Journal of Biomedical Informatics 83 (2018) 178–195

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

A data-driven framework of typical treatment process extraction and evaluation

Jingfeng Chen[a], Leilei Sun[a,b], Chonghui Guo[a,*], Wei Wei[a], Yanming Xie[c]

[a] Institute of Systems Engineering, Dalian University of Technology, Dalian 116024, PR China
[b] School of Economics and Management, Tsinghua University, Beijing 100084, PR China
[c] Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Scieces, Beijing 100700, PR China

ARTICLE INFO

Keywords:
EMR data mining
Set sequence similarity
AP clustering
Treatment process discovery

ABSTRACT

Background: A clinical pathway (CP) defines a standardized care process for a well-defined patient group that aims to improve patient outcomes and promote patient safety. However, the construction of a new pathway from scratch is a time-consuming task for medical staff because it involves many factors, including objects, multidisciplinary collaboration, sequential design, and outcome measurements. Recently, the rapid development of hospital information systems has allowed the storage of large volumes of electronic medical records (EMRs), and this information constitutes an abundant data resource for building CPs using data-mining methods.
Methods: We provide an automatic method for extracting typical treatment processes from EMRs that consists of four key steps. First, a novel similarity method is proposed to measure the similarity of two treatment records. Then, we perform an affinity propagation (AP) clustering algorithm to cluster doctor order set sequences (DOSSs). Next, a framework is proposed to extract a high-level description of each treatment cluster. Finally, we evaluate the extracted typical treatment processes by matching the treatment cluster with external information, such as the treatment efficacy, length of stay, and treatment cost.
Results: By experiments on EMRs of 8287 cerebral infarction patients, it is concluded that our proposed method can effectively extract typical treatment processes from treatment records, and also has great potential to improve treatment outcome by personalizing the treatment process for patients with different conditions.
Conclusion: The extracted typical treatment processes are intuitive and can provide managerial guidance for CP redesign and optimization. In addition, our work can assist clinicians in clearly understanding their routine treatment processes and recommend optimal treatment pathways for patients.

1. Introduction

In the medical health system, clinical pathways (CPs) are regarded as useful tools that ease the tension of the doctor-patient relationship and enable patients to receive correct and timely diagnosis and treatment with controlled medical costs and improved medical quality [1–4]. However, in practice, the creation of such a pathway from scratch is demanding for medical staff [5] because it involves multidisciplinary medical team collaboration (doctors, nurses, and other comedicals), plan-do-check-act (PDCA)-related techniques, and optimal evidence-based medicine. Recently, the rapid development of medical information technology has resulted in the availability of massive electronic medical records (EMRs), which has enabled a new paradigm for optimizing healthcare practices [6,7]. EMRs conceal an untapped reservoir of knowledge concerning specific therapy and treatment activities performed on patients during their treatment processes [8,9]. Thus, the exploration of data-driven methods from EMRs would allow the mining of typical treatment processes for CP design and improvement.

Process mining [10–13], as a general method for extracting information from event logs in business process analysis, has been used to analyse healthcare processes (e.g., CPs) [14,15], and most process mining algorithms can automatically construct process pattern models in a structured manner and are well suited to the understanding, redesign, and continuous improvement of business process [16]. Furthermore, process mining can be divided into three main types: process discovery, conformance checking, and enhancement [11,17]. A previous study [17] explained how automatic process discovery allows the extraction of process models from an event log, how conformance checking allows the monitoring of deviations by comparing a given model with the event log, and how enhancement allows extension or improvement of an existing process model using information of the

* Corresponding author.
E-mail address: dlutguo@dlut.edu.cn (C. Guo).

https://doi.org/10.1016/j.jbi.2018.06.004
Received 16 December 2017; Received in revised form 8 May 2018; Accepted 8 June 2018
Available online 15 June 2018
1532-0464/

actual process recorded in the event log. In particular, ProM, a pluggable tool with the capacity for multiple techniques [18,19], can automatically construct treatment process models aiming to explain the medical behaviour observed in EMRs by executing data-mining algorithms on event logs and can provide advanced visualization and verification capabilities [10,11,14,15]. Recently, considerable research has focused on the development of sequential pattern mining approaches to discover frequent treatment behaviour patterns from clinical data. Such information can reveal which critical clinical activities are performed and in what order and can provide comprehensive knowledge of quantified temporal orders of clinical activities in healthcare processes [20–23]. However, for real patient data, the use of traditional process mining techniques may generate spaghetti-like treatment processes, and sequential pattern mining approaches can create explosions of patterns that are challenging for clinical experts to comprehend. Furthermore, the assumption that processes occurs in a structured fashion with specific timestamps is not valid for clinical activities because many treatment behaviours can occur arbitrarily without a particular order, particularly for doctor orders with time deviations between clinician advice and patient compliance. Therefore, it is necessary to redefine the sequence of treatment activities and design a new process mining technique to effectively extract typical treatment processes.

Data mining is the process of discovering patterns in large datasets using methods at the intersection of machine learning, statistics, and database systems [24]. In recent years, data mining techniques have become increasingly popular in the healthcare field and have the ability to promote EMR knowledge discovery aiming to improve research applications and clinical care, including clinical risk stratification, phenotype stratification, early detection and diagnosis, drug-adverse reaction associations, comorbidity identification, pharmacovigilance, CP discovery, disease progression, and precision medicine [7–9]. While in the data-mining process, the similarity measure is a common technique for providing a measure of distance in the feature space and has been studied extensively in multiple fields. In the healthcare informatics domain, a considerable amount of research has focused on patient (treatment trace or sequence) similarity [25]. Patient similarity, which measures how similar a pair of patients are according to their historical information under a specific clinical context, is expected to enable various healthcare applications possible, such as cohort analysis, case-based reasoning, treatment comparison, disease subtyping and CP analysis [26]. Traditional techniques of patient similarity measures are focused on direct matching between patients by applying common classical distance concepts. However, these traditional approaches may not be appropriate for measuring similarities between patient traces or sequences for treatment process analysis [27]. Thus, novel similarity-measurement methods of patient similarity have been proposed to build and analyse treatment processes, such as the topic similarity measure with cosine distance between patient traces based on latent Dirichlet allocation (LDA) [27], typicalness index [5], duration-aware pairwise trace alignment [28], dynamic time warping (DTW) [29], time-warping similarity of process traces [30], set similarity measure based on doctor order content (SS-DOC) [31], fine-grained patient similarity measure using deep metric learning [32], and multi-view similarity network fusion for aggregating diverse data types [33]. These methods achieve relatively high efficiency for treatment process analysis and discovery, but some limitations remain, such as more specific timestamps for the order of treatment activities, the interpretability and visualization of results, and operability for clinicians. Despite these limitations, research regarding these novel similarity-measurement methods has provided comprehensive theoretical guidance and improvement directions for our work.

In this study, we propose a novel similarity measure method for mining typical treatment processes from EMRs for CPs design. First, we redefine the order of concurrent treatment activities and design a doctor order set sequence (DOSS) similarity measure method to compute the pairwise similarity between treatments. We then implement an affinity propagation (AP) clustering algorithm to cluster DOSSs and extract typical treatment processes using the *K*-nearest-neighbours (KNN) approach according to the AP clustering results. The extracted typical treatment processes are then evaluated from two perspectives, namely, the treatment efficacy and the treatment efficiency (e.g., economy and length of stay). To accomplish this, we validate our methods and evaluate our results with the real EMRs of 8287 cerebral infarction patients collected from three traditional Chinese medicine (TCM) hospitals.

The remainder of the paper is organized as follows. Section 2 presents the preliminary definitions and formalization of EMRs. Section 3 describes the proposed methods for extracting typical treatment processes from doctor orders and designs the general framework for evaluating the extracted results. Section 4 presents our experimental results and discussion. Finally, some conclusions are provided in Section 5.

2. Definitions and formalization

EMRs have been used extensively by general practitioners in multiple developed countries and typically include patient demographic information, diagnostic information, laboratory indicators, doctor orders, and outcomes [31]. The objective of this study is to extract typical treatment processes from EMRs for clinical doctors to understand the relationship between treatments and make better clinical decisions. This section presents the definitions and formalization.

Definition 1 (*Doctor Order*). A doctor order is a medical prescription implemented by a physician in the clinical course of treatment for a patient. In general, the doctor order is a tuple of attributes, such as the drug name, start-end time, therapeutic function (i.e., drug efficacy (DE)), delivery route, dosage each time, and frequency per day. This study focuses on the first two attributes to simplify complexity; thus, a doctor order can be represented as

$$Order = (O^{DN}, t^{O}),$$

where O^{DN} represents the drug name (DN), t^{O} records the active time of the order.

We depict a simple example of two treatments for cerebral infarction patients during a two-week period in Fig. 1. In this example, each treatment consists of a set of doctor orders that are most relevant to cerebral infarction and represented by the DN. Additionally, Fig. 1 shows three characteristics for doctor orders: (1) numerous DNs repeatedly appear in the treatment of a patient; (2) the timestamp of the doctor orders (e.g., start timestamp and end timestamp) are significantly inconsistent with real execution time (i.e., doctor orders with time deviations between clinician advice and patient compliance), but the treatment day or period is correct; and (3) different doctor orders may have the same timestamp due to the prescribing habits of the clinicians and the drug coadministration. These complex characteristics present a challenge for existing process-mining methods; therefore, a new processing method for doctor orders is essential to further mine treatment patters from EMRs.

Definition 2 (*DOSS*). A treatment of a patient is a series of doctor orders related with the patient, which can be represented as

$$T = \{(OS_1, t_1^T), (OS_2, t_2^T), \ldots, (OS_{d^T}, t_{d^T}^T)\},$$

where t_1^T and $t_{d^T}^T$ are the first and last treatment days, respectively. OS_k represents the set of orders that occurred on the *k*-th treatment day.

Considering the characteristics of doctor orders, a treatment can be divided into different periods like treatment courses in clinical practice; thus, for a treatment of a patient, we can define the DOSS by

$$DOSS = \{Init_{P_0}, OS_{P_1}, \ldots, OS_{P_q}, \ldots, OS_{P_Q}, End_{P_{Q+1}}\},$$

where the *Init* and *End* are added to represent the beginning and end period of the patient treatment, respectively, and a doctor order set in