IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received April 27, 2019, accepted August 17, 2019, date of publication August 27, 2019, date of current version September 16, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2937892

Discovering Medication Patterns for High-Complexity Drug-Using Diseases Through Electronic Medical Records

HUIQUN HUANG[1], XIAOPU SHANG[2], (Member IEEE), HONGMEI ZHAO[2,3], NAN WU[3], WEIZI LI[4], YUAN XU[2], YANG ZHOU[2], AND LEI FU[5]

[1]School of Software Engineering, Beijing Jiaotong University, Beijing 100044, China
[2]School of Economics and Management, Beijing Jiaotong University, Beijing 100044, China
[3]Peking University People's Hospital, Beijing 100044, China
[4]Informatics Research Center, University of Reading, Berkshire RG6 6AH, U.K.
[5]Chinese PLA General Hospital, Beijing 100044, China

Corresponding author: Xiaopu Shang (sxp@bjtu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61702023, in part by the Humanities and Social Science Foundation of Ministry of Education of China under Grant 17YJC870015, and in part by the Beijing Natural Science Foundation under Grant 7192107.

ABSTRACT An Electronic Medical Record (EMR) is a professional document that contains all data generated during the treatment process. The EMR can utilize various data formats, such as numerical data, text, and images. Mining the information and knowledge hidden in the huge amount of EMR data is an essential requirement for clinical decision support, such as clinical pathway formulation and evidence-based medical research. In this paper, we propose a machine-learning-based framework to mine the hidden medication patterns in EMR text. The framework systematically integrates the Jaccard similarity evaluation, spectral clustering, the modified Latent Dirichlet Allocation and cross-matching among multiple features to find the residuals that describe additional knowledge and clusters hidden in multiple perspectives of highly complex medication patterns. These methods work together, step by step to reveal the underlying medication pattern. We evaluated the method by using real data from EMR text (patients with cirrhotic ascites) from a large hospital in China. The proposed framework outperforms other approaches for medication pattern discovery, especially for this disease with subtle medication treatment variances. The results also revealed little overlap among the discovered patterns; thus, the distinct features of each pattern are well studied through the proposed framework.

INDEX TERMS Electronic medical record (EMR), medication pattern, discovery, machine learning, high-complexity drug-use pattern.

I. INTRODUCTION

Evidence-based medicine is recognized as an imperative approach to optimize decision-making in medical practice [1]. In the past, the evidence mainly came from well-designed, well-conducted clinical trials and the validated personal experiences of physicians. The most reliable evidence-based medicine is based on randomized controlled trials designed for large populations of patients [2]. However, the increasing number of clinical and biological parameters that must be collected for precision medicine makes it almost impossible to design dedicated trials [3]. A Medical Record (MR), which was a paper document in the past and now is mostly computerized, is the systematic documentation of a patient's medical care history across time within one particular health care provider's jurisdiction [4]. Namely, MRs can be seen as the logs that describe patients' treatments and other hospital activities with outcomes, such as recovery, transfer and death.

As the most important medical log documenting patients' whole treatment processes and evolving statuses, Electronic Medical Records (EMR) contain rich clinical data, information and knowledge. Furthermore, the amount of EMR data has been rapidly increasing with the hospital

The associate editor coordinating the review of this article and approving it for publication was Dalei Wu.

digitalization process. From the view of evidence-based medicine, EMRs are the best clinical evidence of positive (i.e., recovery) or negative (i.e., death) treatment results. EMRs also have great potential for enabling data-driven predictions and further explorations of clinical knowledge if the EMR data can be deeply analyzed. The mining outcomes are helpful to support physicians' decisions during patients' journeys [5]–[7], such as treatment pathway personalization. Recent studies [8]–[10] show that machine-learning methods are essential and powerful EMR-analyzing technologies that can mine the underlying knowledge [8]–[10].

In this research, we aimed to discover medication patterns from highly complex drug treatments and diseases by systematically applying Jaccard similarity, spectral clustering, and Bayes probability techniques for EMR text mining. Cirrhotic ascites EMR data were selected to evaluate our approach since patients with this disease always have many complications and require complex medication patterns [11]. Our approach outperformed other machine-learning-based methods in the following aspects. 1) The approach effectively discovered the major medication patterns from EMR text for highly complex diseases and mixed medication patterns. 2) This approach can separate the highly mixed medication treatments into distinct clustered medication patterns rather than vague future clustering that classifies every item into one of the treatment patterns, although the similarity was weak. 3) Unlike unsupervised deep-learning-based treatment pattern discovery methods, our classification approach results in each step of the framework being interpretable rather than a black box [12]. This approach is important for clinical knowledge discovery (since it is evidence-based and interpretable) to understand the classifying processes of certain drugs for clinical purposes. 4) To the best of our knowledge, this is the first time machine learning methods have been used for medication discovery from cirrhotic ascites EMRs.

The rest of this paper is organized as follows. Section II is the review of related works. Section III describes the proposed framework and methods to discover latent medication patterns. Section IV presents the real data experiments and discussions followed by the conclusions in the last section.

II. RELATED WORKS

EMR text mining is a systematic project that includes data extraction and mining from textual EMR sources (with a focus on the mining part) to discover meaningful knowledge. Similar to an EMR, an Electronic Health Record (EHR) is also a medical record, but it emphasizes daily data across hospitals, social care and self-reported/monitored data for long-term health status. Therefore, this section provides a state-of-the-art review of both EMR and EHR textual data processing and mining research.

A. EMR TEXT PRE-PROCESSING

It has been a challenging task to analyze the textual content of EMRs in current research [13]. Existing research has adopted the Natural Language Processing (NLP) technique that focuses on targeting and extracting useful information from EMRs for clinical research [14]–[18]. The International Classification of Diseases (ICD) and the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) are two widely adopted medical term sets and powerful tools in medical semantics analytics. These coding systems play important roles in targeting specific words or segmentations within EMR text. However, those standardized terms are not always mandatory in EMR records. Some studies [15], [16], [19], [20] use multiple lexicon databases to cover a wide range of words or phrases to create their own dictionaries. Another challenge in processing Chinese EMRs is that there is no official Chinese SNOMED-CT available, while English EMR text mining is relatively easier since every word in a sentence is separated by spaces. Spaces help to improve the precision of information segmentation in text mining, but Chinese sentences are made up of characters with no spaces between words [21]. Some authors [22]–[24] have attempted Chinese EMR text mining based on a special lexicon or a manually established ontology structure. Some meaningful results have been discovered from those approaches, but they require significant manual work to process the textual data. Although word segmentation in Chinese characters is a challenging task, it is a prerequisite step to mine information and knowledge from EMR text directly.

In this study, we focused on mining clinical knowledge from EMR text with a hybrid text-preprocessing approach. To segment Chinese text in EMRs, we first selected the medical terms with Chinese ICD codes. Then, we used a word dictionary from existing hospital information systems and further created the dictionary for this research. The aim of EMR text pre-processing is to extract meaningful words and information from the text in order to provide the computer-readable data for subsequent processing.

B. EMR/EHR DATA MINING

The deep learning approach has attracted significant attention in data representation and prediction from high dimensional EMR data. Reference [25] proposed a deep multi-modality architecture for EHR analysis based on Poisson Factor Analysis modules. This architecture is able to identify Type 2 Diabetes Mellitus (T2DM) patients from a group of all kinds of patients using diagnosis codes and laboratory tests. Reference [26] presented Deepr (short for Deep record), a new end-to-end deep learning system, that learns to extract features from medical records and predict future risks automatically. Reference [27] proposed a deep learning approach for phenotyping from patient EHRs. Existing deep learning approaches always focus on disease prediction or identification based on EMR/EHR non-picture data, such as the work of [28]–[30]. However, in unsupervised deep learning cases, the accuracy varies due to the lack of consideration of uncertainties in complex diseases. Furthermore, deep learning as a black box with obscure reasoning processes has limited its capability of supporting clinical research. Our proposed framework aims to discover hidden patterns that are normally omitted in high-complexity medication diseases and give more semantics and interpretability to identified patterns.

Traditional machine learning methods (e.g., non-deep learning methods) are also powerful tools for mining knowledge from EMR data. Reference [31] demonstrates a framework for identifying subjects with and without T2DM from EHRs with relatively high accuracy and performance. This framework integrates several machine learning methods, such as the k-Nearest-Neighbors, Naïve Bayes, Decision Tree, Random Forest, Support Vector Machine and Logistic Regression. Based on k-means clustering and EMR data, Rajkomar *et al.* [32] characterized the utilization patterns of primary care patients and created weighted panel sizes for providers based on the work required to care for patients with different patterns. Reference [19] classified obesity and obesity types in thousands of EMRs by using the Support Vector Machine (SVM) and Naïve Bayes models, and the experimental results indicate that the SVM outperforms other methods. Reference [33] proposed a flexible hierarchical Bayesian nonparametric model to cluster medical data into groups. This work was inspired by the structure of ICD codes that can present semantic relationships between different diseases. Similarly, an improved Latent Dirichlet allocation (LDA) model [34] that is also a probabilistic model was applied to discover the changing trends of medical behaviors over time from EMRs. The above machine learning models provide effective ways to mine knowledge from EMRs. However, they are context-sensitive, since the same model always has different results in different EMR mining scenarios. These models require enhanced robustness in dealing with context-dependent scenarios.

Following the LDA model, References [35]–[37] proposed the latent treatment pattern discovery model based on the treatment logs to analyze the EMRs of cardiovascular and cerebrovascular diseases. Although the temporal dimensions are important in clinical practice, the LDA has limited capability of identifying the time sequence of treatment activities from the medical logs. To solve this problem, References [36], [37] used the tuple <timestamp, activity> to identify a single treatment activity in the LDA model. In this paper, we discovered medication patterns from EMRs in hepatocirrhosis ascites diseases. The pure LDA model could not adequately identify the medication patterns for hepatocirrhosis ascites. Since there are many complexities among treatments for different patients with this disease, the medication treatments for some patients always have subtle variances, and drug uses are always highly mixed. Moreover, traditional LDA-based methods cannot compare the relative importance of the treatment between different patient groups/patterns. According to the real data test in the following section of this paper, the proposed framework and methods effectively discovered the distinct medication patterns with minor differences for this disease with highly complex medication patterns. Furthermore, the time-density-reduction method can support the relative importance of the same drug in different medication patterns.

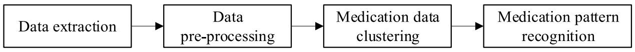


FIGURE 1. General steps to discover medication pattern from EMR text.

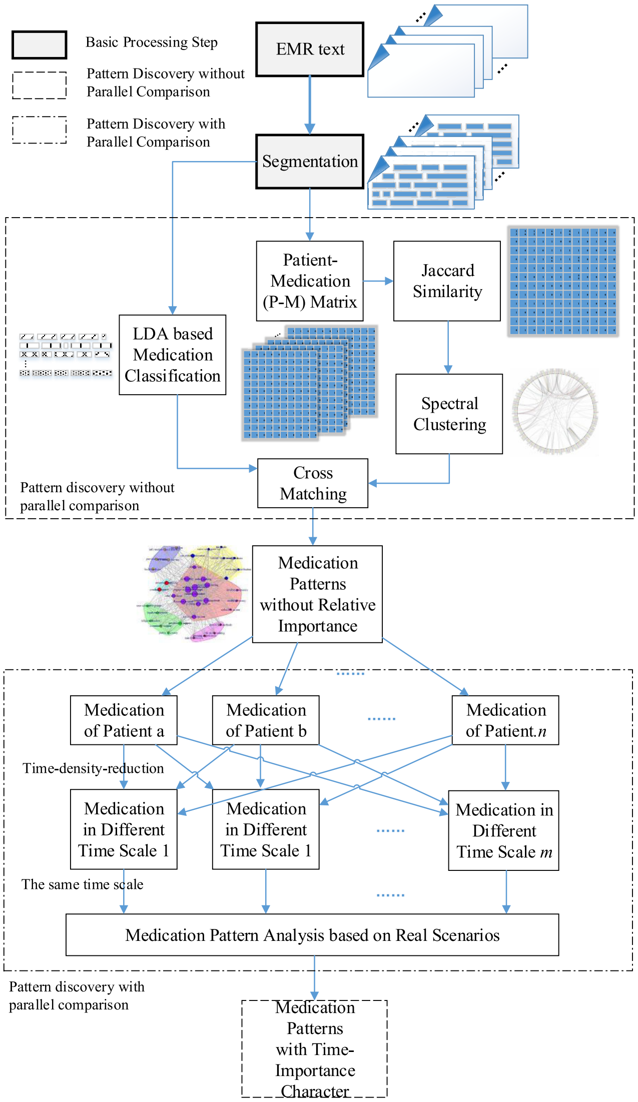

Basic Processing Step
Pattern Discovery without Parallel Comparison
Pattern Discovery with Parallel Comparison
EMR text
Segmentation
Patient-Medication (P-M) Matrix
Jaccard Similarity
LDA based Medication Classification
Spectral Clustering
Cross Matching
Pattern discovery without parallel comparison
Medication Patterns without Relative Importance
Medication of Patient a
Medication of Patient b
......
Medication of Patient.n
Time-density-reduction
Medication in Different Time Scale 1
Medication in Different Time Scale 1
Medication in Different Time Scale m
The same time scale
Medication Pattern Analysis based on Real Scenarios
Pattern discovery with parallel comparison
Medication Patterns with Time-Importance Character

FIGURE 2. Framework of medication pattern discovery.

III. FRAMEWORK AND METHOD TO DISCOVER LATENT MEDICATION PATTERNS

A. GENERAL METHOD TO DISCOVER THE MEDICATION PATTERN

Generally, four steps are needed to discover treatment patterns in EMR text, as shown in Figure 1. These steps include data extraction, data pre-processing for classification, medication data clustering and medication pattern recognition.

In the first step, the challenge lies in precisely extracting meaningful data from the EMR text. An EMR is a

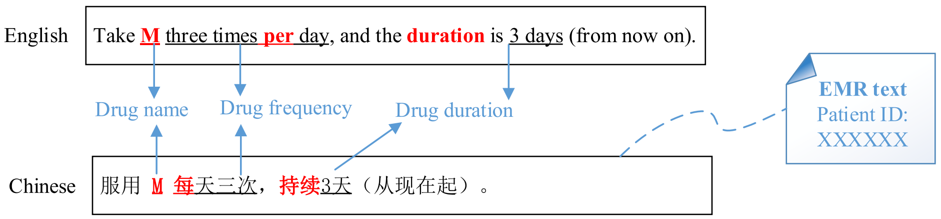
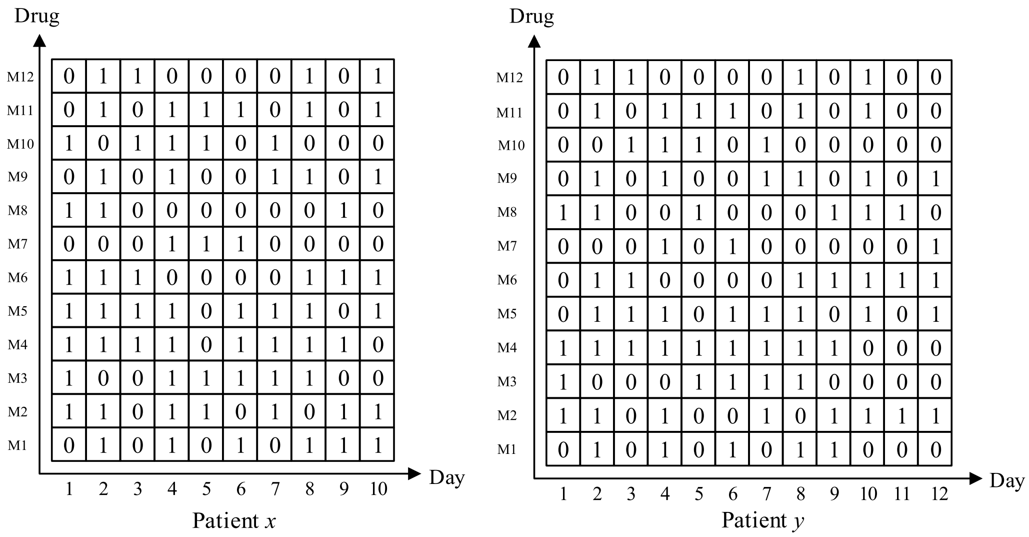

English
Take M three times per day, and the duration is 3 days (from now on).
Drug name
Drug frequency
Drug duration
Chinese
服用 M 每天三次，持续3天（从现在起）。
EMR text
Patient ID:
XXXXXX

FIGURE 3. Medication data extracted from EMR text.



Drug	1	2	3	4	5	6	7	8	9	10
M12	0	1	1	0	0	0	0	1	0	1
M11	0	1	0	1	1	1	0	1	0	1
M10	1	0	1	1	1	0	1	0	0	0
M9	0	1	0	1	0	0	1	1	0	1
M8	1	1	0	0	0	0	0	0	1	0
M7	0	0	0	1	1	1	0	0	0	0
M6	1	1	1	0	0	0	0	1	1	1
M5	1	1	1	1	0	1	1	1	0	1
M4	1	1	1	1	0	1	1	1	1	0
M3	1	0	0	1	1	1	1	1	0	0
M2	1	1	0	1	1	0	1	0	1	1
M1	0	1	0	1	0	1	0	1	1	1

Day

Patient *x*

Drug	1	2	3	4	5	6	7	8	9	10	11	12
M12	0	1	1	0	0	0	0	1	0	1	0	0
M11	0	1	0	1	1	1	0	1	0	1	0	0
M10	0	0	1	1	1	0	1	0	0	0	0	0
M9	0	1	0	1	0	0	1	1	0	1	0	1
M8	1	1	0	0	1	0	0	0	1	1	1	0
M7	0	0	0	1	0	1	0	0	0	0	0	1
M6	0	1	1	0	0	0	0	1	1	1	1	1
M5	0	1	1	1	0	1	1	1	0	1	0	1
M4	1	1	1	1	1	1	1	1	1	0	0	0
M3	1	0	0	0	1	1	1	1	0	0	0	0
M2	1	1	0	1	0	0	1	0	1	1	1	1
M1	0	1	0	1	0	1	0	1	1	0	0	0

Day

Patient *y*

FIGURE 4. P-M matrices with different sizes.

professional and technical document written by physicians in a natural language. Therefore, the key question is where to cut or split a sentence into meaningful words and phrases. In the English language, two words are separated by spaces. However, there are no spaces in Chinese sentences to separate words. Thus, the effectiveness of word segmentation determines the accuracy of the data extracted from EMRs. In the second step, the extracted data in the form of numbers, words and phrases are transformed into a format that the computer can process, such as a word matrix. In the third step, a medication data framework classifies the extracted drugs into different groups. However, in order to reveal the clinical meaning from the clustered data, the fourth step recognizes the medication pattern.

Based on the general process above, we developed a hybrid medication pattern discovery method by introducing the LDA model and the special clustering algorithm. Figure 2 describes the mining process that will be elaborated upon in the subsequent section.

B. BASIC PROCESSING STEPS

Data segmentation and extraction is the basic processing step that begins the mining process. Although Chinese is slightly different from English in an NLP analysis, extracting drug use situations is easier than extracting other elements (i.e., the description of illness) in EMR free text. This is because the name of the drug is stable and consistent. To extract Chinese drug treatment content from EMR text, we used a medicine dictionary provided by our collaborative hospital to extract the names of drugs. Following the names of drugs in EMR text, there are usage instructions, such as the specific drug usage time and duration. We used keywords to extract this information automatically. Figure 3 shows an example that extracts necessary data from EMR text. In this example, red words (English) / characters (Chinese) are keywords that can be used to locate corresponding information in the sentence. The drug name can be extracted from the sentences using a dictionary.

C. PATTERN DISCOVERY WITHOUT TIME SEQUENCE

1) REPRESENT THE MEDICATION TREATMENT OF PATIENTS WITH A P-M MATRIX

Based on the extracted data from EMRs, we used a two-dimensional matrix to represent the medication process for each patient. Columns of the matrix indicate the day of the treatment in the hospital, and each line indicates a drug. Assume Π^{ξ} is a P-D (Patient-Drug) matrix of patient ξ, and π_{ij}^{ξ} is an element in Π^{ξ}. Then, $\pi_{ij}^{\xi} = 1$ if this patient used drug i on the j^{th} day, and otherwise $\pi_{ij}{}^{\xi} = 0$.

Note that the size of these matrices may be different since different patients may have different lengths of stay (LoSs). Figure 4 gives two medication matrices with different sizes for Patient x and Patient y, where the LoS is 10 days and 12 days for Patient x and Patient y, respectively.

The size discrepancy of the P-D matrix has increased the challenge of comparing the similarities among matrices. Although the P-D matrix is an important way to represent and store the medication process, it is difficult to directly cluster the medication treatments of different patients using the P-M matrix. One possible way is to drop the time information from the matrix when calculating the similarity and clustering of the medication patterns. In the following steps, we clustered the medication cases without considering timestamps.

2) MEASURE THE SIMILARITY OF PRESCRIBED MEDICATIONS WITH THE JACCARD SIMILARITY COEFFICIENT

Similarity is an important factor that can cluster patients with similar medication treatments. The Jaccard index [38] is introduced to measure the similarity of the medication experiences of patients who have different LoSs. The Jaccard similarity coefficient J is defined as equation (1):

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

where A and B are sample sets. The definition can be understood as the size of the intersection divided by the size of the union of the sample sets. The similarity of sets A and B can be acquired by (2).

$$J_d(A, B) = 1 - J(A, B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \tag{2}$$

in which J_d (the Jaccard Distance) is the distance between A and B.

According to equations (1) and (2), we calculated the similarity of medication treatments of different patients based on a P-M matrix. A Patient-Patient Matrix (P-P Matrix) can describe the similarity degree. For illustration, we extracted part of the patients into a P-P Matrix PJ shown in (3), as shown at the bottom of this page, which included 20 patients.

It can be seen that an adjacent matrix (3) is symmetric, and each element pj_{ab} in PJ indicates the medication similarity between patient a and patient b.

3) CLUSTER THE PATIENTS WITH SPECTRAL CLUSTERING

Based on the similarity measurement, we can obtain a graph, PG, where all of the patients are connected with each other under certain edge weights. The weight value is the similarity between any two patients. Now, the problem of clustering these patients can be solved by cutting the graph into several sub-graphs. Each of the sub-graphs represents a patient cluster in which the patients probably have similar medication treatments.

During patient clustering, the aim of cutting is to minimize the weights of the connection between two sub-graphs while the weights of the connections inside the sub-graphs are high. Thus, we introduced a special clustering model called Normalized Cut (N-cut) [39]. Considering the connection between the relative density of each group, an N-cut ($Ncut$) can be described as (4):

$$Ncut(C, D) = \frac{cut(C, D)}{vol(C)} + \frac{cut(C, D)}{vol(D)} \tag{4}$$

where C and D are sample sets, vol is the volume in each sample set, and cut is the sum of the weight of the edges that

$$PJ_{20*20} = \begin{pmatrix}
1 & 46.2 & 13 & 26.7 & 18.8 & 20 & 22.6 & 3.8 & 13.5 & 16.7 & 9.1 & 14.3 & 17.6 & 26.9 & 19.4 & 14.8 & 27.6 & 21.1 & 20.6 & 24 \\
 & 1 & 13 & 18.8 & 15.2 & 25 & 18.8 & 3.8 & 20 & 14 & 17.1 & 18.5 & 25 & 22.2 & 16.2 & 19.2 & 23.3 & 17.9 & 24.2 & 19.2 \\
 & & 1 & 23.8 & 18.2 & 10 & 15.6 & 2.5 & 19.1 & 23.5 & 19.2 & 12.2 & 25.6 & 23.7 & 16.3 & 20.6 & 13.3 & 17.6 & 27.9 & 15.4 \\
 & & & 1 & 31 & 11.1 & 22.6 & 0 & 16.7 & 16.7 & 9.1 & 14.3 & 14.3 & 43.5 & 22.9 & 14.8 & 23.3 & 27.8 & 20.6 & 24 \\
 & & & & 1 & 20 & 35.7 & 3.8 & 27.3 & 19.5 & 6.7 & 23.1 & 17.6 & 43.5 & 26.5 & 19.2 & 27.6 & 27.8 & 20.6 & 24 \\
 & & & & & 1 & 20 & 0 & 21.4 & 7.9 & 11.1 & 33.3 & 23.1 & 19 & 12.9 & 12.5 & 26.1 & 15.2 & 22.2 & 21.1 \\
 & & & & & & 1 & 3.8 & 31.3 & 25.6 & 11.6 & 18.5 & 17.6 & 32 & 26.5 & 17 & 32.1 & 27.8 & 24.2 & 29.2 \\
 & & & & & & & 1 & 3.3 & 2.7 & 5.7 & 0 & 0 & 0 & 3.2 & 2 & 4 & 2.9 & 0 & 0 \\
 & & & & & & & & 1 & 26.2 & 20.9 & 16.1 & 29.4 & 23.3 & 20.5 & 24.5 & 24.2 & 31.6 & 32.4 & 20.7 \\
 & & & & & & & & & 1 & 37.2 & 10.3 & 21.4 & 25.7 & 22.7 & 32.7 & 23.1 & 26.7 & 30 & 16.7 \\
 & & & & & & & & & & 1 & 7.7 & 22 & 10.3 & 10.4 & 26.3 & 9.3 & 14.3 & 18.6 & 5.1 \\
 & & & & & & & & & & & 1 & 25.9 & 28.6 & 27.6 & 14.3 & 24 & 17.6 & 25 & 19 \\
 & & & & & & & & & & & & 1 & 25 & 21.6 & 28 & 21.9 & 23.1 & 48.3 & 22.2 \\
 & & & & & & & & & & & & & 1 & 46.2 & 21.3 & 33.3 & 36.7 & 33.3 & 36.8 \\
 & & & & & & & & & & & & & & 1 & 24.1 & 31.3 & 34.2 & 27.8 & 33.3 \\
 & & & & & & & & & & & & & & & 1 & 27.1 & 27.3 & 32.7 & 17 \\
 & & & & & & & & & & & & & & & & 1 & 36.4 & 33.3 & 50 \\
 & & & & & & & & & & & & & & & & & 1 & 32.4 & 25.8 \\
 & & & & & & & & & & & & & & & & & & 1 & 36 \\
 & & & & & & & & & & & & & & & & & & & 1
\end{pmatrix} \tag{3}$$

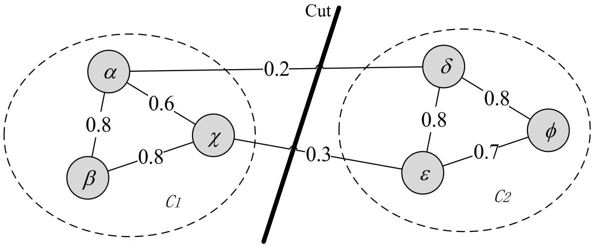


FIGURE 5. An example of graph cutting.

are cut. *vol* and *cut* can be calculated via (5), (6), and (7):

$$vol(C) = \sum_{i\in C, j\in V} w_{ij} \tag{5}$$

$$vol(D) = \sum_{i\in D, j\in V} w_{ij} \tag{6}$$

$$cut(C, D) = \sum_{i\in C, j\in D} w_{ij} \tag{7}$$

The method needs to find the cutting place in the graph to minimize the value of *Ncut* to identify the clusters. The resolution of the problem *min*(*Ncut*(C, D)) can refer to [39].

Figure 5 gives an example of clustering patients by using spectral clustering. Assume α, β, χ, δ, ε, and ϕ are patients, and the edges are the similarities that are obtained by the P-P Matrix. By cutting the edge with the weights of 0.2 and 0.3, *Ncut*(C_1, C_2) can reach the minimum value. Thus, the six patients are separated into two clusters, and the patients in the same cluster have similar medication conditions.

Generally, the patient clustering process can be described as follows.

Step 1: Input the similarity matrix ***PJ*** and the number of clusters K.

Step 2: Calculate the diagonal matrix ***D*** and the symmetric matrix of ***PJ***. Then, obtain the Laplacian matrix ***L and L′***, where $\boldsymbol{L} = \boldsymbol{D}\text{-}\boldsymbol{A}$, $\boldsymbol{L}' = \boldsymbol{D}^{-1/2}\boldsymbol{LD}^{-1/2}$.

Step 3: Calculate the eigenvalue and eigenvector of L': $\boldsymbol{Ve} = \{v_{e1}, v_{e2}, \ldots, v_{em}\}$.

Step 4: Cluster patients according to the first K eigenvector $\boldsymbol{V_K} = \{v_{e1}, v_{e2}, \ldots, v_{ek}\}$.

Note that in Step 1, the number of clusters is required. The number of clusters can be acquired by the LDA method, which we illustrate in section 3.3.4. In Step 4, a clustering method, such as K-means, is required to develop the final clustering result.

The rule of acquiring Spectral Clustering is as follows:

Rule of Spectral Clustering the Patients:

```
Integer k;
            //W is a P-P Matrix
Array W[n][n];
Get W[n][n];
Array D[n][n];
For i = 0 to n-1
   For j = 0 to n-1
      If i = j
         D[i][j] = W[i][0] + W[i][1] + ...... + W[i][n-1];
      Else
         D[i][j] = 0;
Array L[n][n];
For i = 0 to n-1
   For j = 0 to n-1
      L[i][j] = D[i][j]-W[i][j]; Array SL[n][n];
SL[n][n] = D[n][n]$^{1/2}$L[n][n]D[n][n]$^{1/2}$;
Array F[n][m];
Find the m smallest eigenvalues and the
corresponding eigenvectors of F[n][m];
F[n][m]=All of the eigenvectors;
Get k clusters by K-Means(F[n][m],k)
```

4) LDA MINING

LDA [40] is a generative probabilistic model for collections of discrete data, such as text corpora. It is a three-level hierarchical Bayesian model where each item of a collection is modeled as a finite mixture over an underlying set of topics.

LDA models have proven effective in revealing the mixture risk and medical behavior trends from EMRs [34]–[41]. We propose using the LDA to calculate the similarity degree of each medication trace to cluster drugs into different groups. In EMR text mining using the LDA, the EMR text for a patient can be seen as a document; the name of a medication can be seen as a word; and the latent medication patterns can be seen as the topics to be discovered. However, in practice, physicians always use the format of <name of the drug: frequency in one day, duration of the days> to record the medication process in EMR text. Therefore, when sampling the name of drugs, we should label the drug name in each day of the medication duration. Figure 6 is the medication pattern discovery probabilistic graphical model based on LDA.

Similar to the new document generation process in the LDA model, the medication (drug) generation process is as follows.

Step 1: Randomly choose a distribution Φ for a medication pattern, where Φ is subject to a Dirichlet Distribution with parameter β ($\Phi \sim Dir(\beta)$).

Step 2: Randomly choose a distribution δ for a medication pattern and drug use frequency, where δ is subject to a Dirichlet Distribution with parameter v ($\delta \sim Dir(v)$).

Step 3: Randomly choose a distribution ρ for a medication pattern and drug use duration, where ρ is subject to a Dirichlet Distribution with parameter w ($\rho \sim Dir(\eta)$).

Step 4: Randomly choose a distribution θ for a patient's medication treatment D, where θ is subject to a Dirichlet Distribution with parameter α ($\theta \sim Dir(\alpha)$).

Step 5: Choose D drugs by repeating the following three sub-steps.

Sub-step 5-1: Probabilistically draw a medication pattern z from a multinomial distribution θ ($z \sim multi(\theta)$).

Sub-step 5-2: Probabilistically draw a drug use frequency l from δ.

Sub-step 5-3: Probabilistically draw a drug use duration f from ρ.

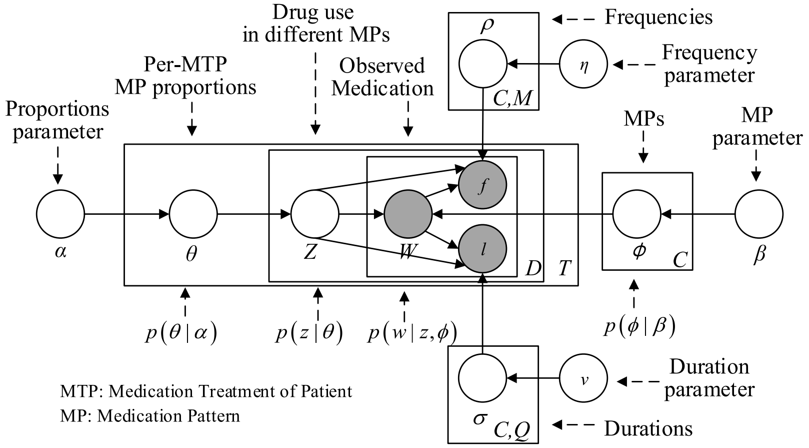


FIGURE 6. Probabilistic graphical model of EMR text mining.

Based on the model and steps above, Gibbs Sampling is an efficient method to resolve the LDA-based problem [42], [43]. Following the idea of Gibbs Sampling method, we need to acquire $p(z, w, f, l|\alpha, \beta, v, \eta)$ and the conditional probability distribution of drug i over medication pattern k. This probability can be represented by $p(z_i = k|z_{\Gamma i}, w, f, l, \alpha, \beta, v, \eta)$, where $z_{\Gamma i}$ is the medication pattern distribution without drug i. Finally, we can get the probability of drug i over a certain medication pattern when Gibbs Sampling is convergent.

According to Figure 6, we can obtain the following joint probability:

$$p\,(z, w, f, l, |\alpha, \beta, v, \eta) = p(z|\alpha)p(w|z, \beta)p(f|w, z, \eta)p(l|w, z, v) \tag{8}$$

where

$$\begin{aligned} p(z|\alpha) &= \int p(z|\theta)p(\theta|\alpha)d\theta \\ &= \int \prod_{\tau=1}^{T} \frac{1}{\Delta(\alpha)} \prod_{c=1}^{C} \theta_{\tau,c}^{n_\tau^{(c)}+\alpha_c-1} d\theta_\tau \\ &= \prod_{\tau=1}^{T} \frac{\Delta(\alpha + n_\tau)}{\Delta(\alpha)} \end{aligned} \tag{9}$$

in which $\Delta(\alpha)$is a Dirichlet delta function and $n_\tau = \{n_\tau^{(c)}\}_{c=1}^{C}$ is the count of allocating the medication pattern c to patient τ.

Similarly, we can use the following equations.

$$p(w|z, \beta) = \prod_{c=1}^{C} \frac{\Delta(\beta + r_c)}{\Delta(\beta)}, \quad m_c = \{r_c^{(w)}\}_{w=1}^{W} \tag{10}$$

$$\begin{aligned} p(f|w, z, \eta) &= \prod_{m=1}^{M} \prod_{c=1}^{C} \frac{\Delta(\eta + x_{m,c})}{\Delta(\eta)}, \\ x_{m,c} &= \{x_c^{w,m}\}_{w=1,m=1}^{W,M} \end{aligned} \tag{11}$$

$$\begin{aligned} p(l|w, z, v) &= \prod_{m=1}^{M} \prod_{c=1}^{C} \frac{\Delta(v + y_{q,c})}{\Delta v(v)}, \\ y_{q,c} &= \{y_c^{w,q}\}_{w=1,q=1}^{W,Q} \end{aligned} \tag{12}$$

where $\Delta v(\beta)$, $\Delta v(\eta)$, and $\Delta v(v)$ are Dirichlet delta functions; $m_c = \{m_c^{(w)}\}_{w=1}^{W}$ is the count of allocating drug w to medication pattern c; $x_{m,c} = \{x_c^{w,m}\}_{w=1}^{W}$ is the count of allocating drug w with using frequency m to medication pattern c; and $y_{n,c} = \{y_c^{w,q}\}_{w=1}^{W}$ is the count of allocating drug w with use duration q to medication pattern c.

Therefore, the joint probability (8) can be represented by (13):

$$\begin{aligned} p\,(z, w, f, l, |\alpha, \beta, v, \eta) &= \prod_{\tau=1}^{T} \frac{\Delta(\alpha + n_\tau)}{\Delta(\alpha)} \cdot \prod_{c=1}^{C} \frac{\Delta(\beta + m_c)}{\Delta(\beta)} \\ &\quad \cdot \prod_{m=1}^{M} \prod_{c=1}^{C} \frac{\Delta(\eta + x_{m,c})}{\Delta(\eta)} \cdot \prod_{m=1}^{M} \prod_{c=1}^{C} \frac{\Delta(v + y_{n,c})}{\Delta v(v)} \end{aligned} \tag{13}$$

According to the characteristic of the Gamma function that $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha + 1)$, we have (14):

$$\begin{aligned} p(z_i = k|z_{\Gamma i}, w, f, l, a, b, v, \eta) &\propto \hat{\theta} \cdot \hat{\phi} \cdot \hat{\rho} \cdot \hat{\sigma} \\ &= \frac{\alpha_c + n_{\tau,\Gamma i}^{(c)}}{\alpha_s + \sum_{s=1}^{C} n_{\tau,\Gamma i}^{(s)}} \cdot \frac{\beta_w + r_{c,\Gamma i}^{(w)}}{\beta_j + \sum_{j=1}^{W} r_{c,\Gamma i}^{(j)}} \cdot \frac{\eta_{w,m} + x_{c,\Gamma i}^{(w,m)}}{\eta_j + \sum_{j=1,\pi=1}^{W,M} x_{c,\Gamma i}^{(j,\pi)}} \\ &\quad \cdot \frac{v_{w,q} + y_{c,\Gamma i}^{(w,q)}}{v_j + \sum_{j=1,\lambda=1}^{W,Q} y_{c,\Gamma i}^{(j,\lambda)}} \end{aligned} \tag{14}$$

Thus, the Gibbs Sampling parameters $\hat{\theta}$, $\hat{\phi}$, $\hat{\rho}$, and $\hat{\sigma}$ are acquired, and we can use these parameters to calculate the distribution of medication patterns. The medication discovery pattern is the reverse of the medication (drug) generation process.

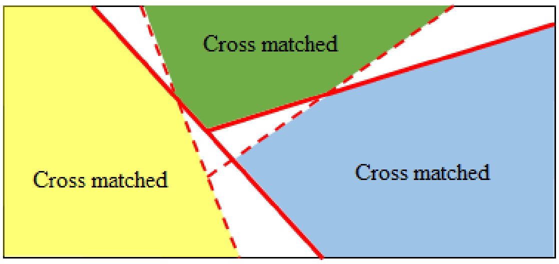


FIGURE 7. An example of cross matching ($K = 3$).

Note that the parameters of Gibbs Sampling and the number of clusters (namely, the number of patterns to be classified) should be given as input parameters when using the LDA-based method. Traditionally, this parameter can be determined by the index of *perplexity* [40] and the real medication scenario. However, some studies [44], [45] concluded that perplexity is not always the best way to determine the number of clusters, and it is sometimes slightly uncorrelated with human judgment. Therefore, in this framework, we recommend determining the number of clusters by using actual clinical scenarios. However, *perplexity* is also acceptable to evaluate the number of clusters when necessary. In IV-B-2, we give an example of determining the number of clusters using real scenarios.

5) CROSS MATCHING FOR MAXIMAL PATTERN COVERAGE

In the EMR text analysis process, the Spectral Clustering method can separate the low weight linkages between patients with low similarity and formulate groups of patients with high similarity. The LDA-based method can calculate the similarity degree of each medication history of patients and then cluster drugs into different groups of different medication patterns. Note that in the LDA-based method, the same drug may appear in different patterns with different memberships. As shown in Figure 7, the medication patterns generated by the LDA-based method may overlap, while the Spectral Clustering method separates patients without any overlap.

As mentioned above, apart from explicitly recognizable differences, there are many subtle medication variances among different patients with similar cirrhotic ascites treatments. This variance has increased the medication complexity since physicians always use drugs with different functions in similar treatment episodes in practice. However, there are very few existing clinical pathway guidelines distinguishing such variances in medical processes. To discover those medication patterns hidden among these medical logs and additional clinical knowledge with less distinct features, we cross-matched the feature results of Spectral Clustering and the LDA-based methods with their cluster sets. This approach was inspired by multiple clustering related studies [46] that emphasize integrative clustering methods to find more stable and robust solutions rather than a single subset with a single algorithm with a single validity metric. For example, consensus clustering takes multiple subsets of a dataset and uses repeated predictions of cluster assignments to gauge stability [47]. HOPACH [48] recursively partitions a dataset while seeking to optimize a clustering measure. Both methods improve from previous single methods by repeatedly examining sub-clusters of a dataset. Moreover, the COMMUNAL method [46] proved that integrating information from multiple clustering algorithms and multiple validity measures would improve the signal and assist in determining stability. In other words, if the number of clusters is determined, cross-matching on different clustering results may increase the focus of each cluster and enlarge the gap between different clusters. To make the discovered medication patterns represent rich clinic meanings, maximal coverage to represent the hidden knowledge that was lost in single clustering is the objective of cross-matching, especially in complicated medication processes. The rules of cross-matching are described as follows.

```
Rules of Cross-Matching:
        //LDAClus[i][0]←patientId
        LDAClus[i][1]←meditation-pattern-Id
Array LDAClus[][];
        //SpectralClus[i][0]←patientId
SpectralClus[i][1]←meditation-pattern-Id
Array SpectralClus[][];
LDAClus[][]←Get data from LDA Clustering result;
SpectralClus[][]←Get data from Spectral Clustering
result;
        //match[i][0]←meditation-pattern-Id of LDA
Clustering
        //match[i][1]←meditation-pattern-Id of Spectral
Clustering
        //match[i][2]←total number of patients that
LDAClus[x][0]=SpectralClus[y][0]
        //Assume that there are N meditation-patterns
after clustering
Array match[N*N][3];
For i = 1 to |LDAClus[]|
        where match[x][0]=LDAClus[i-1][0] and
match[x][1]=SpectralClus[i-1][0]
         match[x][2]++;
        //match[X1][0]∪match[X2][0]∪...∪
match[XN][0] = $\sum_{i=0}^{N-1} LDAC[i][1]$
        //match[X1][1]∪match[X2][1]∪...∪
match[XN][1] = $\sum_{i=0}^{N-1} SpectralClus[i][1]$
Find
max(match[X1][2]+match[X2][2]+...+match[XN][2]);
Find the information of patients that belong to new
cross matching patterns;
Re-clustering N new meditation patterns
```

D. MEDICATION PATTERN DISCOVERY WITH TIME-IMPORTANCE

The results acquired by clustering lack time information since LDA is a word-bag-based method. To mine more

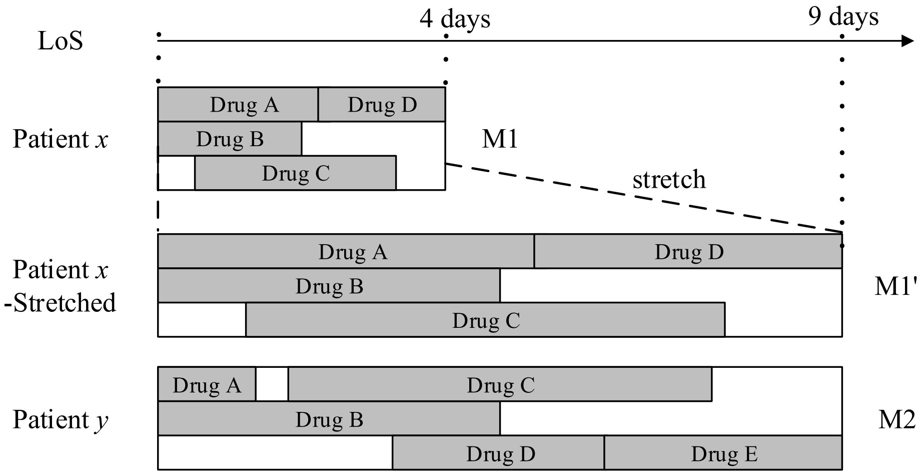


FIGURE 8. Time-density-reduction.

meaningful information related to medication patterns and for the integrity of the framework, this section provides a basic method to discover more information from temporal sequences and evaluate the relative importance of the same drug in different patterns. As denoted in the P-D matrix, the LoSs of different patients vary, which makes measuring the changing drug usage in the same time scale more difficult. For example, a patient used drug *A* for 3 days during his/her total 3-day stay, while another patient used the same drug for 6 days during the first days of his/her 20 day stay. How can we compare the drug use modes in this context? In this scenario, we propose a time-density-reduction method to evaluate the temporal drug use characteristic. In this method, we extended the LoSs of patients whose LoSs were shorter than the maximum LoS to make the medication process even. Following the above example, the LoS of 3 is extended to 20 days, and he/she uses drug *A, D, B,* and *C* throughout the 20 days all the time at the same frequencies. Figure 8 illustrates the time-density-reduction. In this figure, the LoS of Patient *x* is extended to the same as that of patient *y*. For the P-D matrix, the time-density-reduction is equivalent to the extension on the columns.

After standardizing the LoSs of different patients to the same time scale, more work is required to discover medication characteristics according to the scenario's needs, such as the drug use period of a certain patient group. Note that there are also negative effects on the method of time-density-reduction. One problem is the distortion of absolute drug use quantity, which means that the processed data cannot be used for calculating or comparing the absolute drug use quantities of different patients. In this scenario, we are still able to evaluate the relative importance of drugs in different patterns. For example, in Figure 8, the absolute days of using drug D for Patient *x* is smaller than that of Patient *y*, whereas the stretched Patient *x* has a longer use period than Patient *y*. This indicates that Drug D plays a more important role in the treatment of Patient *x* than of Patient *y*, which we call the 'time-density-reduction versus importance' phenomenon.

IV. REAL DATA ANALYSIS

In this data experiment, we selected the patients with cirrhotic ascites. Ascites is the most common complication of cirrhosis, and 60% of patients with compensated cirrhosis develop ascites within 10 years during the disease's progression [49], [50]. Ascites formation is the signal that the illness has progressed into a decompensated period that is a serious stage and requires treatment with medications. Medication for ascites always varies among patients with complex subtle differences with both known and unknown clinical knowledge. Recognizing all the medication variances is also an important part of the whole treatment process for many internal medicine diseases, and it is one of the most crucial components in clinical pathway guideline development.

A. DATA SET AND DATA PRE-PROCESSING

Following the above methods, we analyzed real EMR text from a major hospital in China. We included 998 inpatient EMRs from June 2014 to July 2016. All EMRs were from patients with hepatocirrhosis ascites, which means these patients have a main diagnosis of liver cirrhosis (ICD code: K74.151) and a secondary diagnosis of ascites (ICD code: R18), excluding the cases caused by hepatic carcinoma. In these data, we further excluded the patients who were transferred or dead and kept the ones that obtained effective treatment before being discharged from the hospital.

The data were cleaned and anonymized before analysis, which means the patients' personal information, such as name, age, address, and phone number, were removed. The patient ID was re-coded so that it would be hard to backtrack the ID to the original patient but still be able to identify a unique patient. Figure 9 gives an example of patient ID obfuscation.

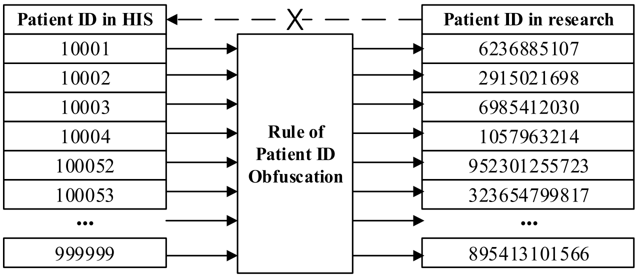


FIGURE 9. Example of patient ID obfuscation.

Based on the masked data, we extracted the medication information from the text. During this process, we paid attention to the duration and frequency of drug use. Namely, the appearance of a drug name in one patient treatment record was in the format of frequency and duration. However, according to the advice of physicians in this data experiment, we ignored the frequency of the drug used in one day because the frequency is always concerned with the degree of the illness and the dosage of drug, which is less important to the clinical pathway and medication scheme. Thus, for each patient, the processed EMR data took the form of a drug list with time stamps.

B. MEDICATION PATTERN DISCOVERY

1) P-M MATRIX

In this process, we first established the P-D matrix for each patient and calculated the similarity between different patients using the Jaccard Index. Figure 10 is scatter diagrams of the P-D matrix from real data. In Figure 10, the vertical axis is the drug, and each drug is coded as a number. The horizontal axis is the LoS, which indicates the time that the drug was used.

2) LDA-BASED MEDICATION CLUSTERING

Based on pre-processed medication data from the EMR text, we used the LDA-based model to extract latent medication patterns. Table 1 shows the patterns discovered from this process. In this work, parameters α, β, δ, and ρ were set to 0.1, 0.01, 0.01, and 0.01, respectively, which are commonly used values in previous studies [40], [41], [51]. As for the number of patterns, we evaluated it from repetition and representation. Figure 11 is the evaluation results of repetition and representation. The horizontal axis is the number of patterns. The left vertical axis is the representation degree, and the right vertical axis is the repetition. The dotted line represents repetition, and the solid line shows representation. Repetition reflects the ratio of drugs that repetitiously appear in different clusters to the total drugs in all patterns. The higher the value, the more overlap among different clusters. The treatments for cirrhotic ascites are similar. According to the physician's advice, the number of patterns should be as small as possible. Figure 11 indicates that the repetition value is similar when the pattern number is in the range of 3 to 10. If the number of patterns continues to decrease, the repetition rapidly rises. Therefore, three patterns are an optimal solution in this context. Moreover, the representation data indicate that

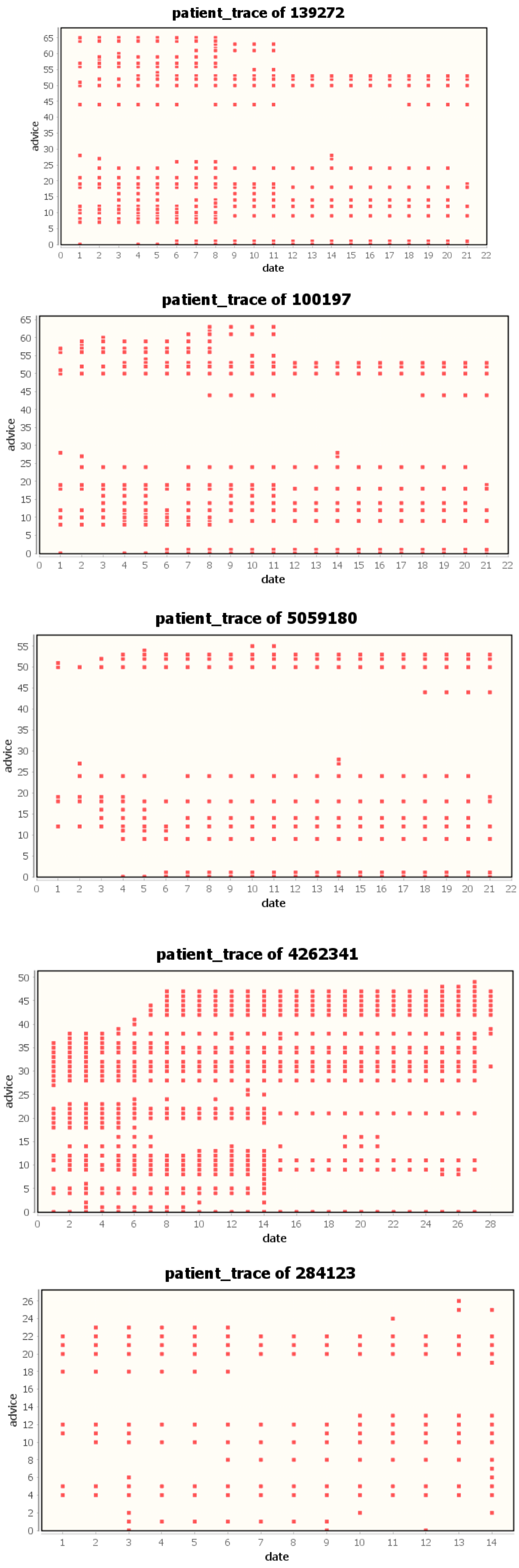


FIGURE 10. Scatter diagrams of P-M matrices.

TABLE 1. Discovered medication patterns.

Pattern	Crossing Matching Type	Medication Contents	Pattern Explanation from Clinic
Pattern No. 1	LDA3&SC(KM)2	Coenzyme Complex for Injection*, Reduced Glutathione Sodium for Injection*, Compound Glycyrrhizin Injection*, Entecavir Tablets**, Leucogen Tablets$^+$, Telbivudine Tablets**, Insulin R Injection$^+$, Compoundα-Ketoacid Tablets$^+$, Ursofalk Capsule*, Vitamin C Injection *	Liver protection: to prevent damage to the liver
	LDA3&SC(DIS)1	Coenzyme Complex for Injection*, Compound Glycyrrhizin Injection*, Reduced Glutathione Sodium For Injection*, Telbivudine Tablets**, Entecavir Tablets**, Leucogen Tablets$^+$, Compoundα-Ketoacid Tablets$^+$, Insulin R Injection$^+$, Ursofalk Capsule*, Silibin Meglumine Tablets*	
Pattern No. 2	LDA1&SC(KM)3	Entecavir Tablets**, Insulin R Injection$^+$, Ursofalk Capsule*, Leucogen Tablets$^+$, Furosemide Injection***, Torsemide Injection***, Coenzyme Complex for Injection*, Furosemide Tablets***, Potassium Chloride Sustained-release Tablets$^-$, Lactulose Oral Solution $^+$	Antiviral: treating viral infections of liver
	LDA1&SC(DIS)2	Entecavir Tablets**, Insulin R Injection$^+$, Ursofalk Capsule*, Leucogen Tablets$^+$, Furosemide Injection***, Torsemide Injection***, Coenzyme Complex for Injection*, Potassium Chloride Sustained-release Tablets $^+$, Furosemide Tablets***, Lactulose Oral Solution $^+$	
Pattern No. 3	LDA2&SC(KM)1	Furosemide Tablets***, Spironolactone Tablets ***, Esomeprazole Sodium for Injection$^+$, Entecavir Tablets**, Ursofalk Capsule*, Omeprazole Sodium for Injection $^+$, Furosemide Injection***, Gefarnate Tablets$^+$, Medium and Long Chain Fat Emulsion Injection $^+$, Human Albumin $^+$	Ease ascites: alleviate fluid accumulation
	LDA2&SC(DIS)3	Furosemide Tablets***, Spironolactone Tablets***, Esomeprazole Sodium for Injection $^+$, Entecavir Tablets**, Omeprazole Sodium for injection $^+$, Ursofalk Capsule*, Gefarnate Tablets$^+$, Furosemide Injection***, Insulin R Injection$^+$, Esomeprazole Magnesium Enteric-coated Tablets $^+$	
Mixed Patterns	LDA3&SC(KM)2 LDA1&SC(KM)3 LDA2&SC(KM)1	Coenzyme Complex for Injection Ursofalk, Capsule Reduced Glutathione Sodium for Injection, Compound Glycyrrhizin Injection, Insulin R Injection Leucogen Tablets, Torsemide Injection, Compound Amiloride Hydrochloride Tablets, Spironolactone Tablets, Irbesartan Tablets； Compound Glycyrrhizin Injection, Torsemide Injection, Ursofalk Capsule Reduced Glutathione Sodium For Injection, Cimetidine Tablets, Sodium Chloride Injection, Potassium Chloride Sustained-release Tablets, Irbesartan Tablets, Estazolam Tablets, Folic Acid Tablets；	Mixed
		Entecavir Tablets, Coenzyme Complex for Injection, Telbivudine Tablets, Spironolactone Tablets, Levofloxacin Mesylate and Sodium Chloride Injection, Compound Glycyrrhizin Injection, Torsemide Injection, Polyene Phosphatidylcholine Injection, Potassium Chloride Sustained-release Tablets, Furosemide Injection	
	LDA3&SC(DIS)1 LDA1&SC(DIS)2 LDA2&SC(DIS)3	Coenzyme Complex for Injection, Ursofalk Capsule Reduced Glutathione Sodium for Injection, Compound Glycyrrhizin Injection, Entecavir Tablets, Potassium Chloride Sustained-release Tablets, Torsemide Injection, Insulin R Injection, Ceftazidime for Injection, Spironolactone Tablets； Insulin R Injection, Ursofalk Capsule, Coenzyme Complex for Injection, Torsemide Injection, Leucogen Tablets, Ornithine Aspartate Injection, Reduced Glutathione Sodium For Injection, Compound Glycyrrhizin Injection, Compound Amiloride Hydrochloride Tablets, Vitamin K1 Injection； Reduced Glutathione Sodium for Injection, Compound Glycyrrhizin Injection, Ursofalk Capsule Spironolactone Tablets, Furosemide Tablets, Coenzyme Complex for Injection, Entecavir Tablets, Torsemide Injection, Folic Acid Tablets, Levofloxacin Mesylate and Sodium Chloride Injection	Mixed

Note: liver protection *; antiviral **; ease ascites ***; adjuvant drug $^+$

the drugs in the patterns are more representative than other cases when there are three patterns. Thus, we selected 3 as the number of patterns.

In the LDA-based method, a patient can belong to different clusters with different memberships. This is the point that we mentioned at the beginning of the paper: some methods are difficult to cluster the patients with subtle treatment variances under similar symptoms. In this case, we first clustered them into different patterns according to their highest membership. Then, in the following steps, we performed cross-matching to ensure that the distinct characteristics of the medications among different patients were focused. The cross-matching results are given in the next section.

Figure 12 shows the results of the LDA-based patient clusters. The round is composed by all patient IDs. Each line establishes the similarity link between two patients, and the lines in different colors indicate the cluster to which a patient belongs.

Figure 13 is the fingerprint map of the LDA-based patient cluster. The three bars demonstrate the patients clustering results acquired by the pure LDA method. We clustered the patients into three groups. The dark blocks indicate a high membership weight of the corresponding cluster to which a patient belongs, while the light blocks represent a low membership weight. The patients are sequenced by ID on the horizontal axis.

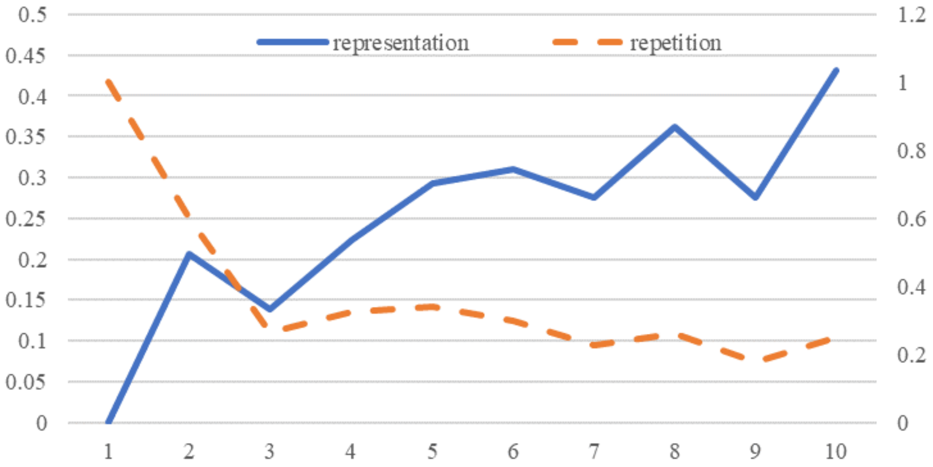


FIGURE 11. Evaluation of the number of patterns.

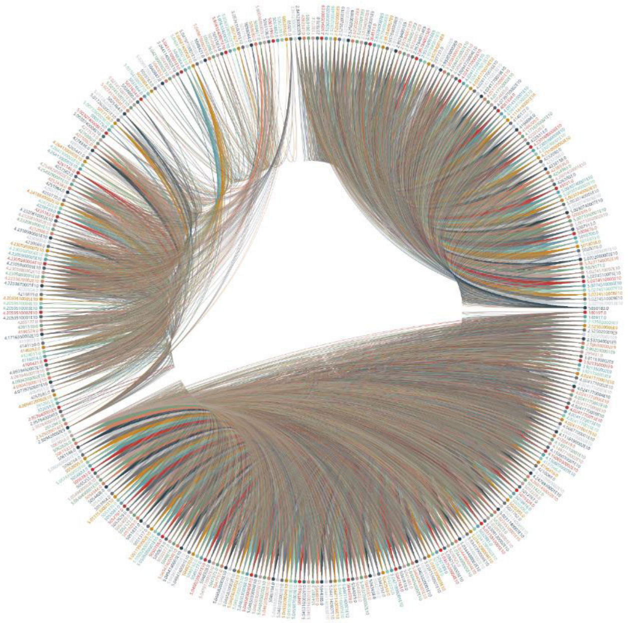


FIGURE 12. LDA-based patient cluster.

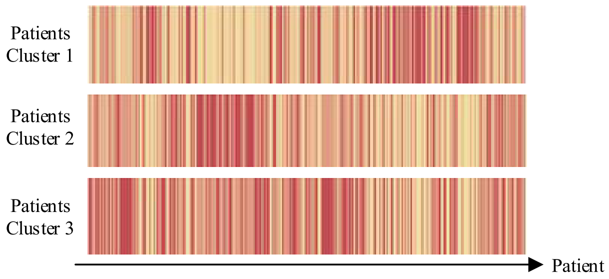


FIGURE 13. The fingerprint map of the LDA-based patient cluster.

3) SPECTRAL CLUSTERING

Based on the P-M matrix, we calculated the Jaccard similarity index and acquired the P-P matrix. Since the matrix was huge, we displayed the matrix using a heat map, as shown in Figure 14. In this figure, the deep colors represent weak similarity, and the light colors represent strong similarity.

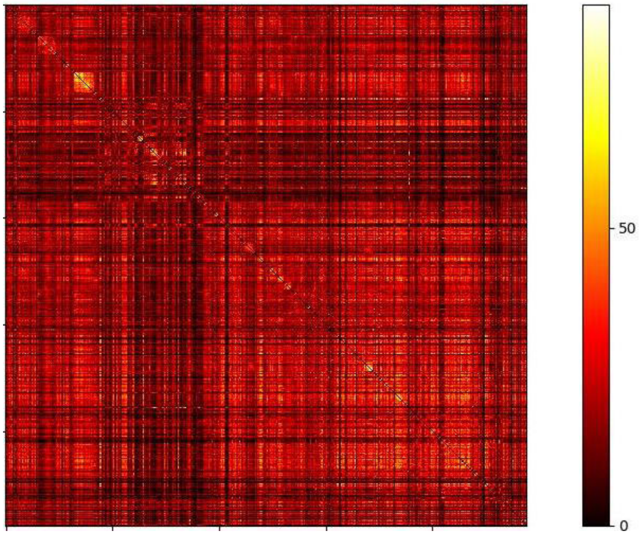


FIGURE 14. The heat map of the P-P matrix.

According to the P-P matrix and the evaluation of cluster numbers above, we clustered the patients into 3 groups using the Spectral Clustering method as shown in Figure 15.

4) CROSS MATCHING AND MEDICATION PATTERNS

Based on the data processing result in section 4.2.1 and the initial clustering results in 4.2.2 and 4.2.3, we cross-matched the features and clustering sets between the Spectral and LDA-based clustering methods to acquire the medication patterns, as shown in Table 1. We used two different methods to acquire the Spectral Clustering results. One method was K-means (KM) based, and the other method was the discretize (DIS)-based method. According to the results, the two approaches yielded similar results. Therefore, in the following discussion, we only used the results from the

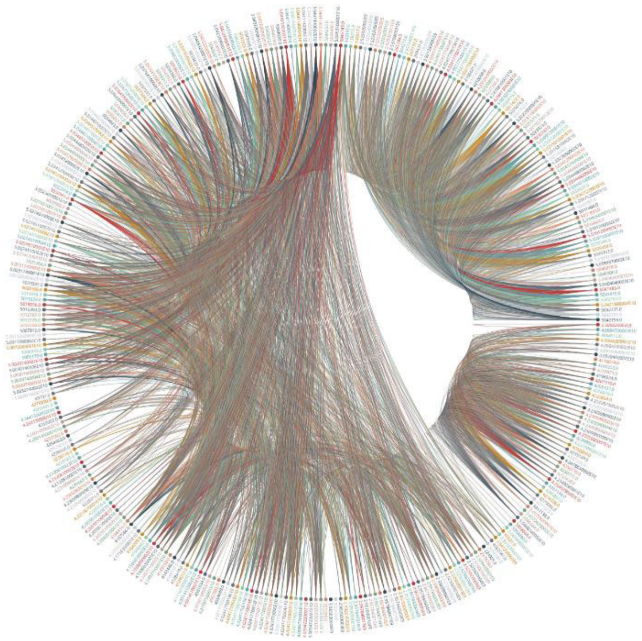


FIGURE 15. Clusters of patients.

K-means-based method marked with gray shading in Table 1. As shown in the table, we acquired three medication patterns with focused clinical meanings and discovered a previously unexplored mixed pattern. Technically, the mixed pattern represents the datasets that are only covered by one clustering method (either LDA or spectral) and represents the residual knowledge, as shown in white areas of Figure 7. From a clinical point of view, patients with relatively complex diseases require treatment with more combinations of various drugs. For these patients, the medication treatment patterns are vague and difficult to be classified into the above patterns. Note that in most of the current methods, all of the sample patients are classified into one specific group. Although this might be possible from the view of data mining, the clinical meaning for each cluster is not always clear. Our method outperforms existing methods in identifying major distinct patterns and mixed patterns. In Table 1, the names of drugs are marked with different symbols to indicate the clinical functions. Clinical functions are an important clue to identify the clinical meaning of each medication pattern. The discovered medication patterns represent drugs that provide liver protection, have antiviral activity, and ease ascites. The drug in each pattern is listed by the sequence of the use probability.

5) MEDICATION PATTERNS WITH TIME-SEQUENTIAL CHARACTERISTICS

Following the medication pattern discovery, we analyzed the time characters in these patterns. Based on the time-density-reduction method, we divided the LoS of each patient into 40 segments and calculated the medication process based on the discovered patterns. For example, the medication process of a patient with a 20-day-LoS will be divided into 40 pieces for 0.5 day is the calculation unit. Figure 15 to Figure 17 show the medication conditions for each pattern with time characteristics. The vertical axis indicates the percentage of

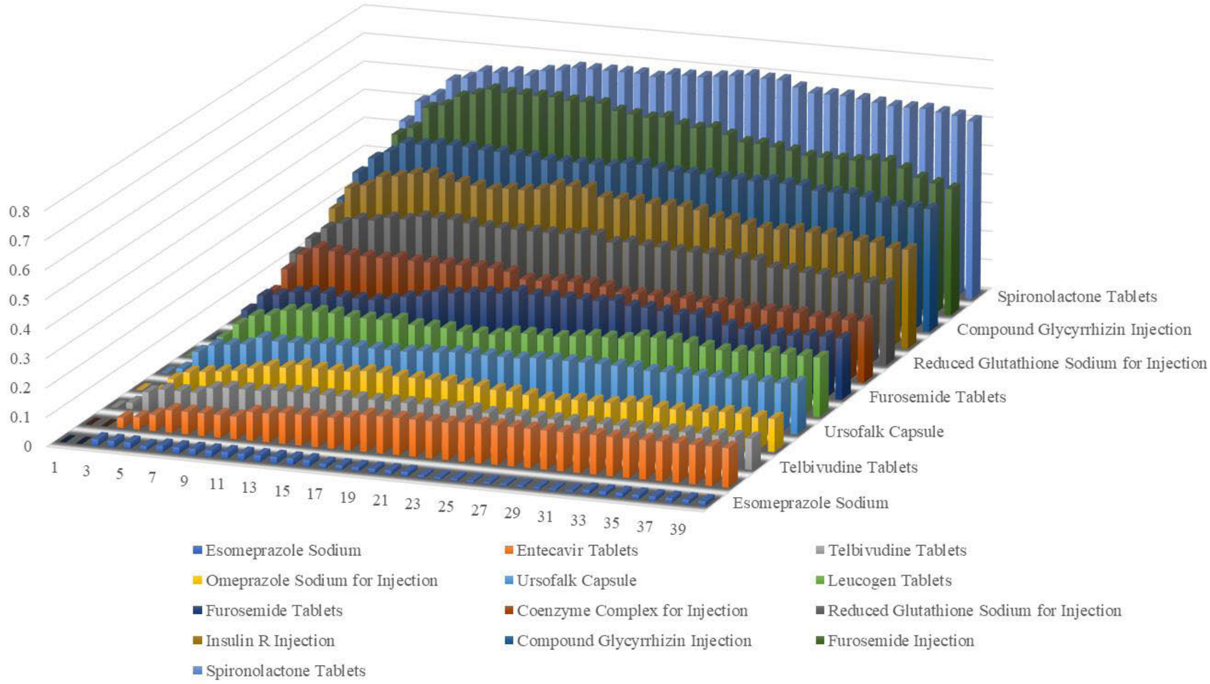


FIGURE 16. Drug usage and time sequential conditions of patients in pattern 1.

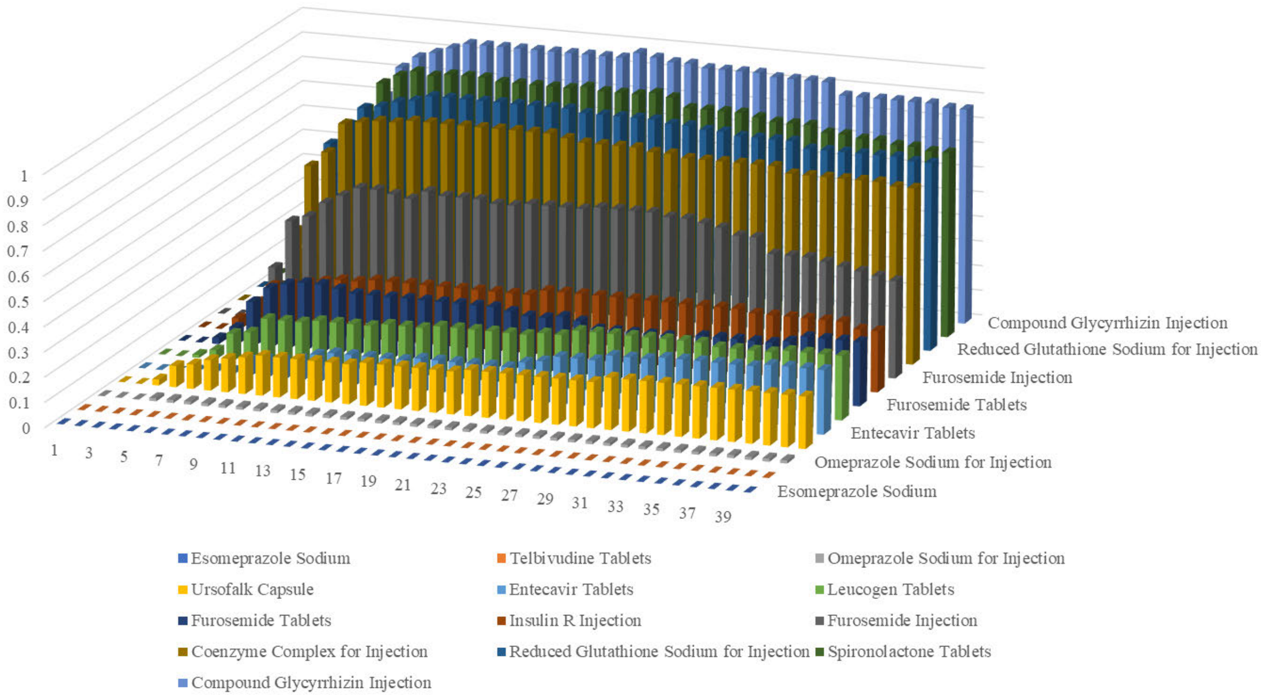


FIGURE 17. Drug usage and time sequential conditions of patients in pattern 2.

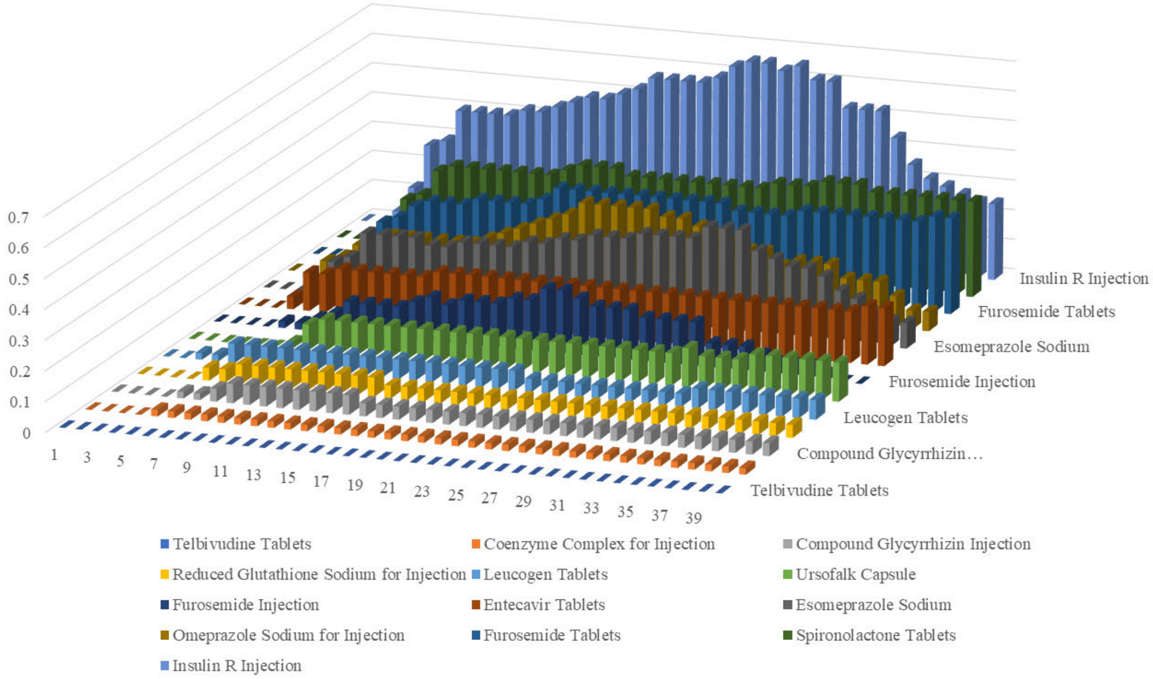


FIGURE 18. Drug usage and time sequential conditions of patients in pattern 3.

patients in the corresponding pattern. The patient number here is not the absolute amount but is amplified by the time-density-reduction method.

As noted above, due to the time-density-reduction method, some drugs that appear in Figures 16-18 were not included in the patterns in Table 1. According to the 'time-density-

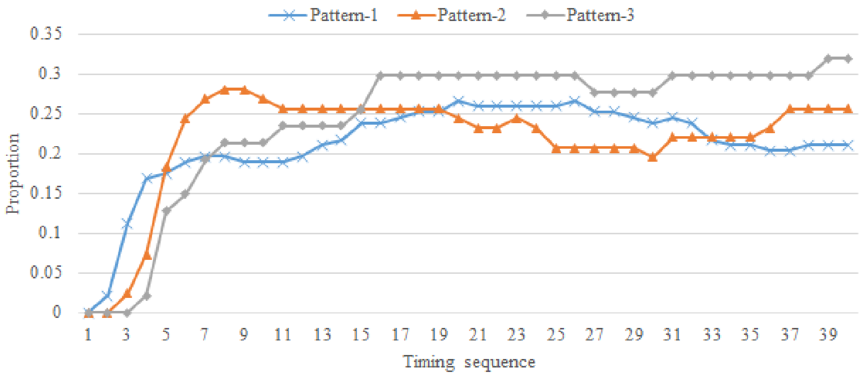


FIGURE 19. The drug usage characteristics among different patterns during the treatment process (Furosemide Tablets).

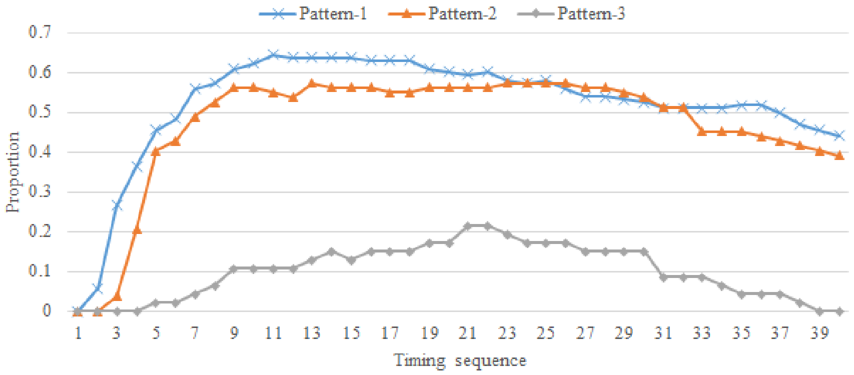


FIGURE 20. The drug usage characteristics among different patterns during the treatment process (Furosemide Injections).

reduction versus importance' phenomenon, these drugs are the drugs often used by patients with short LoSs. For instance, in Figure 16, Spironolactone Tablets are the drug used by most of the amplified patients (patients with extended LoSs). However, Spironolactone is not included in medication Pattern 1 in Table 1. Therefore, we concluded that most of the patients who belong to Pattern 1 and are given Spironolactone Tablets have short LoSs, and this medicine is more important for patients in Pattern 1.

According to the medication characteristics in each pattern, we chose 6 typical drugs (Furosemide Tablets, Furosemide Injection, Ursofalk Capsule, Leucogen Tablets, Insulin R Injection, and Coenzyme Complex for Injection) to analyze the medication patterns further. The reason for selecting these 6 drugs is that they appeared in more than one pattern, and all of them have high usage probabilities. Figures 19 through 23 show the analysis results where the vertical axis represents the proportion of patients using the corresponding drug, and the horizontal axis represents the unified LoSs (which can be understood as timing sequences).

Figure 19 and Figure 20 are both the usage characteristics of Furosemide, which is a diuretic drug. Figure 19 shows the results of Furosemide Tablets, and Figure 20 shows Furosemide Injections. In Figure 19, the drug usage characteristics are similar, but minor differences exist in the different patterns. In Figure 20, significant differences appear among the patterns. The proportions of patients who use Furosemide Injections in Pattern 1 and Pattern 2 are higher than those in Pattern 3. Furosemide is a diuretic drug, but patients who need the treatment to alleviate ascites use less Furosemide. One reasonable explanation is that Furosemide Injections are stronger than Furosemide Tablets. Therefore, the Furosemide Injection is always used as short-term and effective medication treatment for patients without serious ascites. Furthermore, patients with serious ascites take Furosemide Tablets for long-term and surgical treatments.

Figure 21 illustrates the drug use characteristics of the Coenzyme Complex for Injections. Clearly, huge differences exist in these patterns. The high usage proportion indicates that the Coenzyme Complex for Injections is a very important drug for the patients in Pattern 2. According to the function of the Coenzyme Complex for Injections, it is mainly used for curing hepatitis, and the results in Figure 21 are completely consistent with the clinical meaning of Pattern 2.

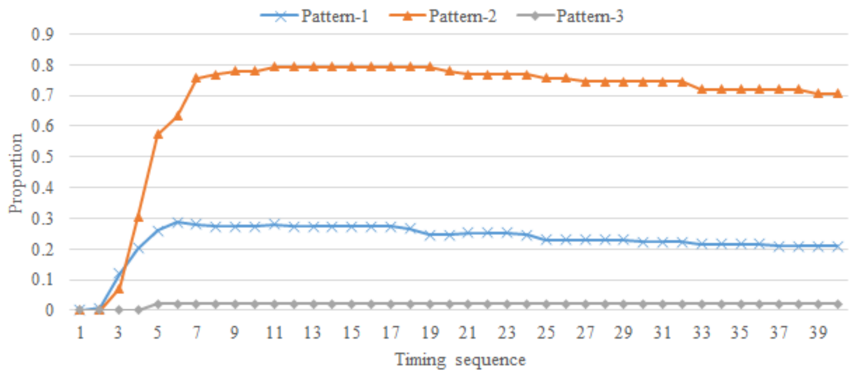


FIGURE 21. The drug usage characteristics among different patterns during the treatment process (Coenzyme Complex for Injections).

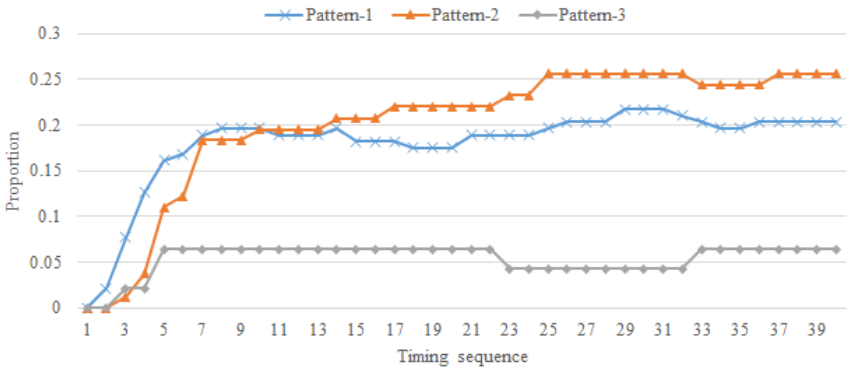


FIGURE 22. The drug usage characteristics among different patterns during the treatment process (Leucogen Tablets).

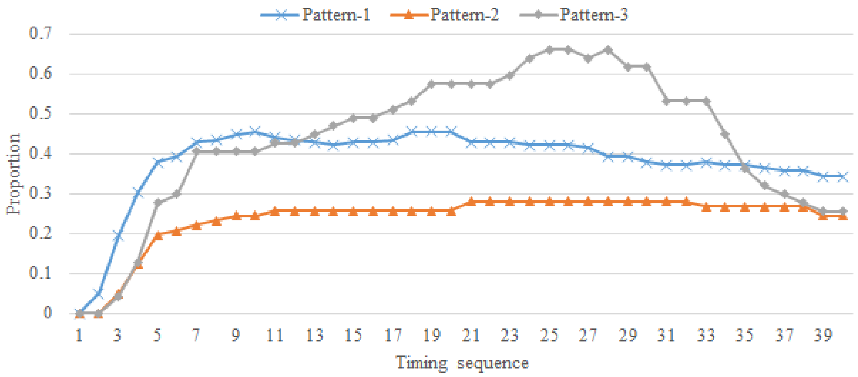


FIGURE 23. The drug usage characteristics among different patterns during the treatment process (Insulin R Injections).

Figure 22 and Figure 23 show the drug usage of Leucogen Tablets and Insulin R Injections. Both of them are commonly used for patients with cirrhotic ascites, especially for complications during the medication treatment. These results reveal that Leucogen Tablets are more important in Pattern 1 and Pattern 2, and Insulin R Injections are more important for patients in Pattern 3, especially in the middle and later treatment periods.

6) EVALUATION OF THE MEDICATION PATTERN DISCOVERING FRAMEWORK

Based on the experiment above, we evaluated the patterns discovered from the EMR text using the proposed framework. Figure 24 is a comparison of the performances of different medication mining methods. The bars in this figure are the clustering results acquired by the pure LDA, pure Spectral Clustering, and the proposed frameworks, respectively.

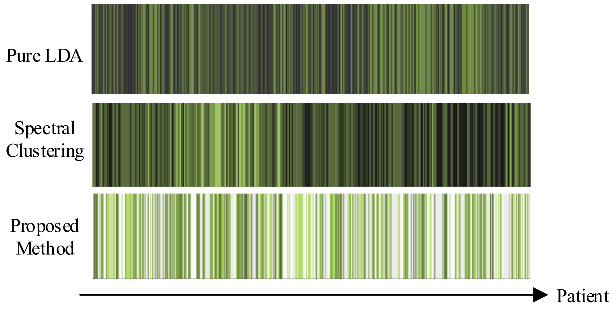


FIGURE 24. Comparison of the results from different clustering methods.

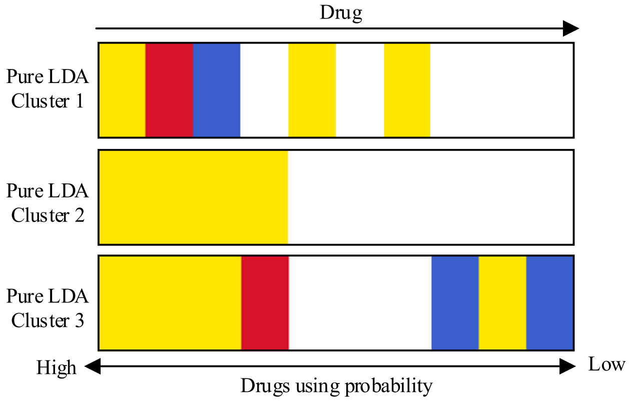


FIGURE 25. Drug use clustered by pure LDA.

We used light green, dark green and black blocks to represent the patients in Pattern 1, Pattern 2, and Pattern 3, respectively. The patients are sequenced by their IDs on the horizontal axis. The first bar represents the LDA clustering results based on the membership weight shown in Figure 13. The second bar shows the results of Spectral Clustering. The last one shows the results of our proposed method in which the white blocks represent the patients filtered by cross-matching the features and clustering sets between the Spectral and LDA methods. Obviously, approximately half of the patients have similar memberships in different clusters, which are difficult to separate. Cross-matching plays an important role in filtering the confusing items from the whole data set.

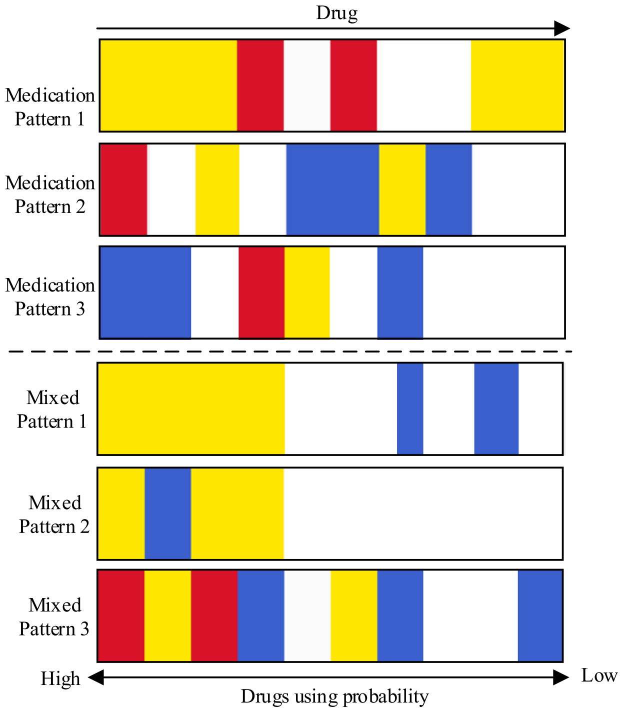


FIGURE 26. Drug use for each medication pattern.

The key goal of the proposed method was to discover medication patterns. Figure 25 and Figure 26 are the comparisons of the drugs from different patterns acquired by the LDA and proposed methods. In these two figures, we classified the drugs into four function groups represented by different color blocks. Those groups represent liver protection (yellow, one star in Table 1), antiviral (red, two stars in Table 1), ascites treatment (blue, three stars in Table 1), and other functions (white, no star in Table 1). For the patterns discovered by the pure LDA methods in Figure 25, we see that it is difficult to distinguish the functional patterns. All three patterns have high probabilities to use the drugs that function to protect

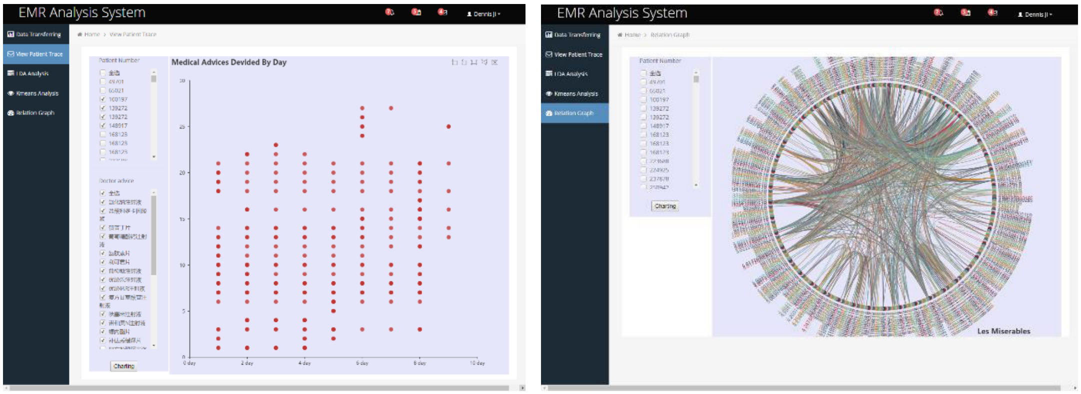


FIGURE 27. Prototype system of mining EMR text.

the liver. This indicates that the characteristics of the patterns are not distinct and obvious. In the results acquired by the proposed method in Figure 26, each discovered medication pattern has a distinct focus on drug functions, starting with different color blocks. The starting blocks indicate the drug function with the highest use probability in the corresponding pattern. By using cross-matching, some medication treatments with similar drug use are separated. We further analyzed these patients and acquired the distribution of drug use probability which is shown as the last three bars in Figure 26. The filtered medications have mixed drug use conditions.

V. DISCUSSION AND CONCLUSIONS

Based on the real EMR textual data from patients with cirrhotic ascites, we proposed a medication pattern discovery framework to help physicians make clinical decisions based on medication patterns. This study effectively identified the medication patterns from EMRs, especially for diseases with minor treatment variances from complex medication treatments. For example, ascites cirrhosis is a common disease, but the clinical practice guidelines released by EASL (European Association for the Study of the Liver) [49] are different from AASLD (American Association for the Study of Liver Diseases) [52]. Both of these associations give professional and official clinical guidelines in their respective regional area. The question is which one performs better for which groups of patients. Regardless of the type of illness, similar questions also arise among different hospitals and physicians. Every hospital may have its own specific clinical guidelines or clinical pathways based on authoritative doctors. Moreover, each physician probably has his/her own clinical treatment style under the guideline. In consideration of those variances, in recent years, physicians and researchers have increasingly focused on personalized pathways to acquire better curative effects [53], [54]. Physicians acknowledge the importance of standard or reference clinical pathways, but they also argue that more attention should be paid to the most suitable treatment for a specific patient rather than a universal standard.

Current existing methods can hardly classify the latent drug use patterns in the above scenarios. The proposed method is based on the Jaccard index, Spectral Clustering and modified LDA method. By cross-matching, the clustered results are well focused on the specific medication patterns with little overlap. This framework filters some medication treatments with high mixed drug usage and then clusters them into a distinct single group. It is different from the current methods in which all sampling data have to be classified into one certain group.

Mining medication treatment patterns is helpful in the process of predicting clinical history, and medication treatments are an important component in clinical pathway development. In the ongoing and following future work, we plan to address the following. First we will introduce more types of data into the clustering process, such as the examination results during the LoS, which can evaluate the effects of corresponding medication processes and mine the best-fit treatment pattern for corresponding patient groups. Second, in this paper, we only consider the medications in the treatment process. In the future, we will consider more kinds of treatment contents, such as clinical examinations, bedside operations, and even nursing contents. These contents are all necessary components of the clinical pathway. Third, we are working on a clinical decision support system that integrates the methods that we have developed. The input data of this system is EMR text, and the outcomes are the medication knowledge hidden in the EMR text. We performed initial work (as shown in Figure 27) based on the proposed method in this paper. One specific format of the decision support system is that a facility needs to be able to automatically recommend clinical pathways with corresponding analyzed data and confidence degrees based on the imported EMR data.

REFERENCES

[1] R. B. Haynes, D. L. Sackett, W. S. Richardson, W. Rosenberg, and G. R. Langley, "Evidence based medicine-how to practice and teach EBM," *Can. Med. Assoc. J.*, vol. 157, no. 6, p. 788, Sep. 1997.

[2] J.-E. Bibault, P. Giraud, and A. Burgun, "Big data and machine learning in radiation oncology: State of the art and future prospects," *Cancer Lett.*, vol. 382, no. 1, pp. 110–117, Nov. 2016.

[3] C. Chen, M. He, Y. Zhu, S. Lin, and X. Wang, "Five critical elements to ensure the precision medicine," *Cancer Metastasis Rev.*, vol. 34, no. 2, pp. 313–318, Jun. 2015.

[4] U.S. Government. (Feb. 2006). Personal health records and personal health record systems. WDC, San Jose, CA, USA. Accessed: Apr. 2012. [Online]. Available: https://www.ncvhs.hhs.gov/wp-content/uploads/2014/05/0602nhiirpt.pdf

[5] L. A. Seaborne, K. Hueneberg, A. Bohler, G. Schroeder, and T. White, "Developing electronic health record (EHR)-based program to deliver survivorship care plans (SCPS) and visits at the UW breast Center," *J. Clin. Oncol.*, vol. 34, p. 56, Apr. 2016.

[6] K. U. Kortüm, M. Müller, C. Kern, A. Babenko, and J. Wolfgang, "Using electronic health records to build an ophthalmologic data warehouse and visualize patients' data," *Amer. J. Ophthalmol.*, vol. 178, pp. 84–93, Jun. 2017.

[7] T. Delespierre, P. Denormandie, A. Bar-Hen, and L. Josseran, "Empirical advances with text mining of electronic health records," *BMC Med. Inform. Decis. Making*, vol. 17, no. 1, p. 127, Dec. 2017.

[8] L. Pan, G. Liu, F. Lin, S. Zhong, H. Xia, X. Sun, and H. Liang, "Machine learning applications for prediction of relapse in childhood acute lymphoblastic leukemia," *Sci. Rep.*, vol. 7, no. 1, Aug. 2017, Art. no. 7402.

[9] O. E. Streeter, P. J. Beron, and P. N. Iyer, "Precision medicine: Genomic profiles to individualize therapy," *Otolaryngologic Clinics North Amer.*, vol. 50, no. 4, pp. 765–773, Aug. 2017.

[10] P. Hamet and J. Tremblay, "Artificial intelligence in medicine," *Met. Clin. Exp.*, vol. 69, pp. S36–S40, Sep. 2017.

[11] F. Wong, "Management of ascites in cirrhosis," *J. Gastroenterol. Hepatol.*, vol. 27, no. 1, pp. 11–20, Jan. 2012.

[12] Z. C. Lipton, "The mythos of model interpretability," 2016, *arXiv:1606.03490*. [Online]. Available: https://arxiv.org/abs/1606.03490

[13] Y. S. Kim, D. Yoon, J. Byun, H. Park, A. Lee, H. I. Kim, S. Lee, H. S. Lim, and R. W. Park, "Extracting information from free-text electronic patient records to identify practice-based evidence of the performance of coronary stents," *Plos One*, vol. 12, no. 8, Aug. 2017, Art. no. e0182889.

[14] S. V. Pakhomov, S. A. Weston, S. J. Jacobsen, C. G. Chute, and R. Meverden, "Electronic medical records for clinical research: Application to the identification of heart failure," *Amer. J. Managed Care*, vol. 13, no. 6, pp. 281–288, Jun. 2007.

[15] K. P. Liao, A. N. Ananthakrishnan, V. Kumar, Z. Xia, A. Cagan, V. S. Gainer, S. Goryachev, P. Chen, G. K. Savova, D. Agniel, S. Churchill, "Methods to develop an electronic medical record phenotype algorithm to compare the risk of coronary artery disease across 3 chronic disease cohorts," *Plos One*, vol. 10, no. 8, Aug. 2015, Art. no. e0136651.

[16] K. P. Liao, T. Cai, G. K. Savova, S. N. Murphy, E. W. Karlson, A. N. Ananthakrishnan, V. S. Gainer, S. Y. Shaw, Z. Xia, P. Szolovits, and S. Churchill, ''Development of phenotype algorithms using electronic medical records and incorporating natural language processing,'' *BMJ Brit. Med. J.*, vol. 350, p. h1885, Apr. 2015.

[17] W. W. Yim, M. Yetisgen, W. P. Harris, and S. W. Kwan, ''Natural language processing in oncology: A review,'' *JAMA Oncol.*, vol. 2, no. 6, pp. 797–804, Jun. 2016.

[18] E. Pons, L. Braun, M. G. Hunink, and J. A. Kors, ''Natural language processing in radiology: A systematic review,'' *Radiology*, vol. 279, no. 2, pp. 329–343, Apr. 2016.

[19] R. L. Figueroa and C. A. Flores, ''Extracting information from electronic medical records to identify the obesity status of a patient based on comorbidities and bodyweight measures,'' *J. Med. Syst.*, vol. 40, no. 8, pp. 1–9, Aug. 2016.

[20] T. Cai, A. A. Giannopoulos, S. Yu, T. Kelil, B. Ripley, K. K. Kumamaru, F. J. Rybicki, and D. Mitsouras, ''Natural language processing technologies in radiology research and clinical applications,'' *Radiographics*, vol. 36, no. 1, pp. 176–191, Jan. 2016.

[21] P. C. Chang, M. Galley, and C. D. Manning, ''Optimizing Chinese word segmentation for machine translation performance,'' in *Proc. WSMT-ACL*, Columbus, OH, USA, Jun. 2008, pp. 224–232.

[22] Z. Huang, T.-M. Chan, and W. Dong, ''MACE prediction of acute coronary syndrome via boosted resampling classification using electronic medical records,'' *J. Biomed. Inform.*, vol. 66, pp. 161–170, Feb. 2017.

[23] D. Hu, Z. Huang, T.-M. Chan, W. Dong, X. Lu, and H. Duan, ''Utilizing chinese admission records for mace prediction of acute coronary syndrome,'' *Int. J. Environ. Res. Public Health*, vol. 13, no. 9, p. 912, Sep. 2016.

[24] Y. Yang, Y. Cai, W. Luo, Z. Li, Z. Ma, X. Yu, and H. Yu, ''An ontology-based approach for text mining of stroke electronic medical records,'' in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, pp. 288–291, Feb. 2014.

[25] R. Henao, J. T. Lu, J. E. Lucas, J. Ferranti, and L. Carin, ''Electronic health record analysis via deep Poisson factor models,'' *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 6422–6453, Apr. 2016.

[26] P. Nguyen, T. Tran, N. Wickramasinghe, and S. Venkatesh, ''Deepr: A convolutional net for medical records,'' *IEEE J. Biomed. Health.*, vol. 21, no. 1, pp. 22–30, Jan. 2017.

[27] Y. Cheng, F. Wang, P. Zhang, and J. Hu, ''Risk prediction with electronic health records: A deep learning approach,'' in *Proc. ICDM-SIAM*, Atlantic City, NJ, USA, Jun. 2016, pp. 432–440.

[28] R. Miotto, L. Li, and J. T. Dudley, ''Deep learning to predict patient future diseases from the electronic health records,'' in *Proc. ECIR*, Padua, Italy, 2016, pp. 768–774.

[29] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzell, ''Learning to diagnose with LSTM recurrent neural networks,'' in *Proc. ICLR*, San Juan, PR, USA, 2016. [Online]. Available: https://arxiv.org/abs/1511.03677

[30] J. M. Liu, M. You, Z. Wang, G. Li, X. Xu, and Z. Qiu, ''Cough event classification by pretrained deep neural network,'' *BMC Med. Inform. Decis.*, vol. 15, no. 4, pp. 1–10, Dec. 2015.

[31] T. Zheng, W. Xie, L. Xu, X. He, Y. Zhang, M. You, G. Yang, and Y. Chen, ''A machine learning-based framework to identify type 2 diabetes through electronic health records,'' *Int. J. Med. Inform.*, vol. 97, pp. 120–127, Jan. 2017.

[32] A. Rajkomar, J. W. L. Yim, K. Grumbach, and A. Parekh, ''Weighting primary care patient panel size: A novel electronic health record-derived measure using machine learning,'' *JMIR Med. Inform.*, vol. 4, no. 4, p. e29, Dec. 2016.

[33] C. Li, S. Rana, D. Phung, and S. Venkatesh, ''Hierarchical Bayesian nonparametric models for knowledge discovery from electronic medical records,'' *Knowl.-Based Syst.*, vol. 99, pp. 168–182, May 2016.

[34] L. Yin, Z. Huang, W. Dong, C. He, and H. Duan, ''Utilizing electronic medical records to discover changing trends of medical behaviors over time,'' *Method. Inform. Med.*, vol. 56, no. S01, pp. e49–e66, Jan. 2017.

[35] Z. Huang, X. Lu, and H. Duan, ''Latent treatment pattern discovery for clinical processes,'' *J. Med. Syst.*, vol. 37, no. 2, pp. 1–10, Apr. 2013.

[36] Z. Huang, W. Dong, L. Ji, C. Gan, X. Lu, and H. Duan, ''Discovery of clinical pathway patterns from event logs using probabilistic topic models,'' *J. Biomed. Inform.*, vol. 47, pp. 39–57, Feb. 2014.

[37] Z. Huang, W. Dong, J. Lei, C. He, and H. Duan, ''Incorporating comorbidities into latent treatment pattern mining for clinical pathways,'' *J. Biomed. Inform.*, vol. 59, pp. 227–239, Feb. 2016.

[38] C. Hennig, ''Cluster-wise assessment of cluster stability,'' *Comput. Stat. Data Anal.*, vol. 52, no. 1, pp. 258–271, Sep. 2007.

[39] J. Shi and J. Malik, ''Normalized cuts and image segmentation,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[40] D. M. Blei, A. Y. Ng, and M. I. Jordan, ''Latent Dirichlet allocation,'' *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[41] Z. Huang, W. Dong, and H. Duan, ''A probabilistic topic model for clinical risk stratification from electronic health records,'' *J. Biomed. Inform.*, vol. 58, pp. 28–36, Dec. 2015.

[42] G. Heinrich, ''Parameter estimation for text analysis,'' Fraunhofer Inst. Comput. Graph. Res., Darmstadt, Germany, Tech. Rep., Sep. 2009.

[43] M. Magnusson, L. Jonsson, M. Villani, and D. Broman, ''Parallelizing LDA using partially collapsed Gibbs sampling,'' *Statistic*, vol. 24, pp. 301–327, 2015.

[44] J. Chang, S. M. Gerrish, C. Wang, J. Boydgraber, and D. M. Blei, ''Reading tea leaves: How humans interpret topic models,'' in *Proc. NIPS*, Vancouver, BC, Canada, 2009, pp. 288–296.

[45] J. Chuang, S. Gupta, C. D. Manning, and J. Heer, ''Topic model diagnostics: Assessing domain relevance via topical alignment,'' in *Proc. Intern. Conf. Mach. Learn.*, vol. 28, no. 3, 2013, pp. 612–620.

[46] T. E. Sweeney, A. C. Chen, and O. Gevaert, ''Combined mapping of multiple clUsteriNg ALgorithms (COMMUNAL): A robust method for selection of cluster number, K,'' *Sci. Rep.*, vol. 5, Nov. 2015, Art. no. 16971.

[47] M. D. Wilkerson and D. N. Hayes, ''ConsensusClusterPlus: A class discovery tool with confidence assessments and item tracking,'' *Bioinformatics*, vol. 26, no. 12, pp. 1572–1573, Jun. 2010.

[48] M. Laan and K. Pollard, ''A new algorithm for hybrid clustering of gene expression data with visualization and the bootstrap,'' *J. Stat. Plan. Infer.*, vol. 117, pp. 275–303, May 2003.

[49] P. Ginés, P. Angeli, K. Lenz, S. Müller, and K. Moore, ''EASL clinical practice guidelines on the management of ascites, spontaneous bacterial peritonitis, and hepatorenal syndrome in cirrhosis,'' *J. Hepatol.*, vol. 53, no. 3, pp. 397–417, Sep. 2010.

[50] P. Gines, E. Quintero, V. Arroyo, J. Terés, and M. Bruguera, ''Compensated cirrhosis: Natural history and prognostic factors,'' *Hepatology*, vol. 7, no. 1, pp. 122–128, Feb. 1987.

[51] T. L. Griffiths and M. Steyvers, ''Finding scientific topics,'' in *Proc. NAS*, Apr. 2004, pp. 5228–5235.

[52] A. M. Runyon, ''Management of adult patients with ascites due to cirrhosis: An update,'' *Hepatology*, vol. 49, no. 6, pp. 2087–2107, May 2009.

[53] G. Fico, A. Fioravanti, M. Arredondo, J. Gorman, C. Diazzi, G. Arcuri, C. Conti, and G. Porini, ''Integration of personalized healthcare pathways in an ICT platform for diabetes management: A small-scale exploratory study,'' *IEEE J. Biomed. Health*, vol. 20, no. 1, pp. 29–38, Nov. 2014.

[54] L. Mertz, ''Ready or not: Personalized medicine is coming. IEEE pulse talks with michael snyder about its potential,'' *IEEE Pulse*, vol. 5, no. 3, pp. 45–47, May 2014.

HUIQUN HUANG was born in Guangxi, China, in 1996. She is currently pursuing the undergraduate degree with the School of Software Engineering, Beijing Jiaotong University.

Her research interests include text mining, knowledge representation, information retrieval, and recommendation.

XIAOPU SHANG received the B.S. degree from PLA Information Engineering University, China, in 2009, and the Ph.D. degree from Beijing Jiaotong University, in 2015.

He is currently a Lecturer with Beijing Jiaotong University. His research interests include health informatics, data-driven healthcare management, information technology and society, and decision making.

HONGMEI ZHAO received the B.S. and M.S. degrees from Peking University and Beijing University of Chinese Medicine, Beijing, China, in 2013 and 2015, respectively.

She is currently pursuing the Ph.D. degree with Beijing Jiaotong University. She is currently an Associate Researcher with Peking University People's Hospital. Her research interests include health informatics and data-driven healthcare management.

NAN WU received the B.S. degree from China Medical University, Shenyang, Liaoning, China, in 2003, and the Ph.D. degree from the Peking University Health Science Center, Beijing, China, in 2008.

She is currently an attending Physician with Peking University People's Hospital. Her research interests include health informatics and data-driven healthcare management.

WEIZI LI received the Ph.D. degree from the Beijing Institute of Technology.

She is currently an Associate Professor of informatics and digital health, the Deputy Director of Informatics Research Centre, Henley Business School, University of Reading. Her research interests include digital health, integrated systems, artificial intelligence, and machine learning applications in healthcare. She is also a Fellow of Charted Institute of IT (FBCS).

YUAN XU received the B.S. degree from Hebei University, Baoding, Hebei, China, in 2017.

She is currently pursuing the master's degree with the School of Economic and Management, Beijing Jiaotong University, Beijing, China. Her current research interests include group decision making, aggregations, health informatics, and data-driven healthcare management.

YANG ZHOU was born in Hangzhou, China, in 1990. He received the bachelor's degree in information management from Beijing Jiaotong University, in 2012, where he is currently pursuing the Ph.D. degree in information management with the School of Economics and Management.

His research interests include (deep) machine learning, NLP, and big data applications.

LEI FU was born in Gansu, China, in 1982. He received the Ph.D. degree in public health from the Chinese PLA Logistics College, Beijing, in 2012.

Since 2014, he has joined Core Laboratory of Translational Medicine, Chinese PLA General Hospital. His research interests include public health, hospital management, and medical informatics.