Data-driven Automatic Treatment Regimen Development and Recommendation

Leilei Sun
Institute of Systems Engineering
Dalian University of Technology
leisun@mail.dlut.edu.cn

Chuanren Liu
Department of Decision Sciences and MIS
LeBow College of Business
Drexel University
chuanren.liu@drexel.edu

Chonghui Guo
Institute of Systems Engineering
Dalian University of Technology
dlutguo@dlut.edu.cn

Hui Xiong*
Department of Management Science and Information Systems
Rutgers University
hxiong@rutgers.edu

Yanming Xie
Institute of Basic Research in Clinical Medicine
China Academy of Chinese Medical Sciences
ktzu2015@163.com

ABSTRACT

The analysis of large-scale Electrical Medical Records (EMRs) has the potential to develop and optimize clinical treatment regimens. A treatment regimen usually includes a series of doctor orders containing rich temporal and heterogeneous information. However, in many existing studies, a doctor order is simplified as an event code and a treatment record is simplified as a code sequence. Thus, the information inherent in doctor orders is not fully used for in-depth analysis. In this paper, we aim at exploiting the rich information in doctor orders and developing data-driven approaches for improving clinical treatments. To this end, we first propose a novel method to measure the similarities between treatment records with consideration of sequential and multifaceted information in doctor orders. Then, we propose an efficient density-based clustering algorithm to summarize large-scale treatment records, and extract a semantic representation of each treatment cluster. Finally, we develop a unified framework to evaluate the discovered treatment regimens, and find the most effective treatment regimen for new patients. In the empirical study, we validate our methods with EMRs of 27,678 patients from 14 hospitals. The results show that: 1) Our method can successfully extract typical treatment regimens from large-scale treatment records. The extracted treatment regimens are intuitive and provide managerial implications for treatment regimen design and optimization. 2) By recommending the most effective treatment regimens, the total *cure rate* in our data improves from 19.89% to 21.28%, and the *effective rate* increases up to 98.29%.

*Corresponding Author

KDD '16, August 13-17, 2016, San Francisco, CA, USA

DOI: http://dx.doi.org/10.1145/2939672.2939866

CCS Concepts

•Information systems → Data mining; •Applied computing → Health care information systems;

Keywords

Treatment Regimen; Treatment Recommendation; Electronic Medical Records; Temporal Sets.

1. INTRODUCTION

The availability of massive Electronic Medical Records (EMRs) has enabled a new paradigm for optimizing healthcare practices. Indeed, it becomes a key to success to improve health and treat disease by exploring the EMRs data. There are various successful examples, such as health surveillance, risk assessment, disease diagnosis, and treatment planning. In particular, the White House's Precision Medicine Initiative will be a data-driven enterprise, using health data of a million or more Americans to catalyze a new era of data-based and more precise medical treatment.

While there are tremendous interests in exploiting EMRs data for improving medical treatment, what have been obtained from the analysis of EMRs data is far less than what EMRs can provide [8]. According to [16], one reason is that a patient's outcome is influenced by a lot of factors, such as the age and gender of the patient, disease severity, and treatment received. Although EMRs data contain comprehensive information about the patients, diagnosis, and treatments, there is no consensus framework for integrating all the related factors for advanced data modeling. Moreover, EMRs data are heterogeneous and longitudinal in nature. For example, a treatment record is series of doctor orders, where each doctor order usually consists of medicine name, delivery route, dosage, starting time, and ending time.[1] Overall, it is a non-trivial challenge to analyze the large-scale and complex EMRs data to extract healthcare knowledge and facilitate decision-making in treatment practices.

[1]The definition of *doctor order* will be given in Section 2.

To this end, in this paper, we focus on two problems: 1) identifying typical treatment regimens from large-scale treatment records; and 2) quantitatively measuring the effectiveness of a typical treatment regimen for a specified patient cohort. Our motivation is that the typical treatment regimens are usually used as prototypes when a clinical doctor designs the personalized treatment plan for a new patient. As a result, the automatic identification and evaluation of the typical treatment regimens are essentially important to support the diagnostic and treatment decisions, and ultimately improve the *treatment effective rate* and the *cure rate*. The identified typical treatment regimens are also helpful for healthcare researchers and clinical doctors to develop new treatment regimens.

To address the aforementioned challenges in identifying and evaluating the typical treatment regimens from the complex EMRs data, first, we need an effective method to measure the similarity between treatment records containing sequential and multifaceted information in doctor orders. Second, based on the similarity measurements between treatment records, we need to group the treatment records to identify the treatment regimens. At the same time, to make the results understandable to the doctors and the patients, we need to extract semantically meaningful descriptions of treatment regimen clusters. In the literature, though many clustering algorithms have been developed, most of the previous efforts focus on assigning each object with a cluster label rather than extracting a semantic description for each cluster [7]. Finally, we also need to estimate the treatment outcome of a typical treatment regimen for a specified patient cohort. As mentioned above, the treatment outcome depends on not only the treatment plans but also other factors, therefore the treatment outcome of the same treatment regimen are often different for different patients. It is necessary to develop a unified framework to measure the treatment outcome by combining different sources of information in the EMRs data.

Our solution to the challenges and our main contribution in this paper are listed as follows:

- We first divide patient-level treatment records into subtreatments according to treatment periods/courses, where each subtreatment is a complex heterogeneous data set consisting of multiple doctor orders. Then, we propose effective similarity measures between subtreatments and between the patient-level treatment records respectively.
- A partial density peaks based clustering algorithm is proposed, which can efficiently cluster large amount of treatment records. For each treatment cluster, a dense *core area* is identified around the density peak to extract a semantically meaningful description for each cluster.
- To evaluate the treatment regimens, we first divide all patients into many patient cohorts according to their demographic information, diagnostic information, and treatment outcomes. The patients in the same cohort are thought to have similar physical preconditions and disease states. Then, we quantify the outcomes of different treatment regimens on the studied patient cohort, and pick out the best one as the recommended treatment regimen for new patients with similar physical preconditions and disease states.

In summary, in terms of applications, our method can help the medical decision-making and improve cure and treatment effective rates through personalized typical treatment regimen recommendation. In terms of theoretical contributions, the developed similarity measure and semantic density-based clustering for complex heterogeneous and longitudinal data is not hardwired with EMR data hence also applicable for other applications. Also, the unified framework for evaluating the treatment outcomes with control factors is a general framework for outcome evaluation in healthcare studies. The framework is also flexible to include more factors if available.

2. DEFINITIONS AND FORMALIZATION

Electronic Medical Records (EMRs) usually contain five categories of information of patients. They are demographic information, diagnostic information, laboratory indicators, doctor orders, and outcomes.

Definition 1 Demographic Information Demographic information is recorded when a patient visits a hospital, which includes the age, gender, address, race and ethnicity, education, and other information of a patient. These information plays an important role in clinical decisions, e.g., therapeutic regimen design and dosage selection. The demographic information of a patient can be formalized as

$$B = \{B^{Age}, B^{Gender}, B^{Address}, \cdots\}.$$

Definition 2 Diagnostic Information Diagnostic information is given by doctors. It consists of disease names and severity of the diseases. Considering a patient (especially aged patient) may suffer from multiple health problems. Diagnostic information can be represented by

$$D = \left\{\{D_1^{Name}, D_1^{Severity}\}, \{D_2^{Name}, D_2^{Severity}\}, \cdots\right\}.$$

Definition 3 Doctor Order An doctor order is a medical prescription, which is implemented by a physician or other qualified health care practitioner in the form of instructions that govern the plan of care for a patient. A doctor order can be represented as

$$Order = \{O^{DN}, O^{Delivery}, O^{Dose}, O^{Freq}, O^{StartT}, O^{EndT}\},$$

where O^{DN} represents the used drug name, $O^{Delivery}$ is the delivery route, which can be by "Intravenous injection" (IV), "Intramuscular" (IM), "Oral" (Per os, PO), and so on. O^{Dose} is the dosage each time, O^{Freq} indicates how many times per day. O^{StartT} and O^{EndT} provides the active time of the order. For example, a doctor order $\{Aspirin, PO, 100, 3, 4, 6\}$ means that the medicine Aspirin is delivered by oral route, 100mg each time (assume the unit is "mg"), three times per day, the order acts from 4-th day to 6-th day.

An order can be rewritten as

$$Order = \{O^{DN}, O^{Delivery}, O^{Dose}, O^{Freq}, t_1^O, t_2^O, \cdots, t_{d^O}^O\},$$

where $t_1^O = O^{StartT}$, $t_{d^O}^O = O^{EndT}$, d^O is the number of an order's lasting days, $d^O = O^{EndT} - O^{StartT} + 1$. For the simplicity of representation, we define

$$O = \{O^{DN}, O^{Delivery}, O^{Dose}, O^{Freq}\},$$

$$t^O = \{t_1^O, t_2^O, \cdots, t_{d^O}^O\},$$

then an order can be simply represented as

$$Order = \{O, t^O\}.$$

Definition 4 Treatment A treatment of a patient is a series of doctor orders related with the patient, which can be represented as

$$T = \left\{ \{t_1^T, OS_1\}, \cdots, \{t_{d^T}^T, OS_{d^T}\} \right\},$$

where t_1^T is the first treatment day, and $t_{d^T}^T$ is the last treatment day. The treatment lasting d^T days. OS_k represents the set of orders happened at k-th treatment day, which is $OS_k = \{O_{k,1}, \cdots, O_{k,m_k}\}$, m_k is the number of active orders in k-th treatment day t_k^T.

Given a series of doctor orders, we can obtain the treatment of the patient by

$$\{t_1^T, \cdots, t_{d^T}^T\} = unique\left(t^{O_1}, t^{O_2}, \cdots, t^{O_n}\right),$$

$$OS_k = \left\{ O_l \mid t^{O_l} \ni t_k, l = 1, 2, \ldots, n. \right\},$$

where $unique()$ is to eliminate the repeated elements in a set, n is the number of orders related with the patient.

Considering a treatment is often divided into different periods (called treatment courses) in clinical practice, a treatment can be rewritten as

$$T = \{OSP_1, OSP_2, \cdots, OSP_p\},$$

where a period is defined as $Period_q = [q^{Start}, q^{End}]$, then

$$OSP_q = OS_{q^{Start}} \,\tilde{\cup}\, \cdots \,\tilde{\cup}\, OS_{q^{End}}.$$

A tilde is added above the logic operation $\cup$, which means we not only get the union set of $OS_{q^{Start}}$ to $OS_{q^{End}}$ but also record the frequency of every O, two Os can be merged if and only if all the elements $O^{DN}, O^{Delivery}, O^{Dose}, O^{Freq}$ are identical. Therefore, OSP_q is a set of quintuples. We use O' to indicate a quintuple,

$$O' = \{O^{DN}, O^{Delivery}, O^{Dose}, O^{Freq}, O^{Times}\}.$$

The symbol prime is used to make the quintuple different from the O defined previously. Finally, a OSP_q can be simply represented by

$$OSP_q = \left\{ O'_{qg} \mid g = 1, \cdots, n_q. \right\},$$

where n_q is the number of O' quintuples.

In order to be easily understood, an example is given, assume a treatment duration is divided into five periods, the first 24 hours, 2-3 days, 4-7 days, 8-14 days, 15th day to the end, which is $T = \{OSP_1, OSP_2, \cdots, OSP_5\}$, every OSP_q contains a number of medicine quintuples, each quintuple corresponds to a doctor order that records the medicine name, delivery route, dosage, frequency per day, and repeating times during the q-th period.

Definition 5 Outcome Outcome is evaluated and presented by doctors when a patient leaves hospital. An outcome of a patient can be "cured", "improved", "ineffective", or "dead". We use R to represent the outcome of a patient in this paper.

Figure 1 illustrates the five categories of information in EMRs. Laboratory indicators are mainly used for judging the severity of disease or evaluating a patient's final outcome, which can be implied by diagnostic information or the outcome, so which are not included in the following model. A treatment of a patient is derived from all the doctor orders of the patient, which is the object studied in this paper.

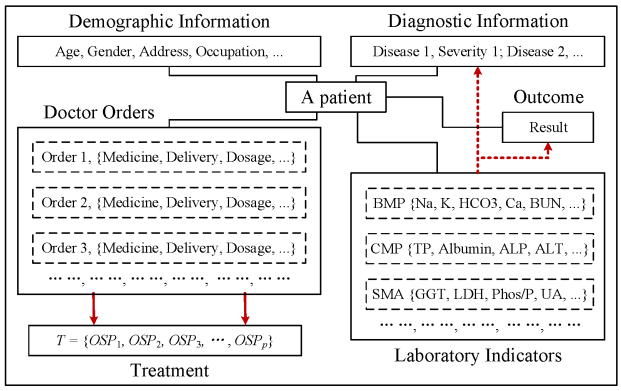


Figure 1: Illustration of five categories of information in EMRs

In the following model, the five categories of information are divided into three groups: demographic information and diagnosis information are used as preconditions, treatment consisted of doctor orders are the studied control variable, outcome is used as target.

3. TREATMENT REGIMEN IDENTIFICATION AND RECOMMENDATION

In this section, we introduce the detail of the proposed methods. Our work consists of four steps: 1) computing similarities between treatments, 2) clustering treatments, 3) extracting typical treatment regimens from treatment clusters, and 4) treatment regimen recommendation. A treatment defined in this paper is much more complex than previously studied ones, which poses non-trivial challenge to similarity measurement and typical treatment extraction. Large volume of EMRs requires efficient clustering algorithm. A lot of related factors are related with outcomes, which means the recommendation should take both preconditions (like age, gender, disease severity) and treatments into consideration. Therefore, we proposed novel methods in each step.

3.1 Similarity Measure for Treatments

Temporal data is usually defined as recorded information with time stamp. From this perspective, a time series is a set of of real values with time stamp, temporal events are a series of nominal terms with time stamp. In a medical treatment, the recorded information includes not only nominal terms like medicine name, delivery route, but also figures like dosage, frequency per day, and repeated times, so the recorded information in a medical treatment are heterogeneous. The time stamp is also more complex than previously studied ones as it records both starting and ending time. In this case, how to compute similarity between two treatments becomes a challenging problem.

According to section 2, a medical treatment can be divided into different periods and represented as

$$T = \{OSP_1, OSP_2, \cdots, OSP_p\}.$$

Therefore, the similarity of two treatments can be defined as weighted average of similarities in different periods,

$$s(i,j) = \sum_{q=1}^{p} w_q \bar{s}_q(i,j), \tag{1}$$

where p is the number of periods, $\bar{s}_q(i,j)$ is the similarity of OSP_{iq} and OSP_{jq}, which is the similarity of two treatments in *q-th* period.

Each OSP contains a number of quintuples, where a quintuple consists of medicine name, delivery route, dosage, frequency, and repeated times. In order to be easily understood, we present a toy example of OSP_{iq} and OSP_{jq} by Table 1 and 2.

Table 1: A toy example of OSP_{iq}

MedName	Delivery	Dosage	Freq.	Times
DrugA	PO	80	2	4
DrugB	PO	6	3	4
DrugC	IV	2	3	2

Table 2: A toy example of OSP_{jq}

MedName	Delivery	Dosage	Freq.	Times
DrugA	IV	120	1	2
DrugB	PO	8	3	4

In order to define similarity between OSP_{iq} and OSP_{jq}, we have to develop a method which can compute similarity between two such tables. As each table consists of some quintuples (each column is a quintuple), we should first define similarity between two quintuples. OSP_{iqg} is used to represent *g-th* quintuple in OSP_{iq}, and OSP_{jqh} is used to represent *h-th* quintuple in OSP_{jq}, then the similarity between OSP_{iqg} and OSP_{jqh} is defined as following:

Firstly, the similarity between OSP_{iqg} and OSP_{jqh} is determined by the Drug Names (we use DN for short), if the Drug Names of two quintuples are same, then delivery route, dosage, and frequency per day are considered in a further step; otherwise, the similarity of two quintuples is set 0. Therefore, the similarity of OSP_{iqg} and OSP_{jqh} contains a multiplying term $\delta\left(DN_{iqg}, DN_{jqh}\right)$, which equals 1 if DN_{iqg} and DN_{jqh} are the same, and equals 0 otherwise.

Secondly, the delivery (we use DE for short) should be taken into account. The similarity between two deliveries is also described by $\delta(\cdot,\cdot)$ function, which is 1 if two deliveries are the same, and equals 0 otherwise.

Lastly, dosage and frequency per day also has large impact on the treatment effect. We use Dosage-per-Day (DD) to describe the similarity in dosage of two quintuples. The DD is defined as $O^{Dose} \times O^{Freq}$. The similarity in Dose-per-Day (DD) is

$$1 - \frac{|DD_{iqg} - DD_{jqh}|}{\max(DD_{iqg}, DD_{jqh})} = \frac{\min(DD_{iqg}, DD_{jqh})}{\max(DD_{iqg}, DD_{jqh})}.$$

To sum up, similarity between OSP_{iqg} and OSP_{jqh} is finally defined as

$$s^e(OSP_{iqg}, OSP_{jqh}) = \frac{\delta\left(DN_{iqg}, DN_{jqh}\right)\left[\delta\left(DE_{iqg}, DE_{jqh}\right) + \frac{\min(DD_{iqg}, DD_{jqh})}{\max(DD_{iqg}, DD_{jqh})}\right]}{2}, \quad (2)$$

where the denominator 2 is to ensure the value of similarity drops in [0,1].

After getting similarities between two quintuples like OSP_{iqg} and OSP_{jqh}, the similarity between OSP_{iq} and OSP_{jq} is essentially a similarity between two complex sets. Different from previous problems, 1) the appearance times of elements in a set and 2) the similarities between elements, should be taken into account. For example, $C_1 = \{a, a, a, c, d, d\}$ and $C_2 = \{a, a, b, b\}$ are two such complex sets, the similarities between elements are given as $s(a,a) = 1.0$, $s(a,b) = 0.2$, $s(c,a) = 0.3$, $s(c,b) = 0.5$, $s(d,a) = 0.6$, $s(d,b) = 0.2$. Then, how to define similarity of C_1 and C_2? To the best of our knowledge, this problem has not been studied ever before. In this paper, we propose a general method for computing similarity between two such complex sets.

Without loss of generality, we still define

$$s^{set}(C_1, C_2) = \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|} = \frac{|C_1 \cap C_2|}{|C_1| + |C_2| - |C_1 \cap C_2|}. \quad (3)$$

However, the $|C_1 \cap C_2|$ is redefined in this paper, which is

$$|C_1 \cap C_2| = \sum_{gh} s^e_{gh} a_{gh}, \quad (4)$$

where s^e_{gh} represents similarity between two elements, $\mathbf{A} = \{a_{gh}\}$ is obtained by solving

$$\begin{aligned} \arg\max_{\mathbf{A}} \quad & z = \sum_{gh} a_{gh} s^e_{gh} \\ s.t. \quad & \sum_{g} a_{gh} \le f^1_h, \\ & \sum_{h} a_{gh} \le f^2_g, \\ & a_{gh} \ge 0. \end{aligned} \quad (5)$$

f^1_h is the appearance frequency of h-th element in set C_1, and f^2_g is that of g-th element in set C_2. The above formula indicates that $\mathbf{A}$ is obtained by allocating the frequencies of elements into a two dimensional table with the goal of maximizing the same part of two sets, therefore $\mathbf{A}$ is called allocation matrix.

Algorithm 1 $SetAllo(\mathbf{S}^e, \mathbf{f}^1, \mathbf{f}^2)$

Input: $\mathbf{S}^e$,$\mathbf{f}^1$,$\mathbf{f}^2$
Output: $\mathbf{A}$;
1: $\mathbf{A} = \mathbf{0}$,$\mathbf{L} = \mathbf{1}$;
2: **while** $\exists\ l_{ij} = 1$ **do**;
3: $(g,h) = \arg\max_{(u,v), l_{uv}=1} s^e_{uv}$;
4: **if** $f^1_h \le f^2_g$
5: **then do** $a_{gh} = f^1_h$, $l_{ih} = 0$ for any i;
6: **else do** $a_{gh} = f^2_g$, $l_{gj} = 0$ for any j;
7: $f^1_h = f^1_h - a_{gh}$, $f^2_g = f^2_g - a_{gh}$;
8: **end**

Algorithm 1 presents a greedy algorithm we proposed for solving optimization problem (5), which is guaranteed to get optimal solution.

Combing the above definition and equation (2) ∼(5), similarity between OSP_{iq} and OSP_{jq} is finally defined as

$$\bar{s}_q(i,j) = \frac{\sum_g \sum_h s^e(OSP_{iqg}, OSP_{jqh}) a_{ijqgh}}{\sum_g f_{iqg} + \sum_h f_{jqh} - \sum_g \sum_h s^e(OSP_{iqg}, OSP_{jqh}) a_{ijqgh}}, \quad (6)$$

where f_{iqg} is the repetition times of g-th order in q-th period of treatment i, $\{a_{ijqgh}\}$ is the allocation matrix for the computing of $|OSP_{iq} \cap OSP_{jq}|$.

It's not difficult to prove that the similarity $\bar{s}_q(i,j)$ has the following properties:
1) The value of $\bar{s}_q(i,j)$ is from 0 to 1;
2) Symmetry. For any i and j, $\bar{s}_q(i,j) = \bar{s}_q(j,i)$;
3) Self-similarity. $\bar{s}_q(i,j) = 1$, if and only if $OSP_{iq} = OSP_{jq}$,

the equal sign indicates the two sets contains same elements and the frequencies of elements are also the same.

It can be easily inferred that similarity of two treatments $s(i,j)$ also holds the three properties. we only present the proof of property 3 here. $s(i,j) = 1$ means $\bar{s}_q(i,j) = 1$ for each q. That is $OSP_{iq} = OSP_{jq}$ for each q. If two treatments they have same OSP in each period, then the two treatments are same. Therefore, $s(i,j) = 1$ if and only if $T_i = T_j$.

3.2 Clustering Treatments

The goal of this paper is to extract effective typical treatment regimens from EMRs. After getting similarities between treatments, we should first divide all treatments into several clusters, and then extract a typical treatment regimen from each cluster. Clustering is a technique of partitioning a set of objects into multiple groups (called clusters) so that objects in the same cluster are more similar to each other than to those in other clusters [5]. In this section, a Map Reduce enhanced Density Peaks based Clustering (MRDPC) is proposed to accomplish this task.

Our method derives from a recently proposed exemplar-based clustering algorithm [18], which is called Density Peaks based Clustering (DPC). The advantage of DPC is that it can discover clusters with complex shapes, while traditional exemplar-based clustering algorithms can only find spherical clusters. In DPC, two indicators are computed for each object: 1) local density ρ and 2) minimum distance (or maximum similarity) between the object and any other object with higher local density γ, where ρ is defined as

$$\rho_i = \sum_j \chi(s_{ij} - s_c), \tag{7}$$

where $\chi(x) = 1$ if $x > 0$ and $\chi(x) = 0$ otherwise, s_c is a cutoff similarity. The meaning of ρ_i is to count the number of objects in object i's s_c-neighborhood.

The second indicator γ is defined as

$$\gamma_i = \max_{j:\rho_j > \rho_i} (s_{ij}), \tag{8}$$

Objects with larger ρ and lower γ values are viewed as exemplars. The intuition is that exemplars are the points that they have the highest density in a relative large range. After identifying exemplars, clustering result can be obtained according to exemplars.

Besides the advantage of discovering clusters with complex shapes, DPC is also an efficient clustering algorithm. Even so, DPC can not be directly used in this task. As we mentioned earlier, one of the most well known challenges of mining EMRs is the large volume of patients it records. Image there are 10,000 patients in EMRs, the scale of similarity matrix can reach 100 million. Such similarity matrix is difficult to store and deal with. Additionally, the computation of similarity matrix will also takes a lot of time and space. In this case, a clustering algorithm which can work with incomplete similarity information becomes very important. The proposed MRDPC is such a clustering algorithm. In MRDPC, the total N patients are first randomly divided into m parts, DPC is implemented on each part to get k potential exemplars; then a partial similarity matrix is obtained by computing similarities of selected potential exemplars and all objects, the scale of which is $mk \times N$; Partial DPC (PDPC) is used to determine K final exemplars according to the partial similarity matrix.

Algorithm 2 $MRDPC(\mathbf{T},m,k,K)$

Input: $\mathbf{T} = \{T_i\}$,m,k,K
Output: $\mathbf{c}$;

1: Divide $\mathbf{T}$ into m parts randomly;
2: **for** g **from** 1 **to** m **do**
3: compute $\mathbf{S}^g$ for treatments in $\mathbf{T}^g$;
4: **for** each $i \in \mathbf{T}^g$ **do**
5: $\rho_i = \sum_j \chi(s_{ij} - s_c)$, $\gamma_i = \max_{j:\rho_j > \rho_i} (s_{ij})$;
6: $\eta_i = \rho_i / \gamma_i$;
7: **end**
8: $PE^g = \{i | \eta_i \geq \epsilon\}$, where ϵ is the k-th largest η value;
9: **end**
10: $PE = \{PE^1, PE^2, \cdots, PE^m\}$;
11: compute $\mathbf{S'} = \{s'_{ij}\}$, where $i \in PE$, $j = 1, 2, \cdots, N$;
12: **for** $i \in PE$ **do**
13: $\rho'_i = \sum_j \chi\left(s'_{ij} - s'_c\right)$, $\gamma'_i = \max_{j \in PE, \rho'_j > \rho'_i,} \left(s'_{ij}\right)$;
14: $\eta'_i = \rho'_i / \gamma'_i$;
15: **end**
16: $E = \{j | \eta'_j \geq \epsilon'\}$, where ϵ' is the K-th largest η' value in PE;
17: **for** each $i \notin E$
18: **if** $i \in PE$ **do**
19: $c(i) = c(\bar{i})$, where $\bar{i} = \arg \max_{j \in PE, \rho'_j > \rho'_i,} \left(s'_{ij}\right)$;
20: **else** $i \notin PE$, $i \in \mathbf{T}^g$ **do**
21: $c(i) = c\left(c(\bar{i})\right)$, where $\bar{i} = \arg \max_{j \in \mathbf{T}^g, \rho'_j > \rho'_i} \left(s'_{ij}\right)$;
22: **end**;

Algorithm 2 presents the proposed MRDPC method. N treatments are divided in step 1, potential exemplars are obtained in step 10. Then, step 11 computes the partial similarity matrix, and step 12∼16 identify final exemplars from potential exemplar set. Finally, step 17∼22 assign a cluster label for each treatment.

We also define the popularity of a treatment cluster as

$$Support(TC_i) = \frac{\sum_j \delta(c(j), E(i))}{N}. \tag{9}$$

where $\delta(x,y) = 1$ if $x = y$, $\delta(x,y) = 0$, otherwise; N is the number of treatments studied.

3.3 Extracting Typical Treatment Regimens

In most of the previous applications of exemplar-based clustering, an exemplar can be directly used to describe the corresponding cluster. However, a treatment can vary in many different directions as a complex temporal and heterogeneous data set, which makes a single object (even the object is an exemplar) can not well describe the cluster it belongs to. In this case, we define the core area of a treatment cluster and extract a semantic description of each treatment cluster by its dense core.

The dense core of a cluster is constructed by k-nearest neighbors of its exemplar. Therefore, a dense core can be represented by a set

$$Core_i = \{T_j | s_{j,e_i} \geq \tau_i\}, \tag{10}$$

where e_i is the exemplar of i-th cluster, τ_i is the similarity of the exemplar with its k-th nearest neighbor.

In order to extract typical treatment regimen from a dense core, we define the support of a drug in a specified period

(e.g., q) of a treatment regimen (e.g., i), which is

$$Support_{iq}(Drug) = \frac{\sum_{j,T_j \in Core_i} \lambda(Drug, OSP_{jq})}{|Core_i|}, \quad (11)$$

where $\lambda(Drug, OSP) = 1$ if $Drug$ is used in OSP ($Drug \in OSP$), $\lambda(Drug, OSP) = 0$ otherwise.

According to formula (11), we can select the drugs used in a specified period of a treatment regimen, which is

$$DrugSet_{iq} = \{Drug | Support_{iq}(Drug) > \sigma\}, \quad (12)$$

where σ is a threshold defined aforehand.

After knowing the main medicines used in a specified period of a treatment regimen, we make clear the usages of drugs, e.g., deliveries, dosages, lasting days in the period. Therefore, we compute support of Dosage and Administration (DA) for every drug,

$$Support_{iqk}(DA) = \frac{\sum_{DA_k, Drug \in DrugSet} \lambda(DA_k, OSP_{jq})}{|Core_i|}, \quad (13)$$

where DA is a triple consists of delivery route, dosage and lasting days of a order.

Based on the clustering result obtained by the method introduced in the previous subsection, formula (10) defines a dense core for each treatment cluster. Then, formula (11)$\sim$(13) extracts a typical treatment regimen from each dense core. A typical treatment regimen includes the names of medicines used in a specified period, the dosages, the delivery roues, and lasting how many days.

3.4 Evaluating Typical Treatment Regimen

One of the most challenging problem in automatic treatment regimen recommendation is how to evaluate a treatment regimen, which is because that 1) the care outcome is actually affected by a lot of factors, 2) for different patient cohorts, the most effective typical treatment regimen may be different.

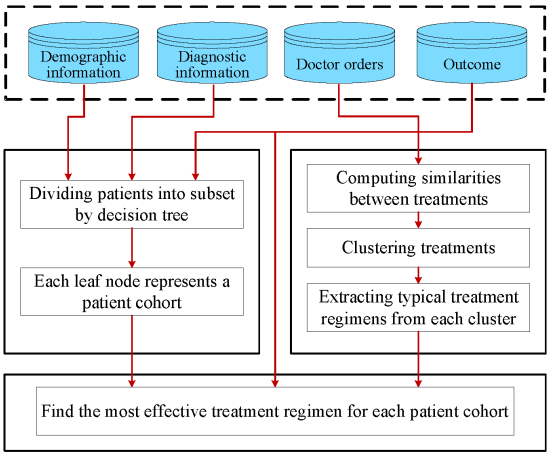


Figure 2: A framework for treatment regimen recommendation.

In this paper, we propose a general framework to overcome this problem. Figure 2 presents the framework. Five categories of information are recorded in EMRs, the laboratory indicators are mainly used to judge disease severity or help to evaluate outcome, which are implied in diagnostic information and outcome. So four kinds of medical information are used in our model.

Doctor orders are mainly used to generate treatments. Then similarities between treatments are computed. Based on the similarities, MRDPC is used to cluster the treatments. After clustering, a typical treatment regimen is extracted from each treatment cluster.

We also divide patients into different groups according to demographic information, diagnostic information and outcomes, which is realized by a decision tree model. The patients in a same leaf node is defined as a patient cohort, which means these patients should have had the same outcome with respect to preconditions like age, gender, ..., and disease severity.

For a specified patient cohort, we observe how many typical treatment regimens have been used on the patients in this cohort, and then figure out which treatment regimen can result in the highest effective rate. This framework combines a lot of information recorded in EMRs together, which can evaluate the effectiveness of a treatment regimen on a patient cohort comprehensively.

4. EMPIRICAL STUDY

In this section, we test our methods by experiments on real-world EMRs data. We first present a brief description of the studied data set, then extract typical treatment regimens from the large-scale treatment records. A comprehensive validation of our results is given with the data of patients from multiple hospitals.

4.1 Experimental Data

The EMRs data used in this paper are collected from Hospital Information Systems (HIS) of 14 *Grade Three Class A* (G3CA) hospitals, where G3CA is a certification for the best general public hospitals in China. The 14 hospitals locate in seven cities: Beijing and Shijiazhuang in north China, Shenzhen in south China, Jinan in east China, Changchun in northeast China, Fuzhou in southeast of China, and Xi'an in northwest of China.

The methods proposed in this paper can be used to discover and recommend treatment regimens for various diseases. To illustrate and test our methods, we focus on the patients with cerebral infarction disease, which is one of the most common diseases in China today. For a cerebral infarction patient, five kinds of information are recorded: 1) demographic information, 2) diagnostic information, 3) doctor orders, 4) clinical laboratory indicators, and 5) treatment outcome. The possible treatment outcome of a cerebral patient can be: "cured", "improved", "ineffective", or "dead". Different kinds of information are associated together by unique patient IDs.

After collecting the EMRs, clinical doctors in China Academy of Chinese Medical Sciences preprocessed the data, they removed the erroneous records and unified the diagnostic and medicine names. Finally, we have demographic information for $27,678$ cerebral infarction patients, and $28,659$ unique patient IDs are found with doctor orders. $27,427$ patients of them have both demographic and doctor order information. In the following, we extract the typical treatment regimens of cerebral infarction from doctor orders of $28,659$ patients. The total number of doctor orders is $1,007,057$.

4.2 Extracting Typical Treatment Regimens

In the doctor orders, 1,090 medicines have been used. However, many of them are used to treat other diseases. For example, most of the cerebral infarction patients are aged people and many of them are suffering from multiple diseases at the same time. Therefore, clinical doctors helped us to select 138 medicines that are most relevant to cerebral infarction to better validate our results. We have 363,674 doctor orders containing the selected medicines, nearly 13 doctor orders per patient.

As discussed in Section 2, we construct the treatment record for each patient, and each treatment record is further divided into four periods. Indeed, for cerebral infarction, the treatment in the first two weeks is most responsible for the treatment result, especially the first 24 hours. Therefore, the four treatment periods are: the first 24 hours, 2-3 days, 4-7 days, and 8-14 days. The weights of different periods in computing similarity between treatment records are $w = (0.4\ 0.2\ 0.2\ 0.2)$ in Equation (1).

All the computed similarities can form a matrix of size $N \times N$, where N is the number of patients with treatment records. As mentioned above, we have $N = 28,659$, so the scale of similarity matrix is quite large. To further improve the efficiency of treatment clustering for the treatment regimen discovery, we use the Map-Reduce enhanced Density Peaks based Clustering method (MRDPC) proposed in Section 3.2. First, all the treatment records are randomly divided into 10 parts, and the Density Peaks based Clustering (DPC) is used to select 100 potential exemplars from each part. Second, the similarities between 10×100 potential exemplars and all the original treatment records are computed, and the Partial Density Peaks based Clustering (PDPC) is used to find the final exemplar treatments.

Our method divided the N treatment records into four clusters, where each treatment cluster corresponds to a typical treatment regimen. In order to extract a semantic description of each typical treatment regimen, we identify the *dense core* in each treatment cluster according to Equation (10) and the frequently used medicines in the doctor orders by Equation (11).

Figure 3 illustrates the four extracted typical treatment regimens for cerebral infarction. The four typical treatment regimens are different from each other with different frequencies of 9 most popular medicines for this disease. In particular, Typical Treatment Regimen 1 (TTR1) is the most widely used typical treatment regimen with support of 58.33%, where two medicines (Aspirin and Xueshuantong) are used in all of the four treatment periods. Indeed, 100% of patients in the dense core of TTR1 used Aspirin, and more than 80% of them used Xueshuantong. The other medicines are not frequently used in TTR1.

Typical Treatment Regimen 3 (TTR3) is the second popular regimen with support of 24.82%. In comparison with TTR1, Aspirin is still widely used in TTR3, but Xueshuantong is replaced by Lumbrokinase and Xuesaitong, which indicates Xueshuantong may have the same therapeutic function with Lumbrokinase or Xuesaitong. Though Aspirin is also used in Typical Treatment Regimen 2 (TTR2), its frequency is much less than in TTR1 and TTR3. The most popularly used medicines in TTR2 are Shuxuetong, Ozagrel, and Cinepazide. The usages of four medicines in different periods are also different. As an emergency medicine, the usage rate of Ozagrel in the first 24 hours is the highest, then it declines in the following periods. In Typical Treatment Regimen 4 (TTR4), many medicines are used, especially the usage rates of Edaravone and Alprostadil are higher than the three other typical treatment regimens, but the other six medicines except Aspirin are not used as popular as they are in other treatment regimens.

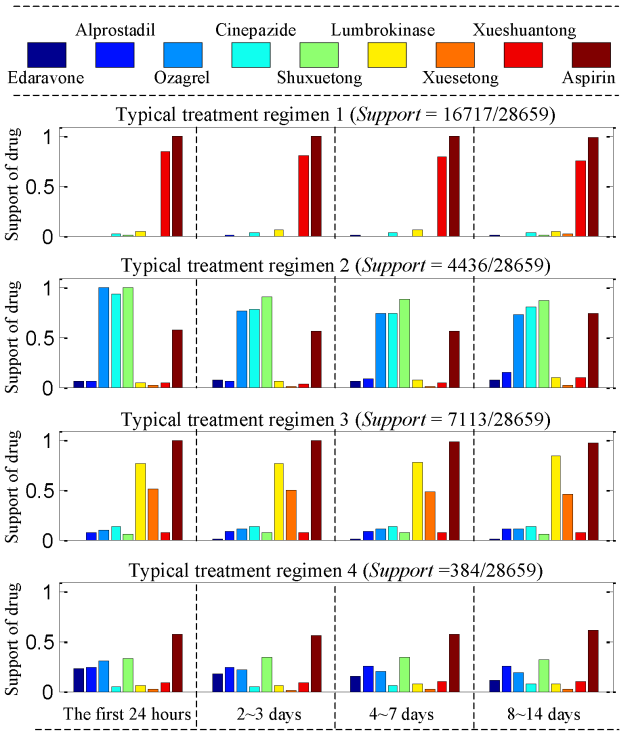


Figure 3: Four typical treatment regimens extracted from EMRs.

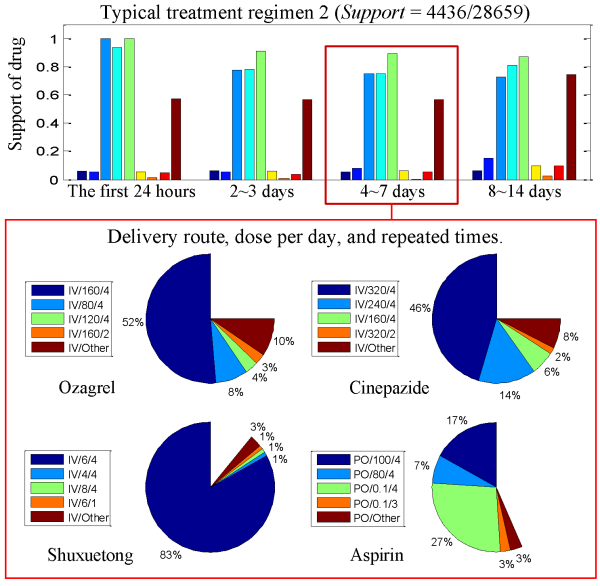


Figure 4: An example of an extracted treatment regimen.

Figure 3 compares the typical treatment regimens by the frequencies of medicines. However, a typical treatment regimen extracted by our method consists of much more information than medicine frequencies. Indeed, the daily dosages, deliveries, and active days are all considered in computing similarity between treatment records (see Equation 2). To better show how these medicines are used, we compute Dosage and Administration support for every medicine according to Equation (13), as shown in Figure 4. We take the third period (4∼7 days) of TTR2 for example. The first pie shows the different usage manners of Ozagrel, where "IV/160/4" means the delivery route is intravenous injection (IV), the daily dosage is 160 units, for four days during this period. This usage manner takes 52% support, and is the most common one. The other usage manners of Ozagrel vary mainly in dosage, which can be 80 or 120 units. The usage manners of Cinepazide are almost the same as that of Ozagrel. The most common usage manner of Shuxuetong is "IV/6/4", which takes 83% support. Aspirin is delivered by "PO" (Per os) instead of "IV".

4.3 Treatment Regimen Recommendation

In Section 3.4, a unified framework is proposed to evaluate the effectiveness of the discovered treatment regimens. According to the proposed framework, we first divide the 27,678 cerebral infarction patients into homogeneous cohorts by a decision tree. In constructing the decision tree, patients' demographic information and disease severity are used as independent variables and the treatment outcome is used as dependent variable.

Intuitively, patients in the same leaf node of the constructed decision tree should have had the same treatment outcome with respect to their physical condition and disease severity. Therefore, for each node, we identify and recommend the treatment regimens which can produce the best outcome based on historical records.

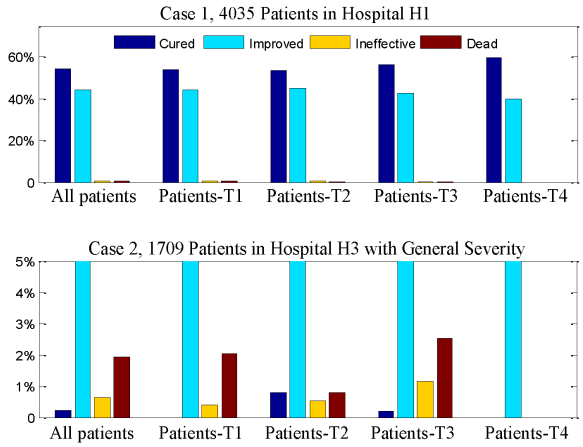


Figure 5: Recommend treatment regimens for two patient cohorts

In our results, there are 36 leaf nodes in the constructed decision tree. Figure 5 presents a visual comparison of four treatments on two cases (leaf node 2 and 17), while Table 3 provides numerical estimates about the cure rate, improved rate, ineffective rate and dead rate obtained by four treatment regimens. Specifically, Case 1 includes 4,035 patients in Hospital "H1", and Case 2 contains 1,709 patients in hospital "H3" with "General" severity (relative to "Emergency" and "Dangerous" severity). In the first case, most of the patients are cured and improved. TTR4 can result in the highest cure rate and lowest ineffective and dead rate. However, only 0.37% of patients in this group accepted TTR4. Though TTR4 may be effective for the patients under this leaf node, it is not robustly validated. TTR3 with popularity of 25.97%, which can produce higher cure rate and lower ineffective and dead rate than TTR1 and TTR2, is the most effective treatment regimen found by our method for the patients in this leaf node.

In Case 2, more than 97% of the patients are diagnosed as "improved". Therefore, we mainly compare different treatment regimens by the cure, ineffective, and dead rate. Similarly, TTR4 is not taken into account because of the low popularity. In comparison with Case 1, TTR3 is not the recommended treatment regimen any longer. Indeed, TTR3 is the worst treatment with highest ineffective and dead rate for patients in this case. TTR2 with highest cure rate (two times more than the follower) and lowest dead rate (less than half of the follower) is the most suitable treatment regimen for this patient cohort.

Table 3: Treatment effects on two patient cohorts.

	Cured	*Impro.*	*Ineff.*	*Dead*	*Popul.*
All Patis	54.62	44.34	0.59	0.45	100.00
Patis-T1	54.24	44.55	0.69	0.52	57.25
Patis-T2	53.57	45.32	0.79	0.32	15.64
Patis-T3	**56.49**	**42.94**	**0.29**	**0.29**	**25.97**
Patis-T4	60.00	40.00	0.00	0.00	0.37
All Patis	0.23	97.19	0.64	1.93	100.00
Patis-T1	0.00	97.58	0.38	2.04	45.87
Patis-T2	**0.79**	**97.91**	**0.52**	**0.79**	**22.35**
Patis-T3	*0.19*	*96.13*	*1.16*	*2.51*	*30.25*
Patis-T4	0.00	100.00	0.00	0.00	0.41

The above results show that, our method can recommend an effective treatment regimen to a specified patient cohort for better treatment outcomes. In addition, by comparing the two cases above, we can conclude that, 1) for patients with different physical conditions and disease severities, the most effective treatment regimens are different; 2) the most widely used treatment regimen in clinical, like TTR1, may be not the actually best one; 3) some rarely used treatment regimen may result in even better outcome. These three observations provide important clues for clinical doctors to develop new treatment regimens or optimize the present ones.

4.4 Overall Treatment Evaluation

The previous experiments have shown that our method can extract treatment regimens from EMRs automatically, and recommend the most effective treatment regimen to a specified patient cohort. In the following, we estimate the recommendation effects on all the 27,678 patients from 14 hospitals, in terms of total cure rate and total effective rate.

The total cure rate and total effective rate is defined as

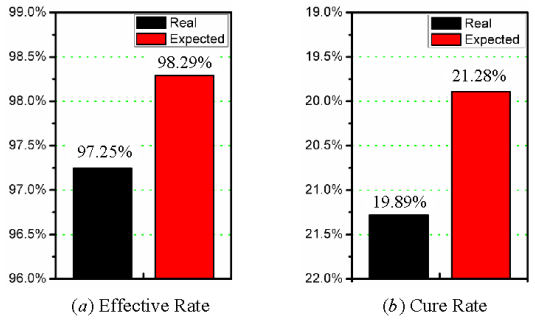
$$r^c = \frac{N^c}{N}, r^e = \frac{N^c + N^{imp}}{N},$$

where N^c and N^{imp} are numbers of patients with cured outcome and improved outcome respectively, N is the number of total patients.

The 27678 patients are divided into 36 patient cohorts according to physical condition and disease severity. We implement our methods on all patient cohorts to find the most effective treatment regimen for each patient cohort. The overall effective rate and cure rate are computed as

$$\hat{r}^c = \frac{\sum_i \hat{r}_i^c N_i}{N}, \hat{r}^e = \frac{\sum_i (\hat{r}_i^c + \hat{r}_i^{imp}) N_i}{N},$$

where $\hat{r}_i^c$ and $\hat{r}_i^{imp}$ are expected cure rate and improvement rate of *i-th* patient cohort, N_i is the number of patients in this cohort. The expected cure and improvement rate of a patient cohort are the cure and improvement rate of the patient cohort with the best treatment regimen respectively, where the best treatment regimen for a patient cohort is identified by method shown in Section 4.3.



(*a*) Effective Rate (*b*) Cure Rate

Figure 6: Our method can help improve effective rate and cure rate.

Figure 6(*a*) demonstrates that, the effective rate of the 27678 patients is promising to be improved from 97.25% to 98.29% by adopting our recommended regimen, while Figure 6(*b*) illustrates that the cure rate is promising to be improved from 19.89% to 21.28%. By this experiment, we can conclude that our method can not only extract typical treatment regimens from large-scale EMRs automatically, but also helpful for improving cure and effective rate.

5. RELATED WORK

In this section, we review the previous work in the literature related with our work. We will also explain the differences between our methods and the previous ones.

Temporal data mining has become an increasingly hot topic [11, 14]. The key challenge of temporal data mining is to represent the temporal data with flexible and informative structures, which enable easy computation of the similarity between temporal sequences, as well as clustering and classification tasks with the sequential data. For continuous time series data, a lot of representation methods and similarity measurement methods have been developed [1, 2, 10, 22]. For discrete event sequence data, there have also been recent works with diverse applications [12, 13, 15, 20]. However, a treatment record in this paper is essentially a sequence of doctor orders, which is much more complex than time series and simple event sequence. Therefore, the previous methods are not directly applicable. In this paper, to discover patterns from large-scale treatment records, we proposed novel representation and similarity measurement methods for sequential and multifaceted event sets.

Based on the similarities between treatments, we clustered all treatments into several groups and extracted a typical treatment regimen from each treatment cluster. Clustering has been studied in data mining community for many years, and various of clustering algorithms have been developed [7]. Among them, a popular category suitable for our task of treatment regimen discovery is the exemplar-based clustering [3]. The exemplar-based clustering algorithm first selects a number of exemplars and then assigns the remaining objects to their nearest exemplars. However, a disadvantage of the exemplar-based clustering is that it can only find spherical clusters. In this paper, due to the temporal and heterogeneous nature of the treatment records, a treatment cluster can be very complicated. Density-Peaks-based Clustering (DPC) is a recently proposed exemplar-based clustering algorithm [18], which can discover clusters with complex shapes. Therefore, we derived our clustering method by further extending the DPC algorithm. In particular, although the original DPC is designed to be efficient, it cannot be directly used if the data size (like number of treatment records in this paper) is truly large. Therefore, a Map-Reduce enhanced Density Peaks based Clustering (MRDPC) method is presented in this paper, which can efficiently cluster large-volume treatments with huge and incomplete similarity matrix. To the best of our knowledge, the MRDPC method has never been discussed in previous literature.

In addition, most of the present clustering algorithms in the literature focus on solely dividing the objects into homogeneous groups, rare work has been done to extract the semantically meaningful descriptions of the identified clusters. Indeed, such semantic extraction from clusters is usually a difficult task, especially for sequential and unstructured data. In this paper, a treatment is much more complex than the objects (or data points) studied in traditional clustering problems. To address this challenge, we extract the semantics by constructing a dense core in each exemplar-based cluster. The most similar work to ours is the dense subgraph discovery [9, 17]. However, their work is mainly for discovering interesting dense regions, while our work aims at extracting semantic descriptions of clusters. In general sense, frequent pattern mining [4] is also related with our work, but it is not applicable in this study because the treatment records vary in many dimensions, leading to no frequent patterns with significant pattern support.

In terms of applications, a lot of recent work has been done in mining the various kinds of EMRs data for actionable insights to improve the quality of healthcare delivery. For example, Zhou et al. [21] proposed PACIFIER method to infer phenotypic pattern from EMRs; Hirano and Tsumoto [6] used occurrence and transition frequency to discover typical order sequences; Liu et al. [12] developed a method to identify most significant and interpretable graphical feature from longitudinal EMRs; Somanchi et al. [19] studied early prediction of cardiac arrest based on demographic information, hospitalization history, vitals and laboratory measurements in patient-level EMRs. However, most of the previous research focused only on part of information recorded by EMRs. In this paper, we would extract typical treatment regimens from treatment records and recommend treatment regimens to new patients according to their demographic as well as diagnostic information. Our work can be deemed a benchmark framework for future healthcare data mining research to integrate all the related factors in EMRs and optimize the treatment effectiveness based on comprehensive evaluation of the typical treatment options.

6. CONCLUSION

In this paper, we investigated how to identify the typical treatment regimens from large-scale treatment records and how to find the most effective treatment regimens for patients. Specifically, we developed an efficient semantic clustering algorithm, based on a new method to measure the similarities between treatment records. The new similarity measurement and the semantic clustering are applicable for general complex heterogeneous and longitudinal data. Applied on large-scale treatment records, we were able to extract the treatment clusters as the typical treatment regimens with semantically meaningful descriptions. Moreover, we designed a unified framework to evaluate the effectiveness of the identified treatment regimens. This framework can recommend the most effective treatment regimens to new patients according to their demographic information and disease severities. Finally, we validated our approach on real-world EMRs data of patients collected from multiple hospitals. Experimental results show that our method can improve the treatment cure rate and effective rate. To the best of our knowledge, this work may be the first step towards the automatic development of treatment regimens and treatment recommendations.

7. ACKNOWLEDGMENTS

This research was partially supported by National Natural Science Foundation of China (71329201, 71171030, and 71421001). Also, it was supported in part by the Rutgers 2015 Chancellor's Seed Grant Program.

References

[1] Iyad Batal, Dmitriy Fradkin, James Harrison, Fabian Moerchen, and Milos Hauskrecht. Mining recent temporal patterns for event detection in multivariate time series data. In *KDD*, 2012.

[2] Gustavo E.A.P.A. Batista, Xiaoyue Wang, and Eamonn J. Keogh. A complexity-invariant distance measure for time series. In *SDM*, 2011.

[3] Brendan J. Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, 315 (5814):972–976, 2007.

[4] Jiawei Han, Hong Cheng, Dong Xin, and Xifeng Yan. Frequent pattern mining: current status and future directions. *Data Min. Knowl. Discov.*, 15(1):55–86, 2007.

[5] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., 3rd edition, 2011.

[6] Shoji Hirano and Shusaku Tsumoto. Mining typical order sequences from ehr for building clinical pathways. In *PAKDD*, 2014.

[7] Anil K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(3):264–323, 2010.

[8] Peter B Jensen, Lars Juhl Jensen, and SÃÿren Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews. Genetics*, 13(6):395–405, 2012.

[9] Victor E. Lee, Ning Ruan, Ruoming Jin, and Charu C. Aggarwal. A survey of algorithms for dense subgraph discovery. In *Managing and Mining Graph Data*. 2010.

[10] T. Warren Liao. Clustering of time series data: a survey. *Pattern Recognition*, 38(11):1857 – 1874, 2005.

[11] Weiqiang Lin, Mehmet A. Orgun, and Graham J. Williams. An overview of temporal data mining. In *Proceedings of the 1st Australian Data Mining Workshop*, 2002.

[12] Chuanren Liu, Fei Wang, Jianying Hu, and Hui Xiong. Temporal phenotyping from longitudinal electronic health records: A graph based framework. In *KDD*, 2015.

[13] Chuanren Liu, Kai Zhang, Hui Xiong, Guofei Jiang, and Qiang Yang. Temporal skeletonization on sequential data: patterns, categorization, and visualization. *Knowledge and Data Engineering, IEEE Transactions on*, 28(1):211–223, 2016.

[14] Theophano Mitsa. *Temporal Data Mining*. Chapman & Hall/CRC, 1st edition, 2010.

[15] Wei Peng, Charles Perng, Tao Li, and Haixun Wang. Event summarization for system management. In *KDD*, 2007.

[16] Adam Perer, Fei Wang, and Jianying Hu. Mining and exploring care pathways from electronic medical records with visual analytics. *Journal of Biomedical Informatics*, 29(3):241–288, 2015.

[17] Lu Qin, Rong-Hua Li, Lijun Chang, and Chengqi Zhang. Locally densest subgraph discovery. In *KDD*, 2015.

[18] Alex Rodriguez and Alessandro Laio. Clustering by fast search and find of density peaks. *Science*, 344 (6191):1492–1496, 2014.

[19] Sriram Somanchi, Samrachana Adhikari, Allen Lin, Elena Eneva, and Rayid Ghani. Early prediction of cardiac arrest (code blue) using electronic medical records. In *KDD*, 2015.

[20] Jingyuan Yang, Chuanren Liu, Mingfei Teng, Hui Xiong, March Liao, and Vivian Zhu. Exploiting temporal and social factors for b2b marketing campaign recommendations. In *ICDM*, pages 499–508. IEEE, 2015.

[21] Jiayu Zhou, Fei Wang, Jianying Hu, and Jieping Ye. From micro to macro: Data driven phenotyping by densification of longitudinal electronic medical records. In *KDD*, 2014.

[22] Hengshu Zhu, Chuanren Liu, Yong Ge, Hui Xiong, and Enhong Chen. Popularity modeling for mobile apps: A sequential approach. *IEEE Trans Cybern*, 45 (7):1303–1314, 2015.