the sequence [41,47,44]. Through the indexes, it can easily divide the searching space and then retrieve all the necessary information for finding frequent sequential patterns [41,47,44].

In order to illustrate the proposed approach practically, we compared the proposed SCP-Miner, SCP-Miner + MATC-Miner + CCP-Miner, and BIDE from the perspectives of discovered pattern numbers, run-times, and scalability, respectively.

Fig. 5 summarized the number of frequent patterns and run-times of bronchial lung cancer clinical workflow log, which consists

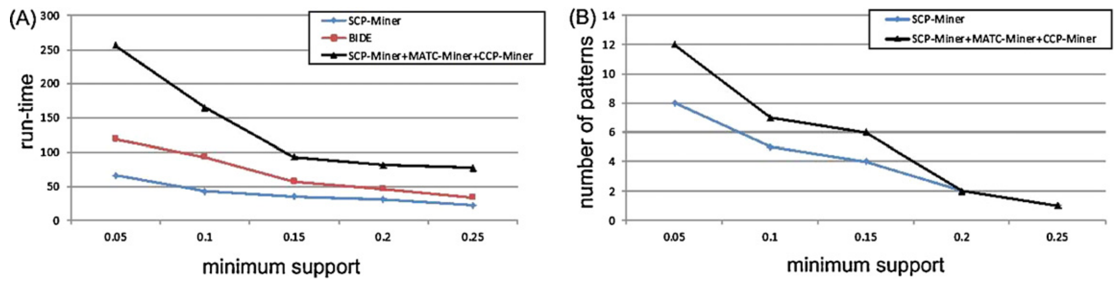

Fig. 5. Bronchial lung cancer clinical workflow log.

of 48 traces. As shown in Fig. 5(A), by applying the algorithm SCP-Miner, it can mine sequences of closed clinical pathway patterns without chronicles information on the generated patterns. It outperforms algorithm BIDE in terms of run-times of mining for closed clinical pathway patterns of bronchial lung cancer. In addition, we can see that by applying SCP-Miner + MATC-Miner + CCP-Miner, the run-time performance approaches algorithm BIDE and SCP-Miner with the increase of minimal support threshold. As indicated in [41], the most important factor influencing run-times of mining frequent patterns is not whether the algorithms or patterns are complicated or not, but whether it generates a large set of patterns, resulting in a longer processing time for these patterns.

Fig. 5(B)[4] shows the experimental results on the discovered number of patterns in comparison with SCP-Miner and SCP-Miner + MATC-Miner + CCP-Miner. As shown in Fig. 5(B), when the minimum support increases, the number of patterns discovered by SCP-Miner + MATC-Miner + CCP-Miner decreases, and is almost equal to the number of sequences discovered by the algorithm SCP-Miner. This results in reducing the processing time. The experimental results confirm this conclusion and reveal the advantages of the proposed approach in mining closed clinical pathway patterns.

Similar to the mining results of bronchial lung-cancer clinical workflow log, the experimental results on the other five diseases, as shown in Figs. 6–10, indicate the feasibility of the proposed approach. In comparison to the relative efficiency of SCP-Miner and BIDE, SCP-Miner always outperforms BIDE. As shown in part (A) of these figures, the run-times of the algorithm SCP-Miner increases very slowly with minimum support decreases. Furthermore, the run-times of SCP-Miner + MATC-Miner + CCP-Miner approaches BIDE and SCP-Miner with the minimum support threshold increases. As indicated in part (B) of these figures, with the minimum support increases, SCP-Miner generates quite a smaller set of patterns even at the low minimum support threshold. As well, the number of patterns discovered by SCP-Miner + MATC-Miner + CCP-Miner is almost equal to the performance of SCP-Miner, especially with the increases of minimum support. The experimental results indicate that the proposed approach is suitable to mine closed clinical pathway patterns.

Next, we will study how the proposed approach performs with the increasing size of a clinical workflow log. Fig. 11 shows how SCP-Miner, SCP-Miner + MATC-Miner + CCP-Miner, and BIDE scale up as the number of input-clinical pathway traces is increased, from 100 to 445. We note that all the experiments were performed on the infarction clinical workflow log with the same minimum support threshold of 0.25%. The execution times are normalized with respect to the time for the 100 input-traces. It can be observed that both SCP-Miner and SCP-Miner + MATC-Miner + CCP-Miner have a linear scalability in terms of the run-times against the increasing number of traces.

5.2. Discussion

We have implemented and tested the proposed approach using Microsoft C#. Fig. 12 depicts a screen-shot of mining results of the breast cancer process. On the top of Fig. 12, clinical activities that belong to one closed clinical pathway pattern are listed sequentially along the time-line of patient LOS. In addition, temporal constraints between the milestone activities (i.e., *admission*, *surgery*,*discharge*) and other interesting activities are shown on the bottom of Fig. 13 . We note that users can either select all clinical activities of one pathway pattern in order to display the temporal relations, or can also select several interesting activities, and display their temporal relations with milestone activities on the Figure. The discovered clinical process patterns have been evaluated by the medical staff at the Zhejiang Huzhou Central Hospital of China, who understand the beneficial effects of the clinical process mining of medical behaviors. They also fully understand the mining results of our approach. They indicate that the mining results of our approach: (1) allow clinical activities to be clearly spread along the time-line of patient LOS; (2) allow for certain temporal relationships to explicitly exist between the activities; and (3) let a clinical process pattern enumerate regular medical behaviors that are expected to occur in patient-care journeys, which serve as checkpoints for the performance of the patient-care journey. We would like to mention that physicians at the Zhejiang Huzhou Central Hospital of China are satisfied with the mined results. The evaluations received from medical staff indicate that the proposed approach has the ability to find a clear characterization of possible clinical pathway patterns for particular diseases.

Finally, we use a simple example to illustrate how the discovered patterns can contribute to clinical pathway redesign. As shown in Fig. 13(A), there is a fragment of bronchial lung cancer clinical pathway recommended by the Chinese Ministry of Health. In Fig. 13(B), there is a fragment of bronchial lung cancer pathway pattern defined by physicians at the Zhejiang Huzhou Central Hospital of China. As well, Fig. 13(C) highlights a fragment of discovered pattern from the collected logs. These three pattern fragments consist of three milestone activities, i.e., *admission*, *surgery* and *discharge*, and the temporal constraints among three activities. We can see that the temporal constraints, in three pattern fragments, reveals the different time spans between any two clinical activities. For example, in the recommended clinical pathway, activity *surgery* is assumed to be performed after 4–7 days of *admission*, and activity *discharge* is assumed to be performed after 8–14 days of *surgery*; in the physicians' defined pattern, activity *surgery* is assumed to be performed after 4 days of *admission*, and activity *discharge* is assumed to be performed after 8 days of *surgery*, while in actual patient-care journeys, the activity *surgery* is occurred between 3 and 4 days after *admission*, and clinical activity *discharge* is occurred

[4] We must mention that the algorithm BIDE discovers the same number of patterns with the algorithm SCP-Miner, since SCP-Miner is designed based on the principle of BIDE except using the forward closure checking scheme instead of BI-Directional Extension. Thus, we have not presented the mined results of BIDE in the discovered number of patterns.

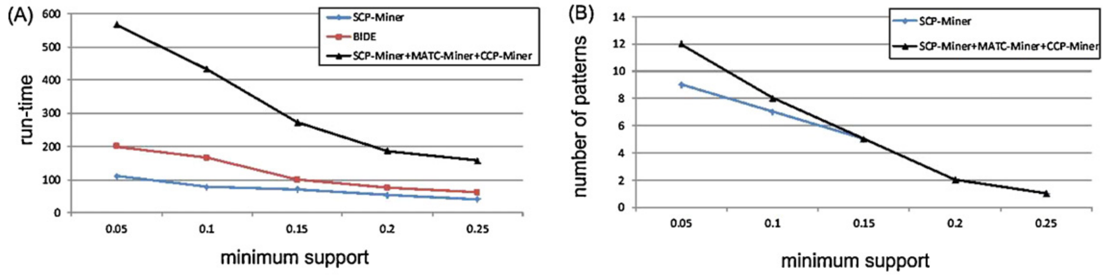

Fig. 6. Gastric cancer clinical workflow log.

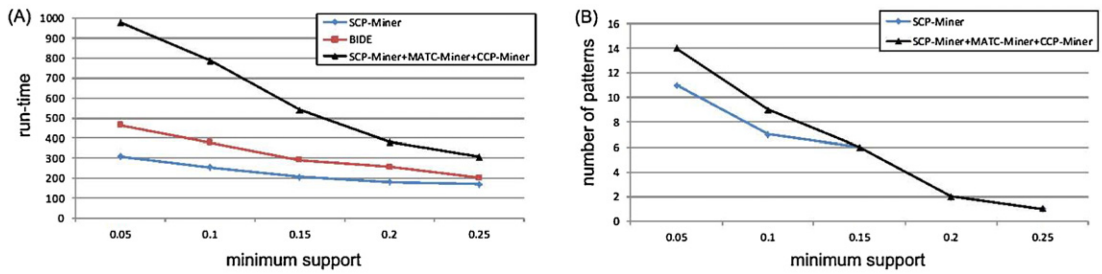

Fig. 7. Cerebral hemorrhage clinical workflow log.

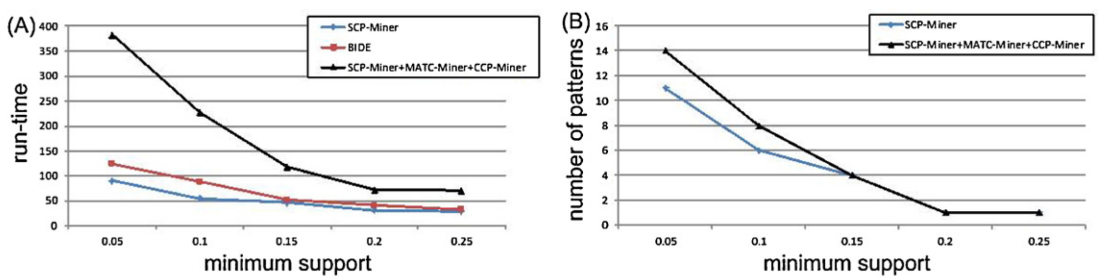

Fig. 8. Breast cancer clinical workflow log.

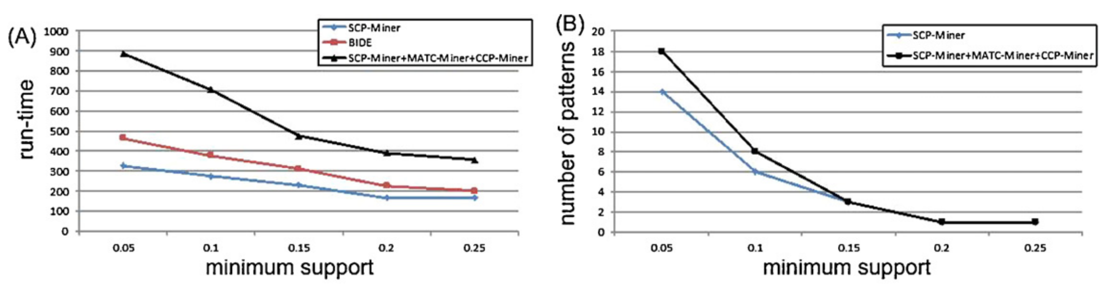

Fig. 9. Infarction clinical workflow log.

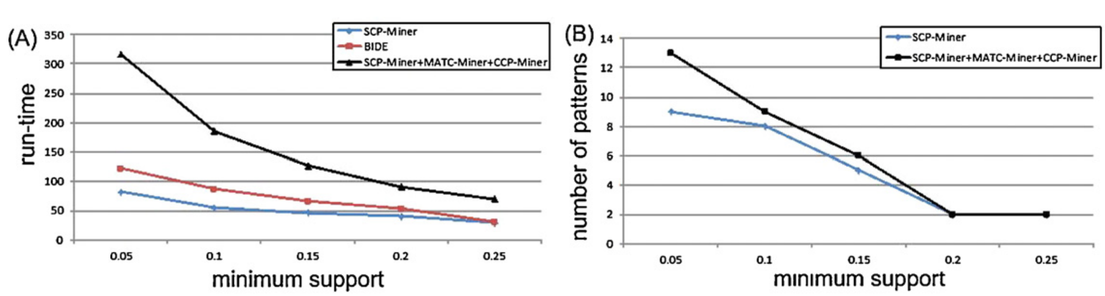

Fig. 10. Colon cancer clinical workflow log.

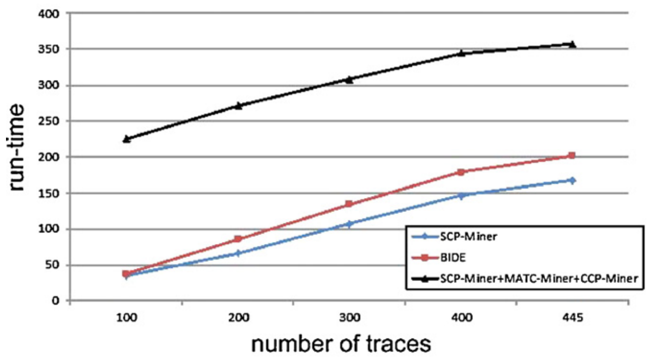

Fig. 11. Scale-up: number of input-traces of infarction clinical workflow log.

between 12 and 17 days after *surgery*. Thus, discovered information, as a reflection of actual medical behaviors in patient-care journeys, can be used to improve clinical pathway design and enactment.

We note that, in real-life situations, physicians or hospital managers are using medical protocols, also named clinical practice guidelines [2,12], to define specific clinical pathway patterns to treat patients more frequently [2]. These base patterns can automatically suggest clinical pathways for individual patients. Note that physicians can deviate from these base patterns when needed. Since medical behaviors of treating individual patients can be discovered by the proposed approach, this suggests that it is increasingly more interesting to compare predefined base patterns with discovered patterns from clinical workflow log. The results of this analysis can assist physicians in continuously (re)designing and optimizing clinical pathways.

It should be mentioned that there are certain limitations to the current approach:

- In this study, we deal with point-based clinical events instead of interval-based events, which may lose expressivity in mining clinical pathway patterns. As we mentioned above, the research on mining temporal patterns from interval-based events attempted to find the temporal relationships between these interval events. However, when using point-based events instead of interval-based data, it is not a straightforward process to derive interval relationships (e.g., overlaps, during, etc.) [48,49,13,50] from a point based representation.
- In a clinical pathway, it may be possible to execute the same activity multiple times. If this happens, this typically refers to a loop in the corresponding model. Note that loops can be used to jump back to any place in the pathway.
- In this study, we assume the information in the clinical workflow log is correct. Although this is a valid assumption in most situations, the log may contain "noise", i.e., incorrectly logged information. For example, it could be that certain events are not recorded or are recorded some time after the event actually took place. The mining algorithm needs to be robust with respect to noise, i.e., medical behaviors should not be evaluated based on a single observation. As a matter of fact, one could argue that the mining algorithm needs to distinguish deviations from the normal pathway. When considering noise, one often has to

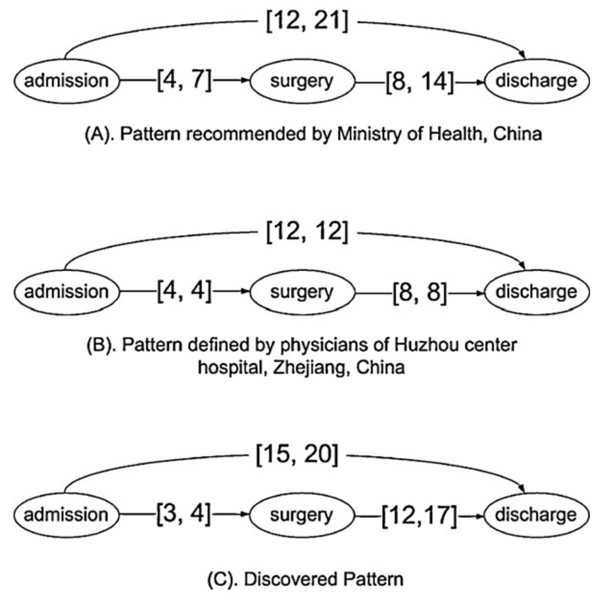

Fig. 13. Comparison between a predefined pattern by physicians and a discovered pattern.

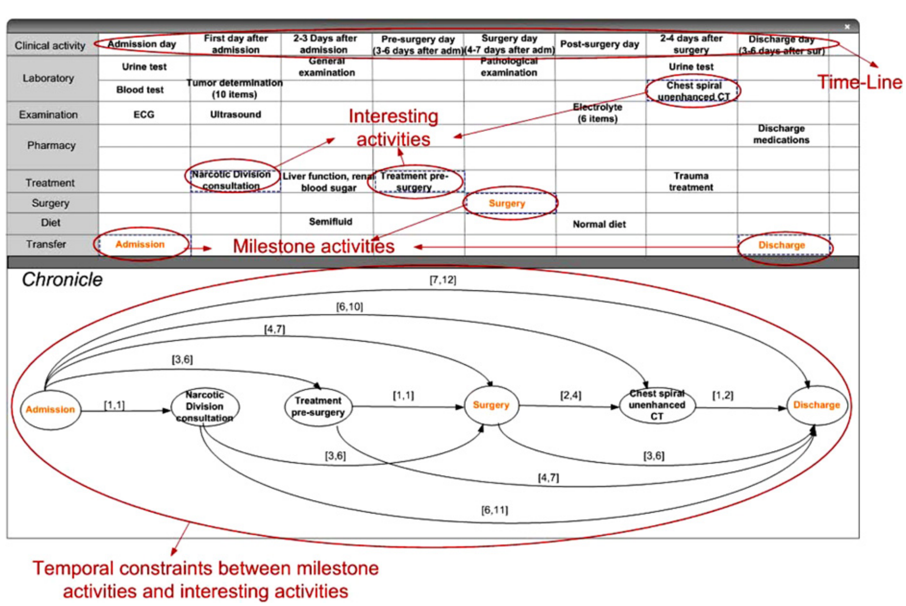

Fig. 12. An example of mining results by using the proposed approach.

determine a threshold value to cut off exceptional or incorrectly logged behavior.

6. Conclusion and future work

Clinical pathways are standardized patient-care processes. Hospital managers have presented their requirements to use tools to analyze medical behaviors in patient-care processes so as to continuously (re)design and optimize clinical pathways [4]. The approach proposed in this study can be viewed as a technology that contributes to this purpose. Our goal is to extract explicit clinical pathway patterns from medical behaviors, which are recorded in clinical workflow logs. Thus, the challenge is to create clinical pathway patterns given a log, such that discovered patterns are consistent with the observed dynamic behavior. The experimental results indicate that the proposed approach provides the ability to discover clinical pathway patterns that cover the most frequent medical behaviors which are regularly encountered in clinical practice.

As mentioned above, discovered clinical pathway patterns have been evaluated by clinical experts and hospital managers from the Zhejiang Huzhou Central Hospital of China. These individuals indicate that the proposed approach can provide a consistent characterization of all possible clinical pathway patterns for particular diseases. Notably, a fully development of a clinical pathway modeling and analysis tool is finishing, which will be employed in the EMR system in the hospital.

Given the relevance of the analysis of medical behaviors and the problems that experts have in making good clinical pathway models, we will continue to work on the topics mentioned in this paper. Note that the bottleneck of clinical pathway pattern mining is not based on whether users can derive the complete set of clinical pathway patterns efficiently, but rather on whether they can derive a compact but high quality set of patterns that can cover most useful medical behaviors in clinical practice. Although our study reveals that the proposed approach is effective in analyzing medical behaviors and discovering efficient clinical pathway patterns, there are even more complex analysis and evaluation tasks that need to be considered.

In fact, clinical experts at the Zhejiang Huzhou Central Hospital of China, have indicated that, even though our approach is efficient for mining precise and complete set of regular medical behaviors in clinical pathways, there are still a number of infrequent behaviors that are missing in the discovered patterns. Note that these infrequent behaviors may represent interesting variants in clinical pathways, and thus need to be discovered and analyzed. Approximate frequent patterns [44,51] could be a possible choice to handle variants in clinical pathways. As well, sequence clustering could also be used to classify and analyze variants in clinical pathways [27,52,53]. However, the interesting questions that remain address the issues of how to design efficient algorithms for mining and detecting variants in clinical pathways, as well as, how to explain these variants in a maximum-informative manner. Much research is still needed to make such mining both effective and efficient.

In addition, to make clinical pathway pattern mining an essential task in clinical practice, much research is needed to further develop pattern-based mining methods. For example, how does one construct an efficient classification model using the discovered clinical pathway patterns in order to specialize in clinical pathways? What sorts of clinical pathway patterns are more effective and discriminative than other patterns in treating particular patients? How does one measure the adherence between the discovered clinical pathway patterns and actual medical behaviors, in order to assist medical staff to analyze clinical pathways? These questions need to be answered before discovered clinical pathway patterns can play an essential role in clinical applications.

Acknowledgements

This work was supported by the National Nature Science Foundation of China under Grant No. 81101126. 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

[1] Wakamiya S, Yamauchi K. What are the standard functions of electronic clinical pathways? International Journal of Medical Informatics 2009;78(8):543–50.
[2] Lenz R, Reichert M. IT support for healthcare processes-premises, challenges, perspectives. Data & Knowledge Engineering 2007;61(1):39–58.
[3] Lenz R, Blaser R, Beyer M, Heger O, Biber C, Bäumlein M, et al. IT support for clinical pathways-lessons learned. International Journal of Medical Informatics 2007;76(3):S397–402.
[4] Schuld J, Schäfer T, Nickel S, Jacob P, Schilling MK, Richter S. Impact of IT-supported clinical pathways on medical staff satisfaction. a prospective longitudinal cohort study. International Journal of Medical Informatics 2011;80(3):151–6.
[5] Quaglini S, Stefanelli M, Lanzola G, Caporusso V, Panzarasa S. Flexible guideline-based patient careflow systems. Artificial Intelligence in Medicine 2001;22(1):65–80.
[6] Hunter B, Segrott J. Re-mapping client journeys and professional identities: a review of the literature on clinical pathways. International Journal of Nursing Studies 2008;45:608–25.
[7] Weiland DE. Why use clinical pathways rather than practice guidelines? American Journal of Surgery 1997;174:592–5.
[8] Uzark K. Clinical pathways for monitoring and advancing congenital heart disease care. Progress in Pediatric Cardiology 2003;18:131–9.
[9] Loeb M, Carusone SC, Goeree R, Walter SD, Brazil K, Krueger P, et al. Effect of a clinical pathway to reduce hospitalizations in nursing home residents with pneumonia. Journal of the American Medical Association 2006;295:2503–10.
[10] Zand DJ, Brown KM, Konecki UL, Campbell JK, Salehi V, Chamberlain JM. Effectiveness of a clinical pathway for the emergency treatment of patients with inborn errors of metabolism. Pediatrics 2008;122:1191–5.
[11] Alexandrou D, Skitsas I, Mentzas G. A holistic environment for the design and execution of self-adaptive clinical pathways. IEEE Transactions on Information Technology in Biomedicine 2011;15(1):108–18.
[12] Huang Z, Lu X, Duan H. Using recommendation to support adaptive clinical pathways. Journal of Medical Systems 2011:1–12, 10.1007/s10916-010-9644-3.
[13] Combi C, Gozzi M, Oliboni B, Juarez JM, Marin R. Temporal similarity measures for querying clinical workflows. Artificial Intelligence in Medicine 2009;46(1):37–54.
[14] Lu X, Huang Z, Duan H. Supporting adaptive clinical treatment processes through recommendations. Computer Methods and Programs in Biomedicine 2011.
[15] Huang Z, Lu X, Gan C, Duan H. Variation prediction in clinical processes. In: Peleg M, Lavrac N, Combi C, editors. Artificial intelligence in medicine, vol. 6747 of Lecture Notes in Computer Science. Berlin/Heidelberg: Springer; 2011. p. 286–95.
[16] Peleg M, Mulyar N, van der Aalst WMP. Pattern-based analysis of computer-interpretable guidelines: don't forget the context. Artificial Intelligence in Medicine 2012;54(1):73–4.
[17] Campbell H, Hotchkiss R, Bradshaw N, Porteous M. Integrated care pathways. British Medical Journal 1998;316:133–7.
[18] Cheah J. Development and implementation of a clinical pathway programme in an acute care general hospital in Singapore. International Journal for Quality in Health Care 2000;12:403–12.
[19] Dousson C, Duong TV. Discovering chronicles with numerical time constraints from alarm logs for monitoring dynamic systems. In: Thomas Dean, editor. Proceedings of the 16th international joint conference on artificial intelligence. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1999. p. 620–6.
[20] Dousson C, Maigat PL, France Telecom R&D. Chronicle recognition improvement using temporal focusing and hierarchization. In: Veloso Manuela M, editor. Proceedings of the 20th international joint conference on artificial intelligence. Menlo Park, CA: IJCAI/AAAI Press; 2007. p. 324–9.
[21] Cram D, Mathern B, Mille A. A complete chronicle discovery approach: application to activity analysis. Expert Systems 2011.
[22] Westbrook JI, Coiera EW, Gosling AS, Braithwaite J. Critical incidents and journey mapping as techniques to evaluate the impact of online evidence retrieval systems on health care delivery and patient outcomes. International Journal of Medical Informatics 2007;76:234–45.

[23] Agrawal R, Gunopulos D, Leymann F. Mining process models from workflow logs. In: Schek HJ, Saltor F, Ramos I, Alonso G, editors. Sixth international conference on extending database technology. London: Springer-Verlag; 1998. p. 469–83.
[24] Cook JE, Wolf AL. Discovering models of software processes from event-based data. ACM Transactions on Software Engineering and Methodology 1998;7(3):215–49.
[25] Yang W, Hwang S. A process-mining framework for the detection of healthcare fraud and abuse. Expert Systems with Applications 2006;31(1):56–68.
[26] van der Aalst WMP, Weijters AJMM, Maruster L. Workflow mining: discovering process models from event logs. IEEE Transactions on Knowledge and Data Engineering 2004;16(9):1128–42.
[27] Rebuge A, Ferreira DR. Business process analysis in healthcare environments: a methodology based on process mining. Information Systems 2012;37(2):99–116.
[28] Lang M, urkle TB, Laumann S, Prokosch H-U. Process mining for clinical workflows: challenges and current limitations. In: Andersen SK, Klein GO, Schulz S, Aarts J, editors. Proceedings of MIE2008 the XXI international congress of the European Federation for Medical Informatics. 2008. p. 229–34.
[29] Günther WC, Rozinat A, van der Aalst WMP. Activity mining by global trace segmentation. In: van der Aalst WMP, Mylopoulos J, Sadeh NM, Shaw MJ, Szyperski C, Rinderle-Ma S, et al., editors. Business process management workshops, vol. 43 of Lecture Notes in Business Information Processing. Berlin/Heidelberg: Springer; 2010. p. 128–39.
[30] Daniyal A, Abidi S. Semantic web-based modeling of clinical pathways using the UML activity diagrams and OWL-S. In: David Riaño, editor. Knowledge representation for health-care. Data, processes and guidelines, vol. 5943 of Lecture Notes in Computer Science. Berlin/Heidelberg: Springer; 2010. p. 88–99.
[31] Bryan S, Holmes S, Prostlethwaite D, Carty N. The role of integrated care pathways in improving the client experience. Professional Nurse 2002;18(2):77–9.
[32] Ye Y, Jiang Z, Diao D, Yang D, Du G. An ontology-based hierarchical semantic modeling approach to clinical pathway workflows. Computers in Biology and Medicine 2009;39:722–32.
[33] Li C. Mining process model variants: challenges, techniques, examples. PhD thesis, University of Twente, The Netherlands; 2010.
[34] Darnton G, Darton M. Business process analysis. CA: International Thompson Business Press; 1997.
[35] Hwang S, Wang C, Yang W. Discovery of temporal patterns from process instances. Computers in Industry 2004;53:345–64.
[36] van der Aalst WMP, Reijers HA, Weijters AJMM, van Dongen BF, Alves de Medeiros AK, Song M, et al. Business process mining: an industrial application. Information Systems 2007;32(5):713–32.
[37] Greco G, Guzzo A, Manco G, Sacca D. Mining and reasoning on workflows. IEEE Transactions on Knowledge and Data Engineering 2005;17:519–34.
[38] Cardoso J, Lenic M. Web process and workflow path mining using the multimethod approach. International Journal of Business Intelligence and Data Mining 2006;1:304–28.
[39] van de Klundert J, Gorissen P, Zeemering S. Measuring clinical pathway adherence. Journal of Biomedical Informatics 2010;43(6):861–72.
[40] Lin F, Chen S, Pan S, Chen Y. Mining time dependency patterns in clinical pathways. International Journal of Medical Informatics 2001;62(1):11–25.
[41] Hu Y, Huang TCK, Yang HR, Chen YL. On mining multi-time-interval sequential patterns. Data & Knowledge Engineering 2009;68(10):1112–27.
[42] Wu SY, Chen YL. Discovering hybrid temporal patterns from sequences consisting of point- and interval-based events. Data & Knowledge Engineering 2009;68(11):1309–30.
[43] Chen YL, Wu SY, Wang YC. Discovering multi-label temporal patterns in sequence databases. Information Sciences 2011;181(3):398–418.
[44] Han J, Cheng H, Xin D, Yan X. Frequent pattern mining: current status and future directions. Data Mining and Knowledge Discovery 2007;15:55–86.
[45] Yan X, Han J, Atshar R. CloSpan: mining closed sequential patterns in large datasets. In: Daniel Barbar, Chandrika Kamath, editors. Proceedings of the third SIAM international conference on data mining. San Francisco, CA, USA: SIAM; 2003. p. 166–77.
[46] Wang J, Han J, Li C. Frequent closed sequence mining without candidate maintenance. IEEE Transactions on Knowledge and Data Engineering 2007;19(8):1042–56.
[47] Tzvetkov P, Yan X, Han J. TSP: mining top-k closed sequential patterns. Knowledge and Information Systems 2005;7:438–57.
[48] Adlassnig K-P, Combi C, Das AK, Keravnou ET, Pozzi G. Temporal representation and reasoning in medicine: research directions and challenges. Artificial Intelligence in Medicine 2006;38(2):101–13.
[49] Sacchi L, Larizza C, Combi C, Bellazzi R. Data mining with temporal abstractions: learning rules from time series. Data Mining and Knowledge Discovery 2007;15(2):217–47.
[50] Combi C, Oliboni B. Visually defining and querying consistent multi-granular clinical temporal abstractions. Artificial Intelligence in Medicine 2012;54(2):75–101.
[51] Li C, Jea K. An adaptive approximation method to discover frequent itemsets over sliding-window-based data streams. Expert Systems with Applications 2011;38(10):13386–404.
[52] Ferreira DR, Zacarias M, Malheiros M, Ferreira P. Approaching process mining with sequence clustering: experiments and findings. In: Alonso G, Dadam P, Rosemann M, editors. vol. 4714 of the Lecture Notes in Computer Science. Berlin/Heidelberg: Springer; 2007. p. 360–74.
[53] Ferreira DR. Applied sequence clustering techniques for process mining. In: Cardoso J, van der Aalst W, editors. Handbook of research on business process modeling, information science reference. IGI Global; 2009. p. 492–513.