

Chapter 2

Big Data Analytics in Healthcare



Chonghui Guo and Jingfeng Chen

2.1 Big Data-Driven Paradigm

The cross-integration of information technology and economic society has led to the rapid growth of data, which has become a basic national strategic resource. Big data is increasingly exerting an important influence on global production, circulation, distribution, consumption activities, economic operation mechanism, social lifestyle, and national governance capacity (Chen et al., 2012; Ji et al., 2017; Lynch, 2008; Naeem et al., 2022; van Elten et al., 2022). In the context of big data, the advantages of the data-driven paradigm are constantly highlighted. Generally speaking, the big data-driven paradigm is described from three aspects: external embedding, technology augmentation, and enabled innovation, reflecting a “correlation + causality” viewpoint in a “data-driven + model-driven” manner (Bakker & Tsui, 2017; Chen, Wu, et al., 2018). Recently, governments, academics, and industries around the world have promoted the research and application of big data to an unprecedented height. In 2008 and 2011, Nature and Science published a special issue on big data respectively, discussing the challenges of big data from multiple perspectives. In 2014, *The Bridge*, the journal of the American Academy of Engineering, organized a special issue to discuss the current situation, challenges, and future trends of big data from the perspective of globalization (Shi, 2014).

C. Guo

Institute of Systems Engineering, Dalian University of Technology, Dalian, China

e-mail: dlutguo@dlut.edu.cn

J. Chen (✉)

Health Management Center, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China

School of Public Health, Zhengzhou University, Zhengzhou, China

e-mail: fccjfchen@zzu.edu.cn

As an important strategic resource, big data contains many key management issues and has its own management characteristics. And in the big data environment, the existing management models should also have further development. The paradigm of scientific research is also shifting to the “data-intensive” fourth paradigm, which fosters research into scientific data management, data analysis, data visualization, and new algorithms and tools (Hey et al., 2009).

2.1.1 The Research Background of Big Data Analytics in Healthcare

Along with the development and popularization of cloud computing, the Internet, various mobile devices, and the Internet of Things, big data analytics has been one of the current and future research frontiers (Chen et al., 2012; Haque et al., 2020). In the medical field, Mayer-Schönberger and Cukier (2013) elaborated on the reforms from two aspects. One is to provide help for the rapid improvement of the collective medical experience of human beings, which will make everyone become the master of their own diseases, and the other is that inexhaustible medical data innovation is dominant, bringing industrial effects with great commercial value.

However, big data analytics in healthcare, in general, lags behind e-commerce business intelligence and analytics applications because it has rarely taken advantage of scalable analytical methods or computational platforms (Miller, 2012). Fortunately, along with the construction and development of healthcare informatization, medical institution informatization, regional medical informatization, and internet plus medical, healthcare, as an important field of big data & big data analytics, is entering a “big data era.” In the clinical sphere, the amount of patient data has grown exponentially because of new computer-based information systems, including clinical data (electronic health records (EHRs), electronic medical records (EMRs), electronic patient records (EPRs), etc.), claims and cost data, pharmaceutical R & D data, and patient behavior data (Groves et al., 2013).

The release of big data analytics in healthcare is transforming the discussion of what is appropriate or right for a patient and right for the healthcare ecosystem, and further changing the paradigm by achieving the new value pathways, as follows. (1) Right living: patients should take more active steps to improve their health; (2) Right care: developing a coordinated approach to care in which all caregivers have access to the same information. (3) Right provider: any professionals who treat patients must have strong performance records and be capable of achieving the best outcomes; (4) Right value: improving value while simultaneously improving care quality; (5) Right innovation: identifying new approaches to healthcare delivery (Groves et al., 2013; Guo & Chen, 2019).

The transformation of the medical paradigm is also accelerating the revolution of the medical model, from one-size-fits-all medicine and stratified medicine to precision medicine, from the bio-psycho-social medical model to the “4P” medical

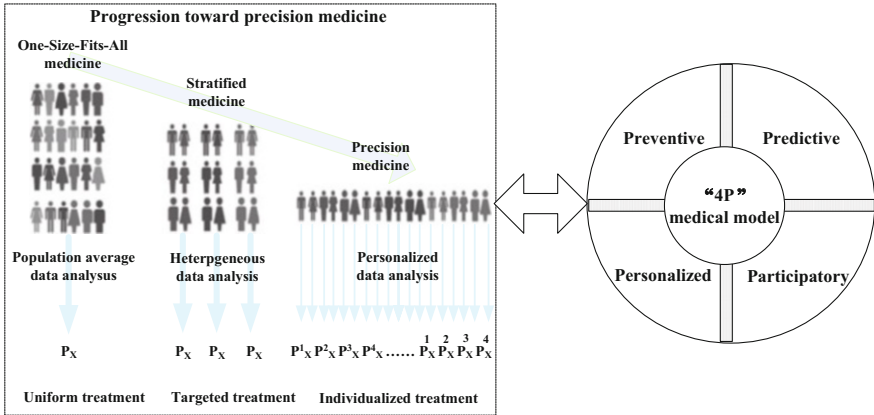


Fig. 2.1 Precision medicine and the “4P” medical model (the left part comes from Hopp et al. (2018))

model, as shown in Fig. 2.1. Specifically, one-size-fits-all medicine requires population average effect analysis, and all patients adopt the uniform treatment; stratified medicine divides patients into groups according to their response to therapy and uses heterogeneous effect analysis to correct for the failure of average effect analysis to account for patient differences; while precision medicine uses personalized effect analysis, which often requires personalized data. In addition, precision medicine, sometimes known as “personalized medicine,” is an innovative approach to tailoring disease prevention and treatment by considering differences in people’s characteristics, environments, and lifestyles. Thus, the goal of precision medicine is to target the right treatments for the right patients at the right time (Hopp et al., 2018).

During the process of achieving the goal of precision medicine, the “4P” medical model is emerging. The “4P” medical model refers to preventive, predictive, personalized, and participatory, emphasizing prevention first, predictive treatment, individualized diagnosis and treatment, and then public participation. It provides patients with a physician guide to medical science as a tool for living healthier, happier, and more productive lives. In the case of the “4P” medical model, it is the power to predict and prevent disease, feel good, slow or even partially reverse biological aging, and optimize patients’ ability to move, think, and perform at patients’ best in all aspects of life, environment, mind, and body (Auffray et al., 2009; Bricage, 2017; Sun et al., 2019; Topol, 2015; Wu et al., 2015).

The application and development of big data in healthcare will promote profound revolutions in the medical service model and greatly improve the quality and efficiency of healthcare services. The application of big data and big data analytics in healthcare will improve healthcare quality, long-term care, and patient empowerment, and using this information and knowledge to analyze the efficacy of clinical diagnosis and treatment and healthcare decision support will bring revolutionary reforms to the medical industry (Chen et al., 2012, 2020).

2.1.2 The Research Framework of Big Data Analytics in Healthcare

Healthcare big data not only have the 4 V (volume, variety, value, and velocity) characteristics of big data, but also high dimensionality, heterogeneity, and relational complexity among data objects. Thus, the existing hypothesis-driven research and reductionist approaches to causality have no capability to adjust for confounding and modifying factors in clinical practice. In recent years, some popular research frameworks or the modeling processes of big data analytics in healthcare have been proposed to promote the transformation from data to knowledge. For example, in a data-intensive healthcare environment, Hey et al. (2009) proposed a unified modeling approach that can take full advantage of a data-intensive environment without losing the realistic complexity of health. Based on the cross-industry standard process for data mining (CRISP-DM), Niaksu (2015) and Esfandiari et al. (2014) proposed an extension of the CRISP-DM to address specific challenges of big data analytics in healthcare, and described some specialized tasks and activities for each phase, respectively. Considering healthcare as an adaptive system with a combination of three essential components—decision making, decision informatics, and human interface, Tien and Goldschmidt-Clermont (2009) proposed a decision-making framework from data to information, knowledge, and wisdom, and also a decision informatics paradigm with a feedback loop among multiple data sources, abstracted information, and real-time decision.

From the perspective of systems engineering and service engineering, we put forward the paradigm of big data analytics in healthcare, as shown in Fig. 2.2. Firstly, the fusion and analysis of multi-source heterogeneous data can be used as input for data-driven decision modeling on the one hand, and for building a knowledge map on the other hand. Secondly, descriptive modeling and predictive modeling are carried out by using data mining methods and technologies, where the descriptive modeling mainly includes the feature extraction of objects from high-dimensional sparse data and the complex relation representation between individual objects, while the predictive modeling mainly includes statistical inference and prediction model. Then, normative modeling for obtaining the knowledge is carried out by integrating the knowledge map into the results based on descriptive and predictive modeling. Finally, knowledge can provide decision support for the practical problems in the operation and management of the medical service system.

While based on the literature records related to data mining for EHRs, Chen et al. (2017) adopted the Latent Dirichlet Allocation (LDA) and Topics Over Time (TOT) models to extract topics and analyze topic evolution trends and further summarized the general research framework of data mining for the medical domain by combining the topic co-occurrence relations and domain knowledge, including the data, methods, knowledge, and decision levels, as shown in Fig. 2.3. This research framework can provide a high-level insight for scholars in the medical domain field and guide their choices of medical data mining techniques in healthcare knowledge discovery, medical decision support, and public health management.

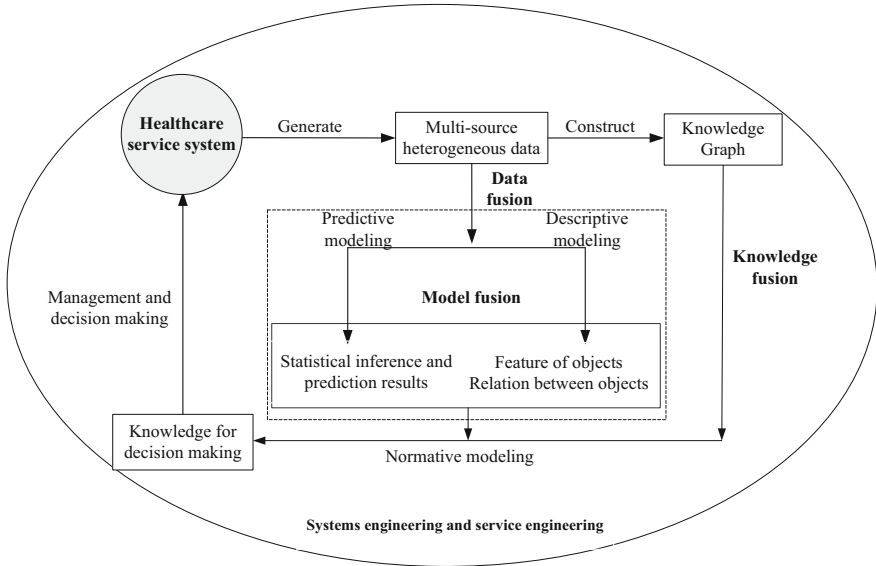


Fig. 2.2 Big data analytics in healthcare from the perspective of systems engineering and service engineering

Further, as the core medical big data, EMRs have become the core foundation of smart hospital construction, and the research on the analysis and utilization of EMRs is of great significance. In order to promote the analysis and utilization of EMRs, an integrated research framework for the generation, analysis, and utilization of electronic medical records was proposed in Fig. 2.4. We found that EMR analysis was helpful to the construction of higher-level hospital intelligent service, and further improve the intelligent service level of the hospital by relying on data mining methods such as classification, recommendation, association rules, text mining, and natural language processing.

In summary, these research frameworks of big data analytics in healthcare are similar, emphasizing data collection and preprocessing methods, big data analytics and modeling techniques, and knowledge for decision support discovery methods to optimize the medical process and further achieve the profound reforms of the medical model.

2.1.3 Analysis of Clinical Diagnosis and Treatment Process

In clinical practice, it is necessary to formulate and implement standardized diagnosis and treatment processes in order to effectively improve the efficiency of medical staff, promote the quality of hospital medical services, and achieve a patient-centered service concept. Shortliffe and Cimino (2006) proposed a clinical diagnosis and

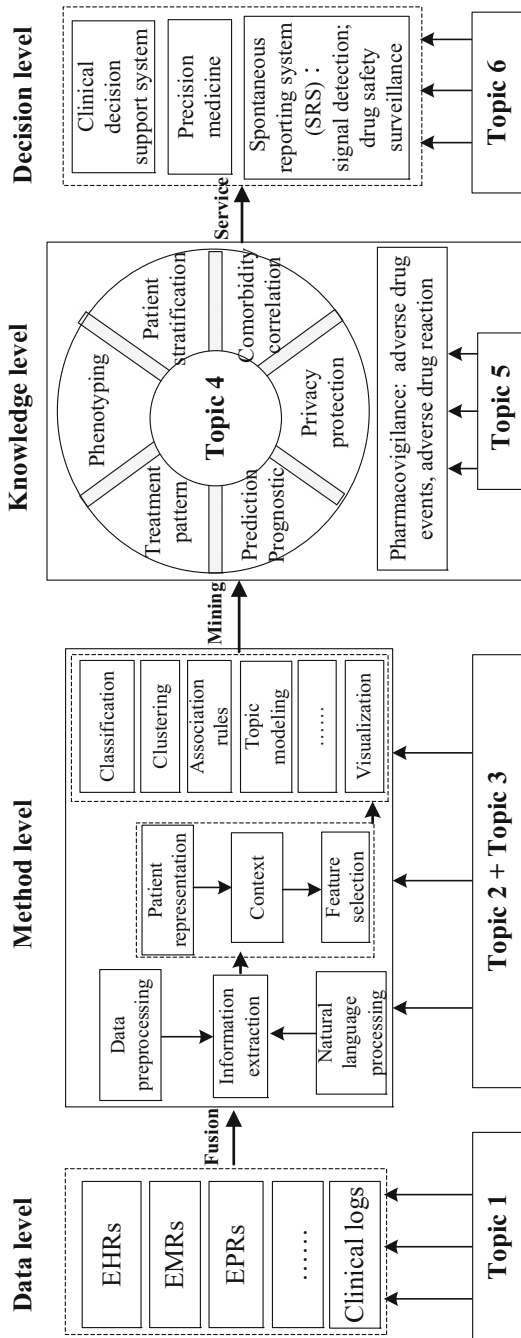


Fig. 2.3 The general research framework of data mining for the medical domain

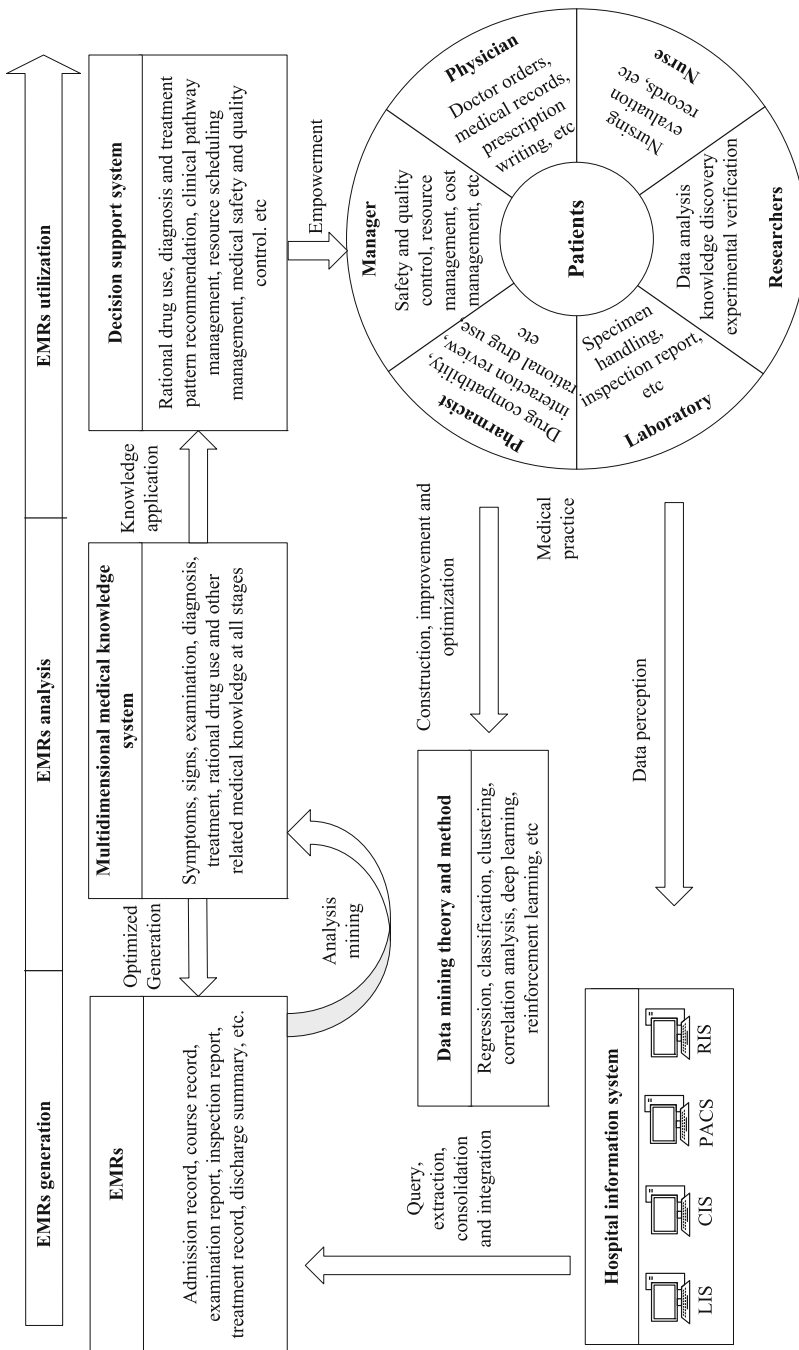


Fig. 2.4 Analysis and utilization framework of EMRs under the background of smart hospital construction

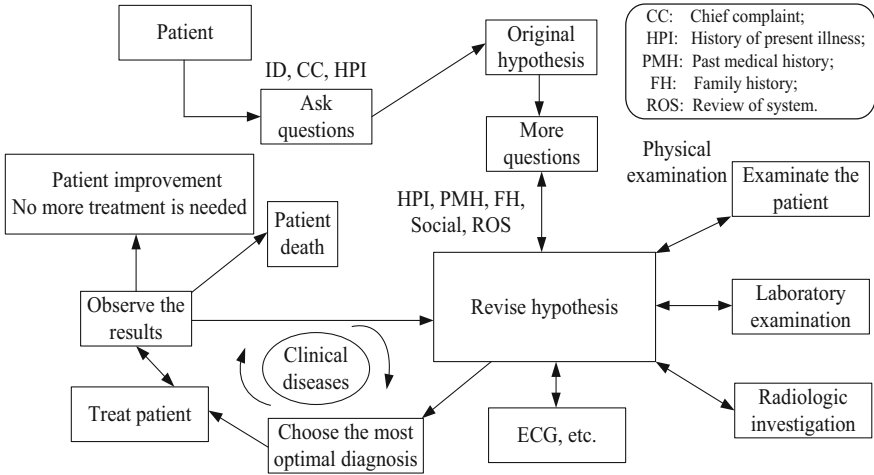


Fig. 2.5 The clinical diagnosis and treatment process based on hypothetical-deductive methods

treatment process based on hypothetical-deductive methods when admitted to a hospital, as shown in Fig. 2.5. First, when a new patient is admitted to the hospital with chief complaints (symptoms or diseases), the doctor forms the initial hypothesis (diagnosis) by asking some questions and further revises the hypothesis based on the patient’s history of present illness, past medical history, family history, social history, and review of the system. Then, when the patient completes the medical examination, the hypothesis lists revised by the doctor may be effectively reduced to determine the appropriate treatment. Finally, the doctor determines the source of the patient’s problems and develops a specific treatment regimen to treat the diseases and observe the outcomes. In addition, when clinical diseases have not been effectively improved, the doctor needs to further revise the hypothesis and treat the patient again.

Whereas clinical data describing patient phenotypes and treatment remains an underutilized source of data, it holds tremendous potential for advancing research and optimizing clinical diagnosis and treatment regimen (Jensen et al., 2012; MIT Critical Data, 2016; Yadav et al., 2018). Thus, we design a clinical diagnosis and treatment process based on data-driven methods to reduce medical costs and improve medical service quality, as shown in Fig. 2.6. Firstly, according to the research framework of big data analytics in healthcare described in Sect. 2.1.2, we can mine diagnosis and treatment patterns from EMRs by data-driven methods, and build two types of rule bases: Admission Information-Diagnosis rule base, and Diagnosis-Treatment rule base. Secondly, when a new patient is admitted to the hospital, the doctor can retrieve the most similar diseases from the Admission Information-Diagnosis rule base on demographic information, symptoms, and laboratory indicators of the patient. Thirdly, the doctor can recommend the most effective treatment pattern for the patient based on the Diagnosis-Treatment rule

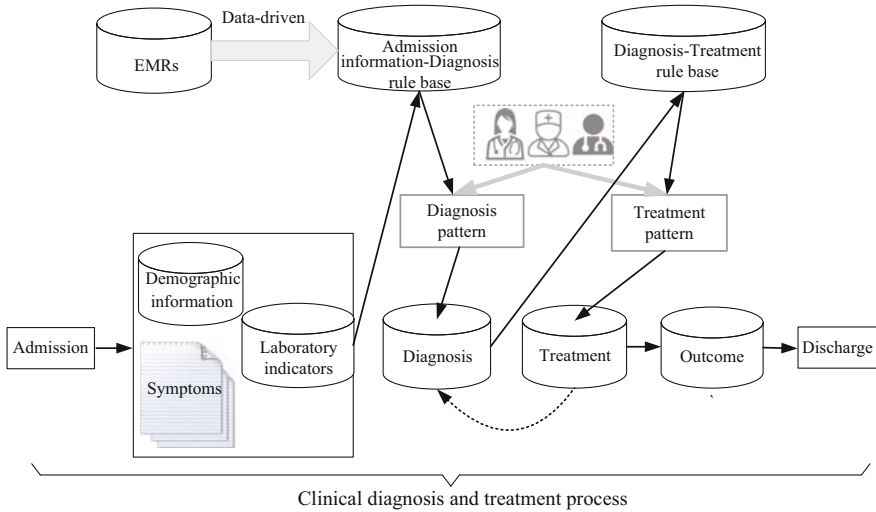


Fig. 2.6 The clinical diagnosis and treatment process based on data-driven methods

base. Finally, if the outcome of the patient is not effectively improved, the doctor needs to further revise the disease type and the corresponding treatment pattern.

Obviously, in the data-driven clinical process, diagnosis-treatment pattern plays an important role to reduce the inflammation that triggers patients’ signs and symptoms and improve long-term prognosis by limiting complications. Meanwhile, the diagnosis-treatment pattern should also meet the requirements of rational drug use. Rational drug use requires that “patients receive medications appropriate to their clinical needs, in doses that meet their own individual requirements, for an adequate period of time, and at the lowest cost to them and their community” (World Health Organization, 2012). The goal of rational drug use is also to achieve “5R”: right patient, right drug, right dose, right route, and right time. Thus, according to the above analysis, we further describe the role of data-driven diagnosis-treatment pattern mining in the healthcare environment in Fig. 2.7. Concretely, on one hand, after collecting the medical evidence (e.g., pyramid of evidence), medical experts adopt the evidence-based medicine (EBM) approach to design clinical guidance, which can be applied to the diagnosis and treatment process proposed in Fig. 2.5. On the other hand, we can mine the diagnosis and treatment rule database from clinical data by data-driven methods, which is suitable for the clinical diagnosis and treatment process described in Fig. 2.6. Then the clinical guidance can guide the feasibility implementation of diagnosis-treatment patterns by providing domain knowledge, and diagnosis-treatment patterns can enrich, supplement, and perfect the clinical guidance, which both can achieve the goal of “5R” in medical (i.e., rational drug use) and “5R” in healthcare (i.e., new value pathways in the healthcare paradigm discussed in Sect. 2.1.1).

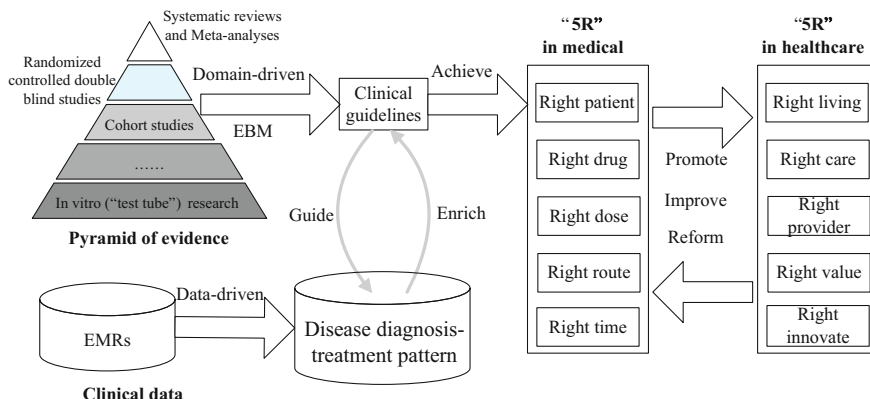


Fig. 2.7 The role of data-driven diagnosis-treatment pattern mining

2.1.4 The Literature Summary of Diagnosis-Treatment Pattern Mining

Data-driven diagnosis-treatment pattern mining is receiving increasing attention in the field of healthcare management. Diagnosis-treatment patterns, as actionable knowledge latent in EMRs representing the best practice for most patients in most time of their clinical processes, can be exploited to help physicians better understand their specialty and learn from previous experiences for clinical guidance improvement (Huang et al., 2015). To the best of our knowledge, unifying diagnosis (UD), clinical pathway (CP), and rational drug use are the main research directions of diagnosis-treatment pattern mining.

2.1.4.1 The Related Work of Unifying Diagnosis (UD)

In medical practice, clinicians are encouraged to seek a UD that could explain all the patient’s signs and symptoms in preference to providing several explanations for the distress being presented (Herman, 1994). A UD is a critical pathway to identifying the correct illness and crafting a treatment plan; thus, clinical experience and knowledge play an important role in the science of diagnostic reasoning. Generally, from a brief medical history of a patient, clinicians can use the intuitive system in their brain and rapidly reason the disease types, whereas, for complex and multi-type abnormal results, clinicians must use the more deliberate and time-consuming method of analytic reasoning to deduce the UD, raising the risk of diagnostic errors.

The coexistence of multiple diseases is pervasive in the clinical environment, particularly for patients in the intensive care unit (ICU) (Sareen et al., 2020). According to the statistical results of the MIMIC-III database, which is a freely accessible critical care database, the average number of diagnosis codes for patients

in the ICU is 11. Additionally, diagnosis codes are highly fine-grained, closely related, and extremely diverse (Johnson et al., 2016). Thus, it is trivial and difficult for clinicians to make a consistent, accurate, concise, and unambiguous diagnostic decision reasonably.

Furthermore, although the inter-relation of diagnosis codes was considered in previous studies, the researchers commonly used the first three digits of ICD-9 codes to assign diagnosis codes for patients (Diao et al., 2021; Wu et al., 2022); hence, the complexity may increase and prediction performance may reduce when considering all digits of the ICD-9 codes. Additionally, in those studies, reasonable complicated and confusing diagnosis codes could not be classified into a UD using a data-driven method. A UD is the basic principle of clinical diagnostic thinking. Its basic idea is that when a patient has many symptoms and if these symptoms can be explained by one disease, it will never explain different symptoms using multiple diseases. A UD reflects the integrity of the patient and the professionalism of clinicians; however, in previous studies, the main focus was on the UD of a category of diseases from the clinical perspective, such as mood/mental disorders (Malhi et al., 2020), intracranial mesenchymal tumor (Sloan et al., 2021), and arrhythmogenic right ventricular cardiomyopathy (Liang et al., 2016).

2.1.4.2 The Related Work of Clinical Pathway (CP)

CPs are regarded as useful tools that ease the tension of the doctor-patient relationship and enable patients to receive correct and timely diagnosis and treatment with controlled medical costs and improved medical quality (Chen, Sun, et al., 2018). In general, process mining is the most popular method to mine CPs from massive EMRs. When process mining technology is applied to clinical environments, treatment behavior can be measured from EMRs that regularly record patient execution information. What is more, due to strict mathematical logic and reasoning ability, process mining can be used as an objective way to analyze clinical pathways (Rebuge & Ferreira, 2012). For instance, Mans et al. (2008) applied process mining technology to discover the treatment workflow of stroke patients. Bouarfa and Dankelman (2012) proposed a process mining algorithm to extract a consensus model from multiple clinical activity logs, which can automatically detect the abnormal behavior of CPs without the prior knowledge of clinical experts. Lakshmanan et al. (2013) designed a process mining approach for mining CPs correlated with patient outcomes that involve a combination of clustering, process mining, and frequent pattern mining. Huang et al. (2013) presented a process mining method for constructing CP summaries from the collected event logs which regularly record various kinds of medical behaviors by hospital information systems. Yang et al. (2017) presented a process analysis and recommendation framework to extract medical prototypes from activity logs.

In addition, sequential pattern mining and probabilistic topic model have also been applied to discover CPs. For instance, Perer et al. (2015) used a frequent sequence mining algorithm to explore care pathways from EMRs with visual

analytics. Huang et al. (2014, 2015) developed a probabilistic topic model to mine treatment patterns hidden in EMRs for clinical pathway analysis and improvement. Hirano and Tsumoto (2014) designed a typicalness index method to mine typical order sequences from EHRs for building clinical pathways. While in clinical practice, considering the complexity of actual treatment activities, variations are widely existent in different stages of CPs. Li et al. (2015) proposed an automatic method to detect CP variation patterns in EMRs and statistically examined their correlation with patient outcomes. Ainsworth and Buchan (2012) developed a collaborative online CP investigation tool that combines the required specialist knowledge and skills from different disciplines, providing a network-based CP variation analysis tool for clinicians and health service managers.

2.1.4.3 The Related Work of Rational Drug Use

Rational drug use is also an important research direction of treatment patterns mining, which requires that the right patient receives the right drug with the right dose and the right route at the right time. EMR data mining technology has been proven that it has good results to analyze drug use efficiency and various drug treatment regimens. For instance, Wright et al. (2015) used sequential pattern mining to automatically infer temporal relationships between medications, visualize these relationships, and generate rules to predict the next medication likely to be prescribed for a patient. Jin et al. (2018) developed a treatment engine to predict next-period prescriptions based on disease conditions, laboratory results, and treatment records of the patient. Chen, Li, et al. (2018) presented a disease diagnosis and treatment recommendation system to recommend medication treatments based on the given inspection reports of patients.

In general, EMRs are heterogeneous and longitudinal in nature, including demographic information, diagnostic information, laboratory indicators, doctor orders, and outcomes. A treatment record is a series of doctor orders, and each doctor's order usually consists of a drug name, delivery route, dosage, start time, and end time. However, in the existing studies, a doctor's 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 (Sun et al., 2016). In this chapter, considering the diversity, temporality, and dynamicity of EMRs, we propose the concept of typical treatment patterns, which can reflect the complexity of EMRs better and enhance the interpretability of mining results.

The rest of the chapter is organized as follows. Section 2.2 highlights the challenges to analyze the large-scale and complex EMRs to mine typical diagnosis-treatment patterns. Section 2.3 describes the UD unifying diagnosis identification and prediction method embedding the disease ontology structure from electronic medical records. Section 2.4 provides four clinical pieces of research on typical treatment patterns in rational drug use and CPs, and discusses the examination of typical treatment pattern mining approaches, limitations, and open issues. Section 2.5 presents the conclusions as well as the challenges.

2.2 Challenges for Typical Diagnosis-Treatment Pattern Mining

EMRs usually contain five kinds of information about patients, such as demographic information, diagnostic information, laboratory indicators, doctor orders, and outcomes. Concretely, demographic information includes the age, gender, address, race and ethnicity, education, and other information of a patient. Diagnostic information includes diagnosis code, disease names, and severity of the diseases. Laboratory indicators record the detailed results of laboratory tests to evaluate the health status of a patient, such as blood routine, urine routine, stool routine, liver function, and kidney function. A doctor order is a medical prescription, including drug name, delivery route, dosage, starting time, and ending time, and a treatment record is a series of doctor orders related to the patient. The outcome is evaluated by doctors when a patient is discharged from the hospital, including treatment efficacy (cured, improved, ineffective, and dead) and treatment efficiency (payment and length of stay) (Chen, Sun, et al., 2018; Dang & Ho, 2017; Sun et al., 2016).

After summarizing our previous works (Chen, Guo, et al., 2018; Chen, Sun, et al., 2018; Sun et al., 2016), we propose a general framework of data-driven typical treatment pattern mining, as illustrated in Fig. 2.8. Our framework has two stages: typical treatment pattern mining and typical treatment pattern evaluation and recommendation. The former includes (1) similarity measure among diagnosis and treatment records; (2) clustering diagnosis and treatment records based on similarity matrix; and (3) typical diagnosis and treatment pattern extraction from each cluster. The latter includes (1) patient cohort division by classification methods; (2) evaluation of diagnosis and treatment records in each patient cohort; and (3) recommendation of the most effective diagnosis and treatment pattern for each patient cohort. In this process, three key technical challenges for the general framework emerge, including how to measure similarity among diagnosis and treatment records, how

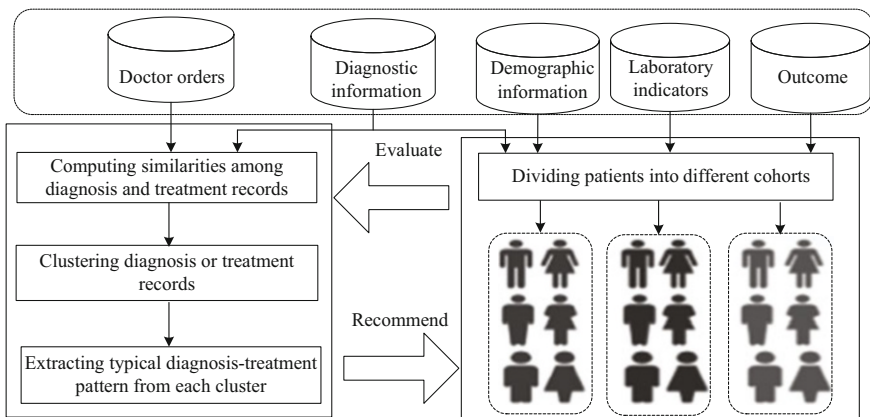


Fig. 2.8 The general framework of data-driven typical treatment pattern mining

to extract typical diagnosis and treatment patterns from EMRs, and how to evaluate and recommend diagnosis and typical treatment patterns.

2.2.1 Measuring Similarity Among Diagnosis and Treatment Records

2.2.1.1 Similarity Measure of Patients' Diagnostic Records

Diagnostic information is one of the most important clinical data. Diagnostic information refers to a record of disease diagnosis made by clinicians based on the health condition of a patient admitted to the hospital. It is stored in the patient's EMR data in the form of a diagnosis code (e.g., ICD-9 and ICD-10). How to calculate the similarity between disease diagnosis codes is a problem to be solved. Diagnosis code is a semantic concept, not a specific numerical value. ICD code of disease diagnosis concept is classified data with a hierarchical structure, which contains medical knowledge. The distance between the two concepts in medical semantics can be judged according to the position of the disease diagnosis concept in the ICD coding tree.

In the real EMR dataset, patient diagnostic information is typically a set of diagnosis codes, as shown in Fig. 2.9. Thus, patient similarity can be transformed into the similarity of the diagnosis code set. Generally, for binary code-level similarity, we can use classical methods, such as Dice, Jaccard, cosine, and overlap, to calculate set-level similarity. However, these methods cannot fully embed semantic similarity. Thus, it is critical to measure the similarity of patients' diagnostic records by fusing the information content measure of diagnosis codes, diagnosis code similarity measure, and diagnosis code set similarity measure.

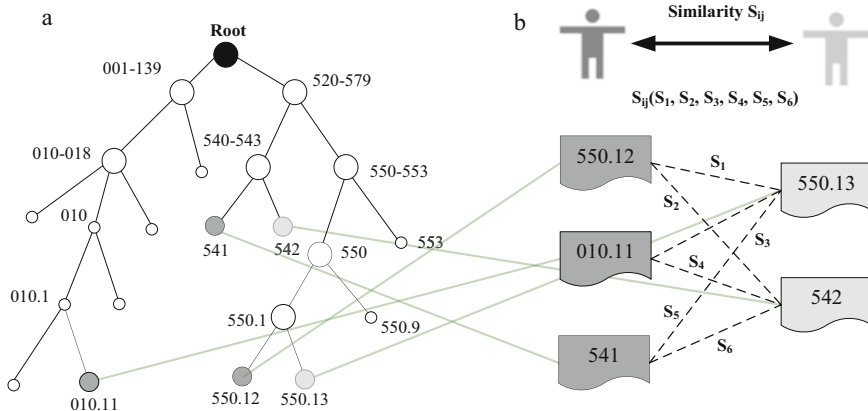


Fig. 2.9 Example of two patients' diagnostic records in the ICD-9 ontology structure

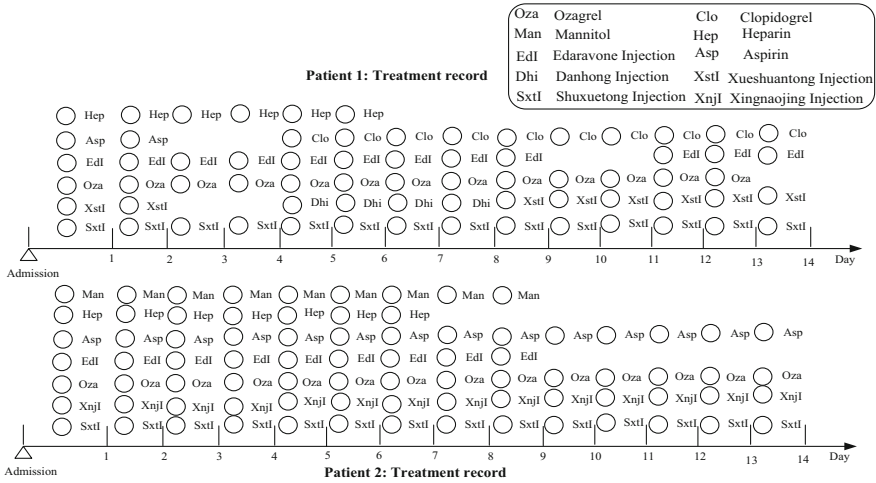


Fig. 2.10 Treatment records of two cerebral infarction patients

2.2.1.2 Similarity Measure of Patients' Treatment Records

The similarity between pairwise treatment records measures how similar a pair of treatment records are according to their doctor order information under a specific clinical context. As discussed in Sect. 2.1.4, a treatment record is a series of doctor orders with timestamps, which can be seen as a temporal event, as shown in Fig. 2.10. In general, the treatment information not only includes nominal terms like drug name, and delivery route, but also figures like dosage, frequency per day, and repeated times, so the recorded information in a treatment record is heterogeneous. The timestamp is also more complex than previously studied as it records both start and end times. In this case, how to measure similarity between pairwise treatment records has become a challenging problem (Sun et al., 2016, 2021).

After analyzing the characteristics of treatment records in Fig. 2.10, there exist three categories of differences illustrated in Fig. 2.11: including (1) doctor order content difference: each doctor order is a set of seven tuples, including drug name, drug efficacy, delivery route, daily dosage, frequency, start and end time; (2) doctor order duration difference: the usage and duration time of the same doctor order are various in different treatment records; and (3) doctor order sequence difference: certain temporal relations exist between doctor orders. Thus, it is necessary to take these differences into full consideration when designing similarity measure methods of pairwise treatment records (Chen, Guo, et al., 2018; Chen, Sun, et al., 2018; Htun & Sornlertlamvanich, 2017; Sun et al., 2016, 2021).

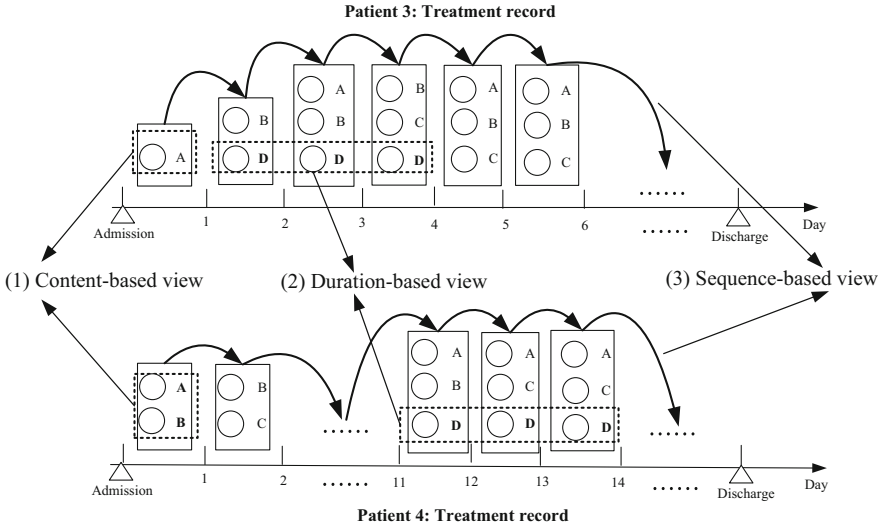


Fig. 2.11 Three-view analysis for treatment records of two patients

2.2.2 Extracting Typical Diagnosis-Treatment Patterns from EMRs

After obtaining the similarity matrix for all diagnosis and treatment records, we first divide all diagnosis and treatment records into several groups by clustering algorithms and then extract a typical diagnosis and treatment pattern from each cluster. Clustering is a technique of partitioning a set of objects into multiple groups (i.e., clusters) so that objects in the same cluster are more similar to each other than to those in other clusters (Cho & Kim, 2017; Han et al., 2011; Wang et al., 2018; Xu & Tang, 2018). For the research on cluster analysis in data-driven management and decisions, Sun, Chen, et al. (2017) discussed the three most popular clustering categories, such as centroid-based clustering, connectivity-based clustering, and density-based clustering, analyzed and addressed five challenges for cluster analysis in new business environments, including clustering dynamic data, clustering a large-scale data set, finding the representatives, handling arbitrary-shaped clusters, and validation measures and consensus clustering, and further provided three practical cases relating to management and decisions, for instance, clustering enhanced information extraction, data-driven operations research, and clustering assisted knowledge discovery.

2.2.2.1 Typical Diagnosis Pattern Extraction from Clustering Results

Some previous studies have proved that defining the core zone of a cluster is an effective approach to extracting stable clustering results (Chen et al., 2020).

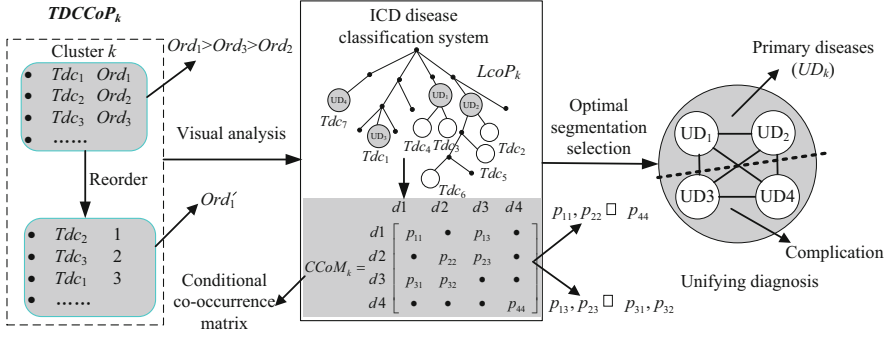


Fig. 2.12 The extraction process of the UD from diagnostic records

Additionally, considering the complex semantic relations among different diagnosis codes, the feature of a cluster cannot be fully described when the diagnostic information (cluster center or exemplar) of only one patient is used. Thus, the core zone of each cluster can be defined to select a group of patients (i.e., core patients) using the k -nearest neighbor method, and further, identify typical diagnosis code co-occurrence patterns (TDCCoP) from each cluster by defining a threshold and a sorting function.

To extract typical diagnosis patterns (i.e., UD) from patients' diagnostic records, categorizing the TDCCoP of each cluster reasonably according to the disease taxonomy is a critical step. Chen et al. (2022) proposed a UD identification method, as shown in Fig. 2.12. Specifically, for the TDCCoP_k of cluster k , all typical diagnosis codes were visualized in the reconstructed ICD ontology structure and marked in their orders. Then the least common ancestor (LCA) method was used to categorize these codes and define their LCA and the corresponding orders. Furthermore, the conditional co-occurrence matrix was calculated using patient diagnostic information to select the optimal segmentation between primary diseases and complications, where the primary diseases were regarded as UD.

2.2.2.2 Typical Treatment Pattern Extraction from Clustering Results

Clustering large-scale treatment records is also a big challenge to extracting typical treatment patterns. Sun et al. (2016, 2021) proposed a MapReduce enhanced density peaks-based clustering (MRDPC) to address this challenge, as shown in Fig. 2.13. MRDPC is a two-stage procedure. First, the total N patients are first randomly divided into m parts, DPC is implemented on each part with an $N_0 \times N_0$ similarity matrix to obtain k potential exemplars (i.e., representative objects); then a partial similarity matrix with $m \times k \times N$ is obtained by computing similarities between the selected potential exemplars and all objects, and partial DPC (PDPC) is used to determine K final exemplars according to the partial similarity matrix (Sun et al., 2016, 2021).

Then after clustering all treatment records, a typical treatment pattern can be identified from each cluster. In most of the previous applications of exemplar-based

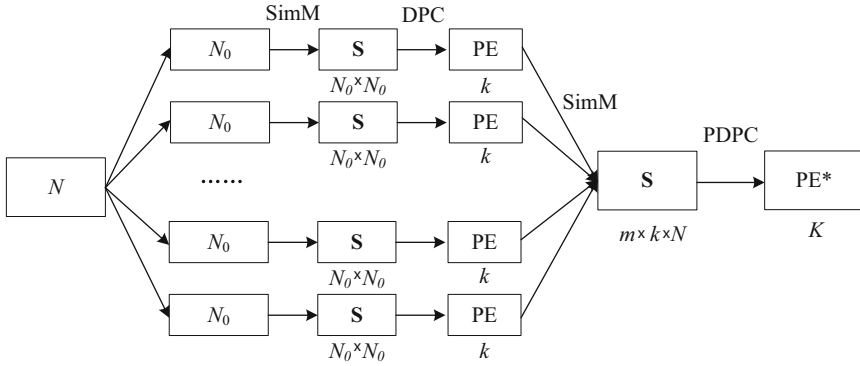


Fig. 2.13 Clustering treatment based on MapReduce Enhanced DPC method

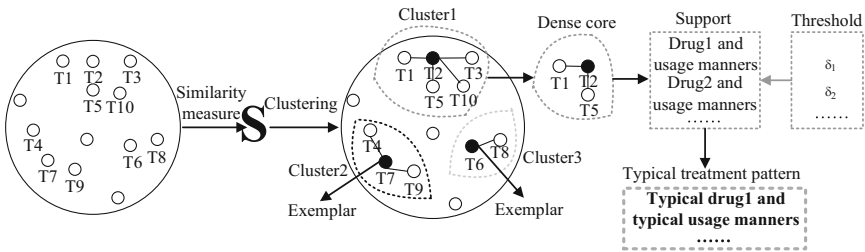


Fig. 2.14 The extraction process of typical treatment patterns from EMRs

clustering (e.g., affinity propagation (AP) and density peaks-based clustering (DPC)), an exemplar can be directly used to describe the corresponding cluster. However, a treatment record can vary in many different directions as a complex temporal and heterogeneous data set, and the exemplar of each cluster cannot well describe the cluster it belongs to. In this case, Sun et al. (2016, 2021) defined the core area of a treatment cluster and extract a semantic description of each treatment cluster by its dense core. Further, the typical treatment pattern can be extracted from the dense core based on the trade-off between the support of drug or usage manners of drug and a threshold defined beforehand, as shown in Fig. 2.14.

2.2.3 Predicting Typical Diagnosis Patterns

After extracting the typical diagnosis pattern (i.e., UD), Chen et al. (2022) further proposed the prediction task based on the health condition of a patient admitted to the hospital, exploring the important features to assign the most possible UD to new patients. Figure 2.15 shows the proposed UD prediction method. First, three

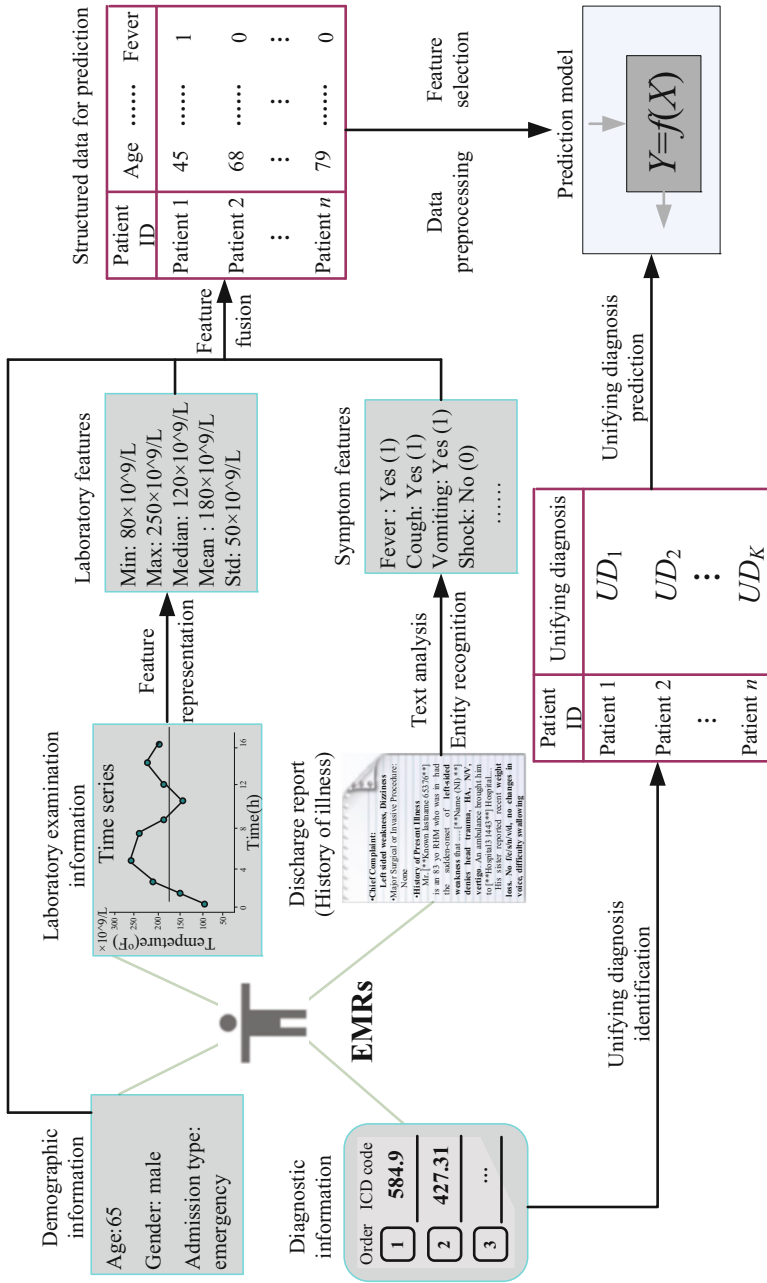


Fig. 2.15 Prediction of typical diagnosis pattern

categories of features using time series feature representation and text analysis methods were fused into structured data for further prediction. Then after data preprocessing and feature selection, all patients were labeled with a UD. Finally, some classical prediction models were adopted to perform the UD prediction task.

2.2.4 Evaluating and Recommending Typical Treatment Patterns

Before recommending typical treatment patterns (TTP) to patients, how to evaluate their effectiveness is also one of the most challenging problems, since the treatment outcome is affected by a lot of factors, and for different patient cohorts, the most effective typical treatment patterns may be different. Sun et al. (2016, 2021) presented a general framework with three stages to address this challenge shown in Fig. 2.16. First, according to demographic information, laboratory indicators, diagnostic information, and outcomes of all patients, we divide patients into different groups by a decision tree model. The patients in the same leaf node are defined as a patient cohort. Then, for a specified patient cohort, we observe how many typical treatment patterns have been used on the patients in this cohort, and further figure out which treatment pattern can result in the highest effective rate. Finally, we can recommend the best typical treatment pattern for each patient cohort.

In addition, Chen, Guo, et al. (2018) and Chen, Sun, et al. (2018) proposed a brief evaluation and recommendation framework. First, we use treatment outcomes to

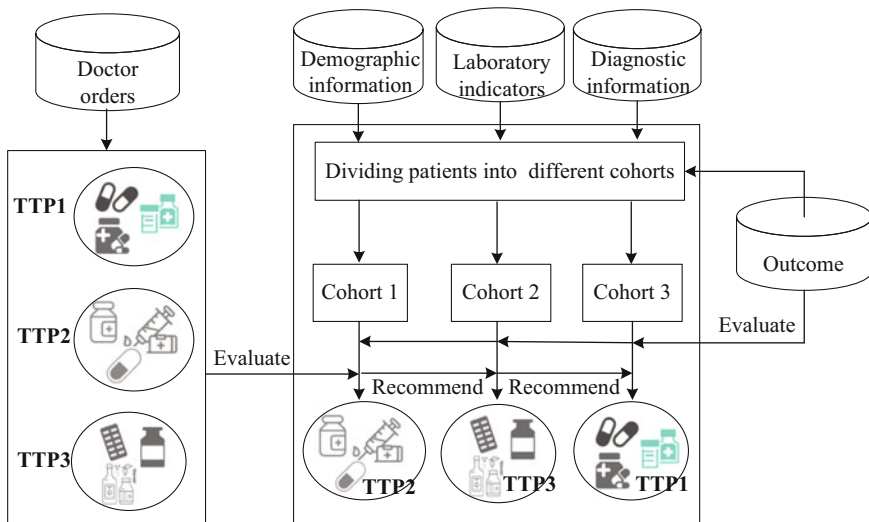


Fig. 2.16 Evaluation and recommendation of typical treatment patterns

evaluate the effectiveness of the extracted typical treatment patterns, such as treatment efficacy and treatment efficiency. Then, we also analyze demographic information, laboratory indicators, diagnostic information of each pattern, and identify some representative characteristics. Finally, for a specific patient cohort with these representative characteristics, we can recommend the most effective typical treatment pattern for new patients.

2.3 Typical Diagnosis Pattern Mining for Clinical Research

This section provides a clinical case of data-driven typical diagnosis pattern mining and predicting (i.e., UDIPM) from EMRs in our previous studies (Chen et al., 2022). In clinical practice, the reasonable classification of a large number of distinct diagnosis codes can clarify patient diagnostic information and help clinicians to improve their ability to assign and target treatment for primary diseases. Thus, the accurate identification and prediction of the UD from a large number of distinct diagnosis codes and multi-source heterogeneous patient admission information in EMRs can provide a data-driven approach to assist in better coding integration of diagnosis. Chen et al. (2022) proposed a research framework for data-driven UDIPM from EMRs, as shown in Fig. 2.17.

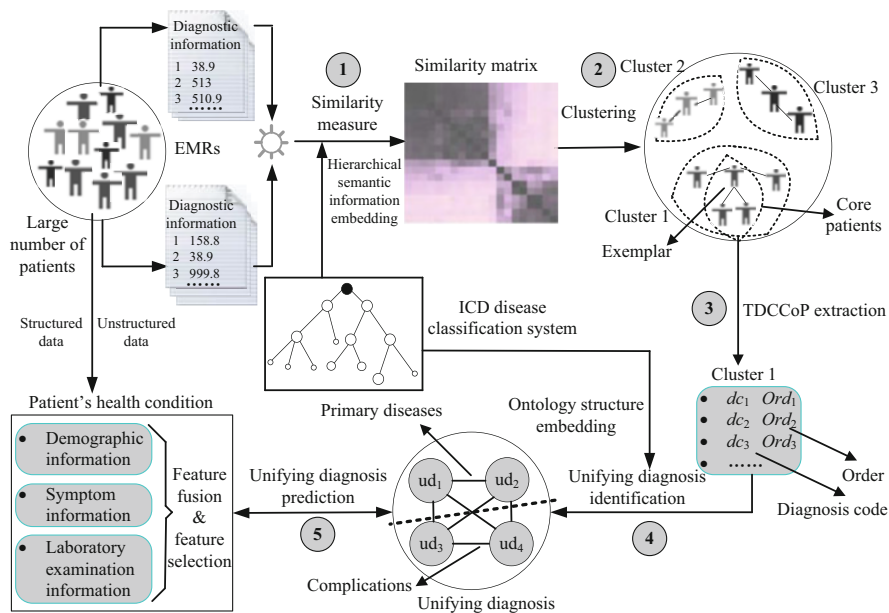


Fig. 2.17 Research framework for applying the proposed UDIPM to EMRs

This study adopted diagnostic information to identify the UD and used demographic information, symptom information, and laboratory examination information to predict the UD. First, a set of similarity measure methods was applied to a large number of patients by embedding the semantic relation of the ICD classification system (Task 1). Second, a clustering algorithm was adopted to divide patients into different groups, and further obtain the exemplar and core patients of each cluster (Task 2). Third, the typical diagnosis code co-occurrence patterns (TDCCoPs) were identified from each cluster by defining a threshold and a sorting function (Task 3). Fourth, the visual analysis and conditional co-occurrence matrix (CCoM) were combined to extract the UD by selecting the optimal segmentation (Task 4). Finally, after obtaining the health condition of the patient admitted to the hospital, a UD prediction using multi-class classification methods was achieved (Task 5).

After applying the AP clustering algorithm, we first divided the 4418 sepsis patients into two clusters, where clusters 1 and 2 contained 1391 and 3027 patients with the support of 31.48% and 68.52%, respectively. After obtaining TDCCoPs, we visualized all the TDCs in the ICD-9 ontology structure and obtained the LCA co-occurrence pattern (LCoP), as is shown in Fig. 2.18. Then we calculated the CCoM₂ of the LCoP₂ based on the diagnostic information of 800 core patients in cluster 2, as described in Table 2.1. Thus, diseases of the respiratory system (460–519, order: 3) and diseases of the circulatory system (390–459, order: 5)

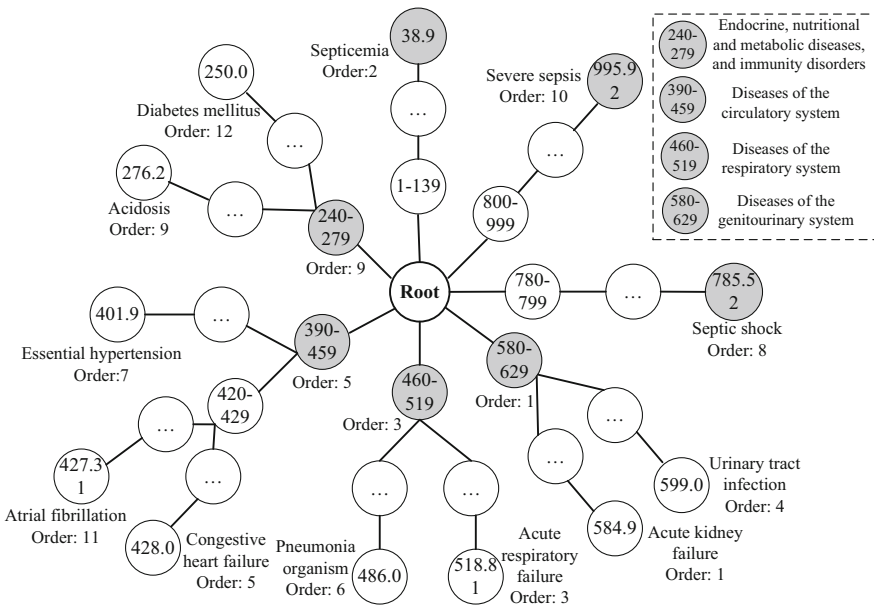


Fig. 2.18 LCoP₂ identified using the visualization of TDCCoP₃ in the ontology structure

Table 2.1 CCoM obtained based on the result of LCoP in Fig. 2.18

$p_2(d_j/d_i)$	580-629	38.9	460-519	390-459	785.52	240-279	995.92
785.52 (8)	0.99	0.60	0.65	0.93	0.73	0.60	0.97
240-279 (9)	0.74	0.63	0.67	0.91	0.71	0.61	0.94
995.92 (10)	0.74	0.61	0.66	0.92	0.75	0.61	0.94

Note Values in brackets are the orders of the seven diseases, and bold values on the master diagonal denote the occurrence probabilities of the seven diseases

were likely to be the optimal segmentation between primary diseases and complications, and the first three diseases were considered to be the UD (UD₂) of cluster 2.

Further, we applied feature fusion and feature selection using the IG method and performed five classifications to predict a UD based on patient admission information and identify important features for the constructed prediction models. Figure 2.19 shows the classification performance of the proposed UDIPM, including the area under the ROC curve (AUC), accuracy (Acc), precision (Pre), recall (Rec), and F1-score (F1).

The experimental results indicated that the proposed UDIPM achieved better prediction performance, where the AUC values were all above 0.8, except for the decision tree method. Similarly, the best Acc, Pre, Rec, and F1 among all classifications were XGBoost, at approximately 80%, followed by random forest, SVM, and logistic regression, whereas the decision tree was last, at approximately 66%. Consider the random forest as an example. We obtained the feature importance results to better understand the prediction model. First, we found that demographic information (i.e., age) and laboratory examination information were more important than symptom information. Then some disease severity indicators were very important, such as SAPS and SAPS-II. Finally, the variance distribution (i.e., Var) of the laboratory examination indicators was more important than the mean, median, minimum, and maximum values. To summarize, the proposed UDIPM not only identified a UD from patient diagnostic information but also predicted a UD based on the health condition of a patient admitted to the hospital.

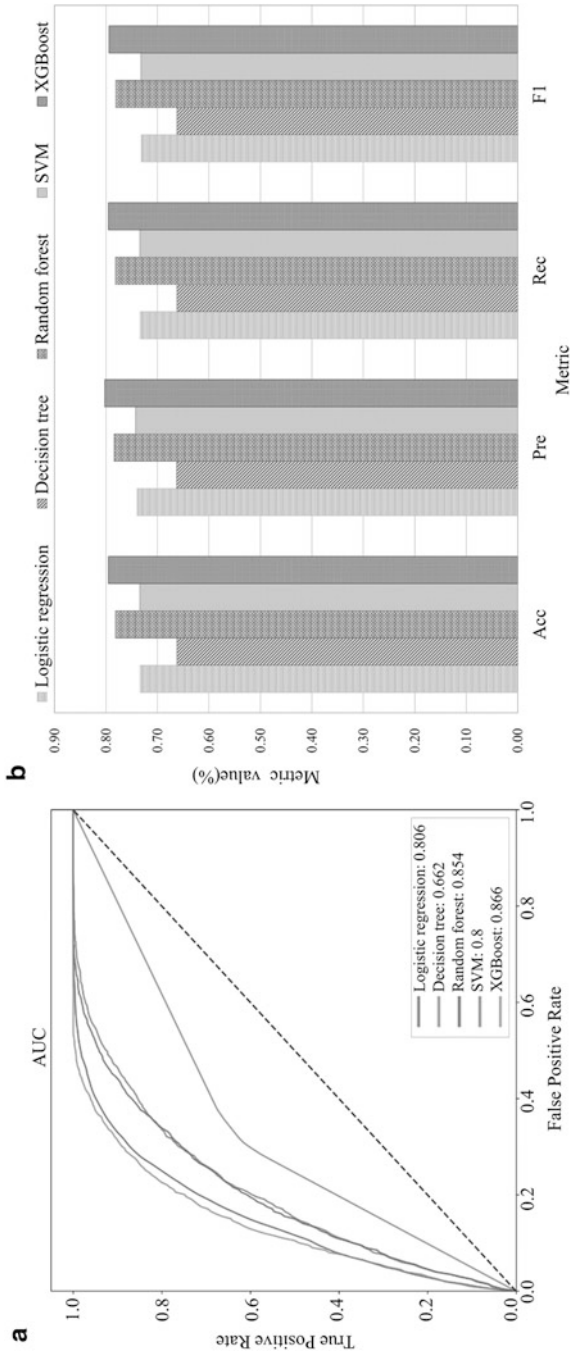


Fig. 2.19 Classification performance of the proposed UDIPM. (a) AUC. (b) Acc, Pre, Rec, and F1

2.4 Typical Treatment Pattern Mining for Clinical Research

This section provides three clinical cases of data-driven typical treatment pattern mining from different views in our previous studies (Chen et al., 2020; Chen, Guo, et al., 2018; Chen, Sun, et al., 2018; Sun et al., 2016, 2021). The first case proposes a data-driven typical treatment regimen mining approach from a doctor order content view, which is published in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Sun et al., 2016) and Transactions on Knowledge and Data Engineering (Sun et al., 2021). The second case designs a data-driven typical drug use pattern mining approach from a doctor order duration view, which is published in proceedings of the 19th International Symposium on Knowledge and Systems Sciences (Chen, Guo, et al., 2018) and *Journal of Systems Science and Systems Engineering* (Chen et al., 2019). The third case discusses the context of clinical pathways and presents a data-driven typical treatment process mining approach from a doctor order sequence view in the *Journal of Biomedical Informatics* and our work can provide managerial guidance for clinical pathway redesign and optimization (Chen, Sun, et al., 2018). The fourth case proposes a fusion framework to extract typical treatment patterns based on the multi-view similarity network fusion method in Artificial Intelligence in Medicine (Chen et al., 2020). Furthermore, all proposed methods have been validated on real-world EMRs of the cerebral infarction dataset and MIMIC-III dataset (Johnson et al., 2016). In addition, a typical treatment regimen, typical drug use patterns, and typical treatment process can be regarded as one of the typical treatment patterns according to different research questions. Thus, both cases are in the context of rational drug use, and the methods we proposed can contribute to achieving the “5R” goal, namely right patient, right drug, right dose, right route, and right time.

2.4.1 Typical Treatment Regimen Mining from Doctor Order Content View

A typical treatment regimen usually refers to a series of doctor orders with a high frequency of occurrences (i.e., typical doctor orders) in a group of patient treatment records, and each typical doctor order also includes the drug name, delivery route, daily dosage, frequency, start and end time. Sun et al. (2016, 2021) presented a research framework of data-driven typical treatment regimen mining from doctor order content view shown in Fig. 2.20. This process has been discussed in Sect. 2.2, except for the similarity measure methods. In this work, we developed a novel method that can compute the similarity between two doctor orders by an orderly combination of a drug name, delivery route, and dosage-per-day, and further proposed a complex set similarity measure method for computing the similarity between two treatment records.

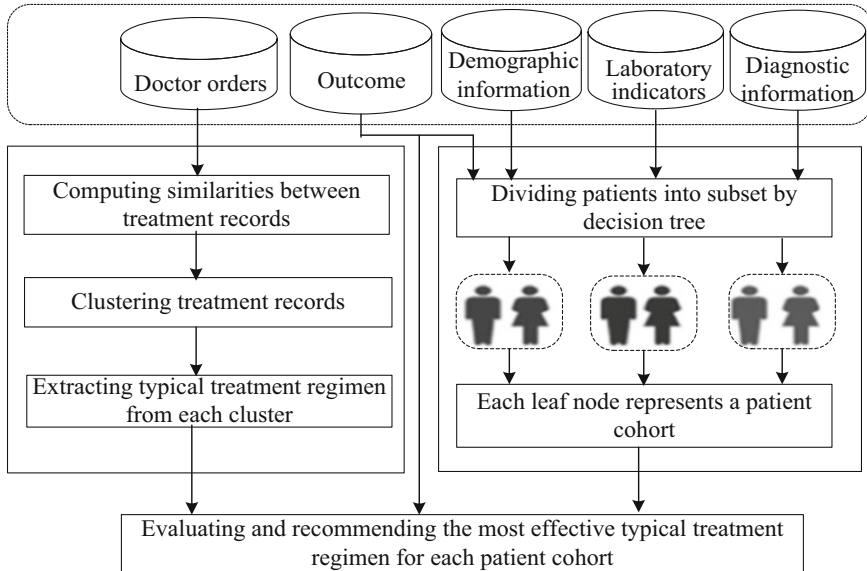


Fig. 2.20 The research framework of data-driven typical treatment regimen mining

After clustering treatment records, we extracted typical treatment regimens from each cluster. For instance, Fig. 2.21 shows the extraction results of typical treatment regimen 2, where each bar denotes a typical drug. Concretely, the support of typical treatment regimen 2 is 15.5%, and the most typical drugs are Shuxuetong, Ozagrel, Cinepazide, and Aspirin. The usages of four medicines in different periods are also different. Further, taking the third period (4–7 days) for example, each pie denotes the different usage manners of the typical drug with its support, such as “IV/160/4” of Ozagrel refers to that the delivery route is an intravenous injection (IV), the daily dosage is 160 units, four days are used during the third period, and the support is 52%.

Next, after extracting typical treatment regimens and dividing the patients into homogeneous cohorts by the decision tree method, we can evaluate and recommend the most effective typical treatment regimen for each patient cohort. For instance, Fig. 2.22 shows the evaluation and recommendation for two patient cohorts. Specifically, for Case 1 (leaf node 2 with 4035 patients), most of the patients are cured and improved. Typical treatment regimen 4 (Patient-T4) is the best regimen with the highest cure rate and improved rate, but only 0.37% of patients in this cohort used this regimen. Typical treatment regimen 3 (Patient-T3) with higher support of 25.97% is regarded as the most effective treatment regimen because it can obtain a higher cure rate and lower ineffective and dead rate than a typical treatment regimen 1 (Patient-T1) and 2 (Patient-T2). Similarly, we can recommend typical treatment regimen 2 to the patient cohort with leaf node 17.

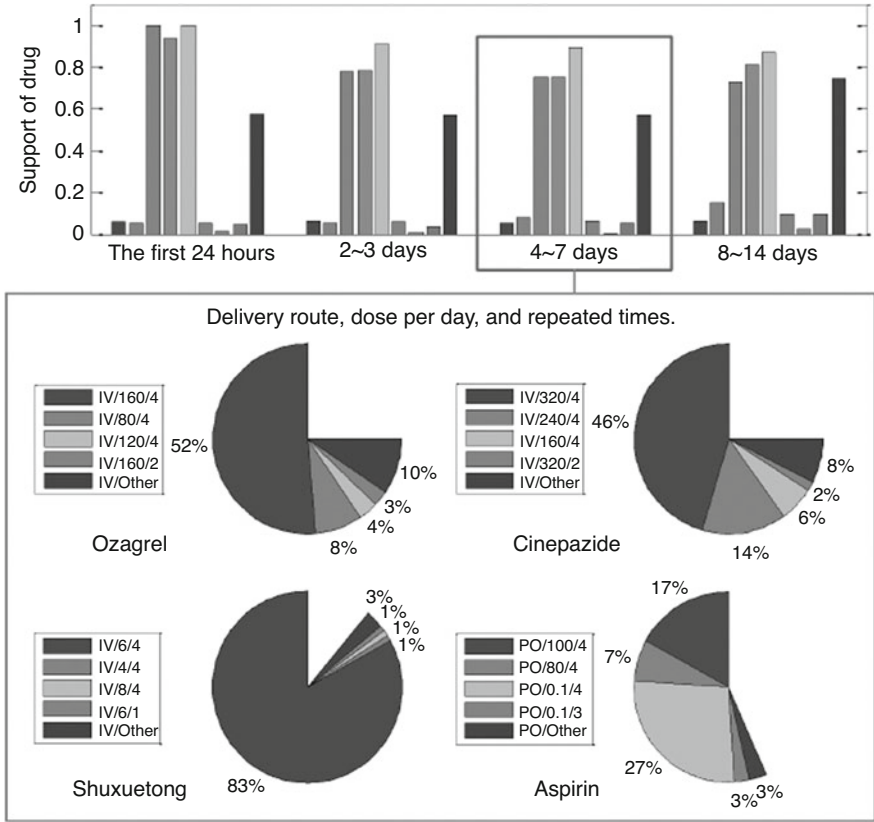


Fig. 2.21 The extraction results of the typical treatment regimen

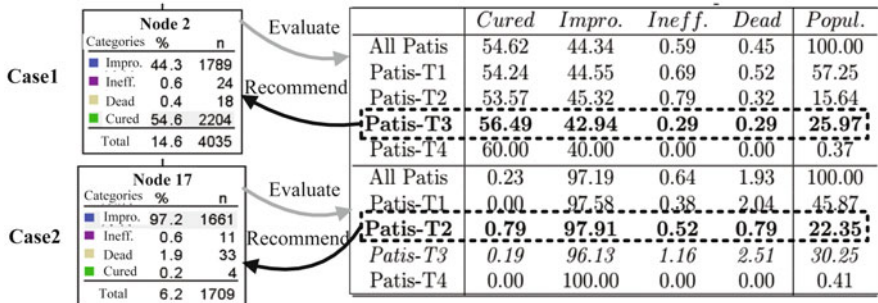


Fig. 2.22 Evaluate and recommend typical treatment regimens for two patient cohorts

2.4.2 Typical Drug Use Pattern Mining from Doctor Order Duration View

Rational drug use also requires that patients receive medications for an adequate period of time. The adequate duration time of medications not only improves the therapeutic effect of medicines but also reduces the side effects and adverse reactions of medicines. Chen, Guo, et al. (2018) and Chen et al. (2019) proposed a research framework of data-driven typical drug use pattern mining from the doctor order duration view shown in Fig. 2.23. The main process has been also discussed in Sect. 2.2, except for the representation of the drug use distribution feature vector (DUDFV) from doctor orders and the similarity measure methods. In this work, in order to analyze the duration time characteristic of medications, we first defined the drug use distribution feature with a quintuple for each drug, including the mean, the variance, the lasting days, and the first and last day of drug use. Then we represented the DUDFV of each patient by the ordered combination of DUDFs for all drugs and further used the Euclidean distance to measure the similarity between pairwise DUDFVs.

After clustering DUDFVs, we extracted three typical drug use patterns (i.e., pattern 1, pattern 2, and pattern 3). For instance, Fig. 2.24 shows the extraction results of pattern 2, where each black bar in Fig. 2.24 (1) and Fig. 2.24 (2) denotes a drug and drug use day, respectively; each white bar in Fig. 2.24 (1) and Fig. 2.24 (2) denotes a typical drug and effective drug use day when exceeding a threshold defined beforehand, respectively; and the curve in Fig. 2.24 (2) is the effective drug use days and DUDF of Heparin. Concretely, the support of pattern 2 is about 55% with 19 typical drugs, and the support of each typical drug is different. For Heparin

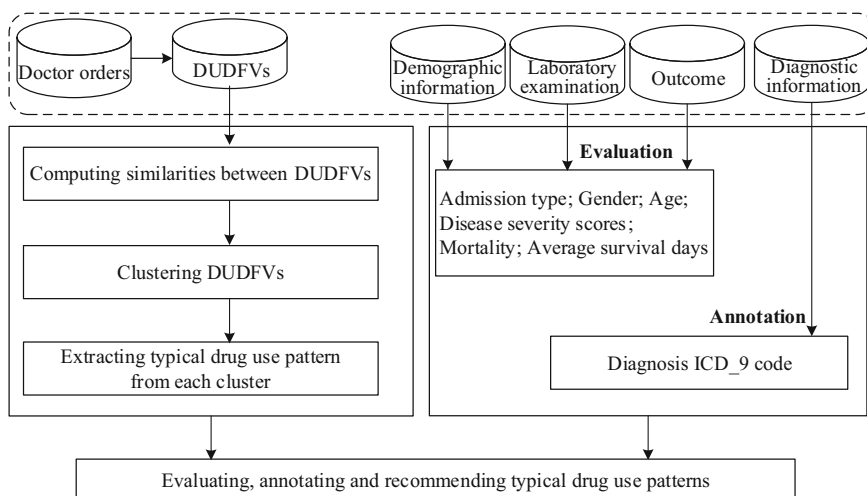


Fig. 2.23 The research framework of data-driven typical drug use pattern mining

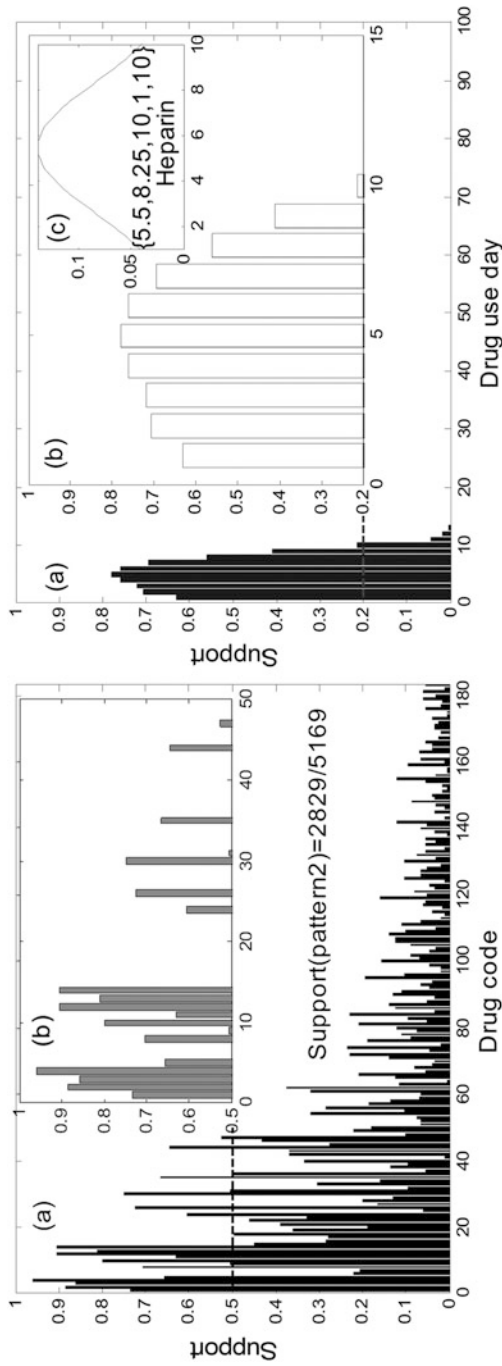


Fig. 2.24 The extraction results of pattern 2. (1) Typical drugs of pattern 2. (2) Effective drug use days and DUDF of Heparin

selected from pattern 2, the DUDF is {5.5, 8.25, 10, 1, 10} indicating that the mean, the variance, the lasting days, and the first and last day of Heparin use are 5.5, 8.25, 10, 1, and 10. Similarly, we can also obtain the DUDFs of all typical drugs and provide clinical guidance for the duration time of drug use.

Then, we further evaluated the extracted typical drug use patterns based on demographic information, laboratory examination and outcome, annotated diagnosis codes for each typical drug use pattern according to diagnostic information, and proposed a recommendation work for the patients with the same patient condition and disease types shown in Fig. 2.25. In Fig. 2.26, we deem patterns 2 and 3 to be the

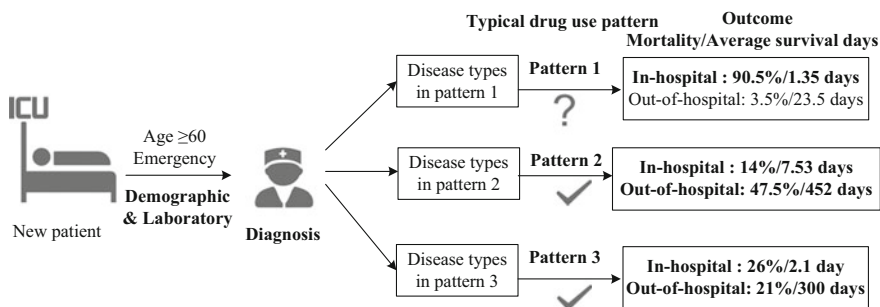


Fig. 2.25 Recommendation of typical drug use patterns

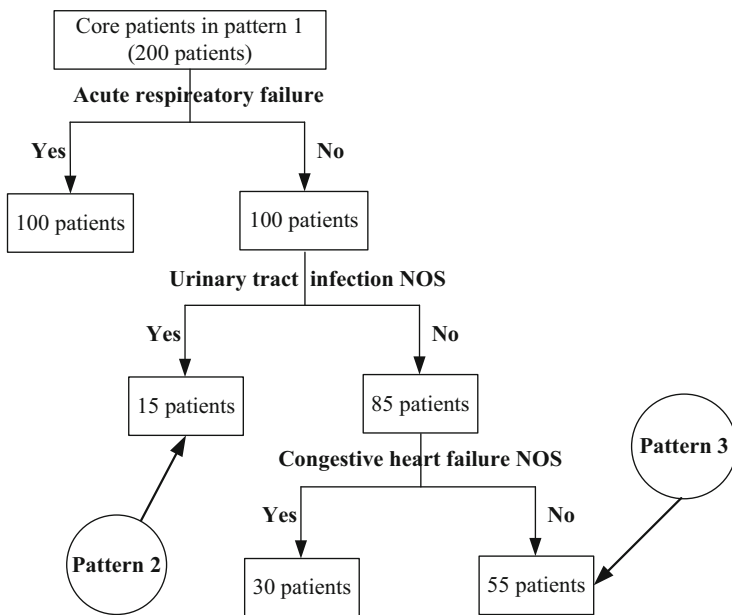


Fig. 2.26 Recommendation for the patients in pattern 1

effective typical drug use patterns because of lower in-hospital and out-of-hospital mortality and a longer average survival time than that of pattern 1. Thus, for patients in pattern 1, we further analyze their disease types and recommend pattern 2 and pattern 3 to the seventy patients in Fig. 2.26, which can effectively improve their treatment outcomes.

2.4.3 Typical Treatment Process Mining from Doctor Order Sequence View

A clinical pathway (CP) defines a standardized care process for a well-defined patient group, aimed at improving patient outcomes and promoting patient safety (Huang et al., 2015). Figure 2.27 shows the process of CP design and implementation. However, in clinical practice, creating such a pathway from scratch is demanding for medical staff as it involves multidisciplinary medical team collaboration, plan-do-check-act-related techniques, and optimal EBM (Chen, Sun, et al., 2018). In addition, due to the difference in disease severity, complication, multi-pathogenesis, and reaction to therapy, the variation of CPs often occurs when implementing them for patients.

In order to build CPs from EMRs, Chen, Sun, et al. (2018) proposed a research framework of data-driven typical treatment process mining from the doctor order sequence view shown in Fig. 2.28. This process has been also discussed in Sect. 2.2, except for the representation of doctor order set sequence (DOSS) from doctor orders and the similarity measure methods. In this work, considering the treatment courses in clinical practice, we divided treatment into different periods and defined DOSS, then generated a set transition matrix sequence from DOSS based on Markov chain theory, and further adopted Manhattan distance to compute the similarity between two treatment records.

After clustering all DOSSs, we can extract the typical treatment process from each cluster. For instance, Fig. 2.29 shows the extraction result of typical treatment processes from dataset 3 (i.e., patients in critical condition), where each circle denotes a typical drug, and each line represents the transition probability of two doctor orders in the adjacent period. Specifically, we identified four categories of typical treatment processes with seven drugs for cerebral infarction patients in the critical condition. For typical treatment process 1, we can extract a high-frequency typical treatment process (HF-TTP: support $\in [0.7,1)$, black thin line), namely, {Admission, {Lum, GBEP, Asp}, {Lum, GBEP, Asp}, {Lum, GBEP, Asp}, {Lum, GBEP, Asp}, Discharge}, which can be contribution to build CPs. Similarly, we can also extract some HF-TTPs from the typical treatment processes 2, 3, and 4.

Next, we further evaluated the treatment efficacy and efficiency, analyzed demographic and diagnosis information of typical treatment process, and discussed a recommendation work for the patients with the same patient condition and disease severity. Figure 2.30 shows the recommendation of typical treatment processes,

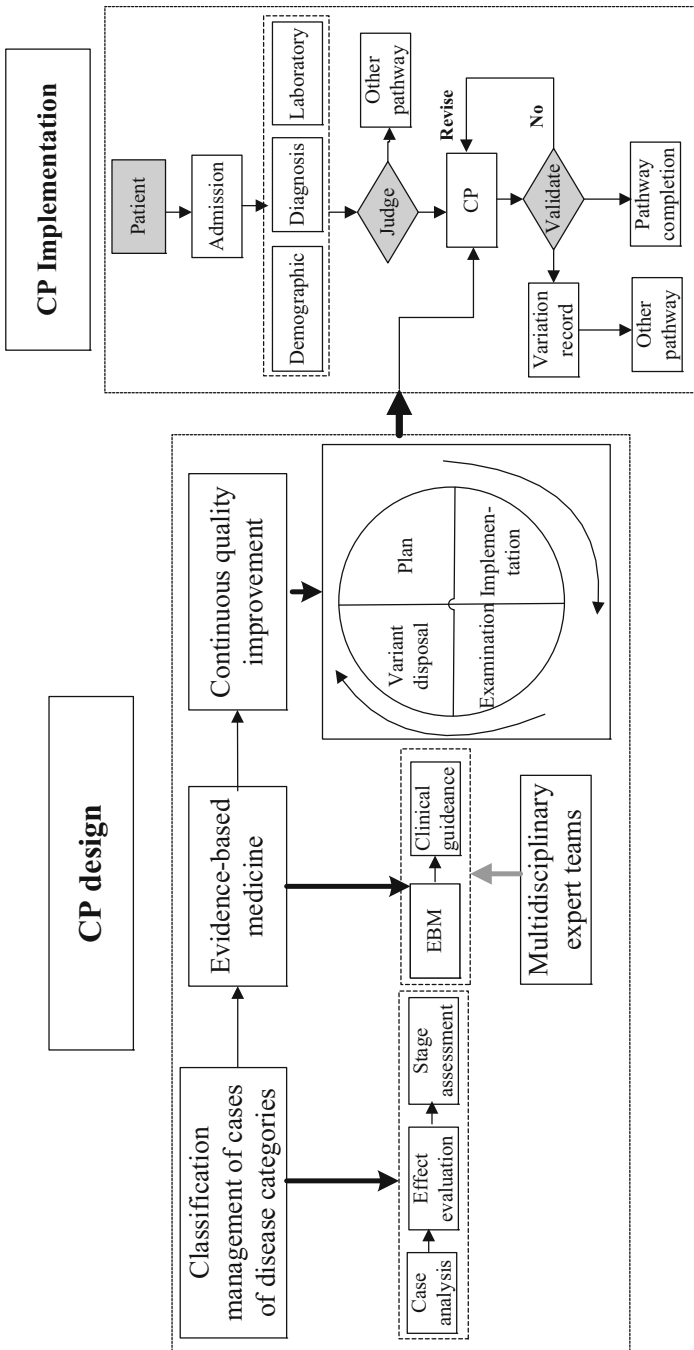


Fig. 2.27 The process of CP design and implementation

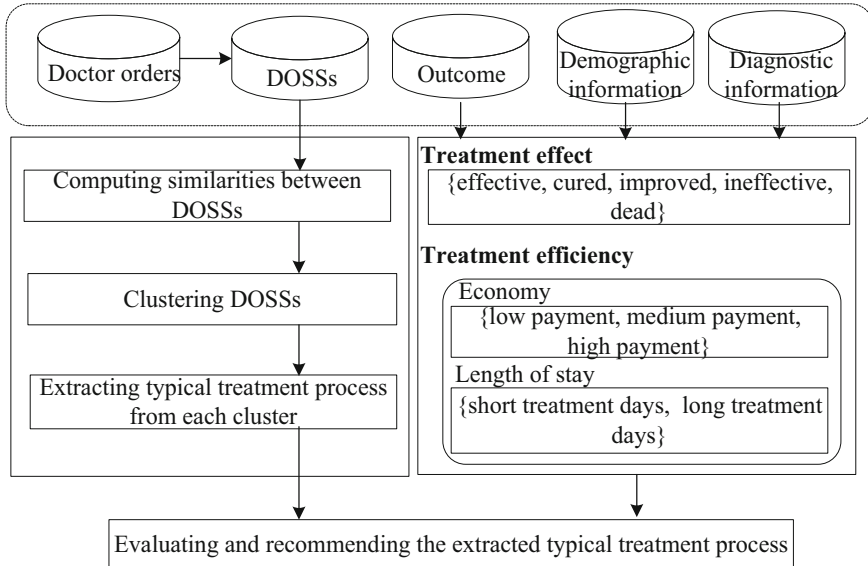


Fig. 2.28 The research framework of data-driven typical treatment process mining

where F-TTP, S-TTP, and C-TTP denote the extracted typical treatment processes from dataset 1 (patients in fair condition), dataset 2 (patients in serious condition), and dataset 3 (patients in critical condition), respectively. For instance, when a new patient is admitted to hospital A, and diagnosed with cerebral infarction in fair condition after demographic information and laboratory examination by clinical doctors, the F-TTP1 can be recommended for this patient, and treatment outcome is predictable, such as the cured rate is as high as 74%, the probability of payment [¥4000, ¥24,000] is 86%, and the probability of length of stay less than two weeks is 78%. Similarly, we can recommend the best treatment for different patients according to Fig. 2.30.

2.4.4 Typical Treatment Pattern Mining from Multi-View Similarity Network Fusion

In clinical practice, rational drug use means that patients receive medications appropriate to their clinical needs, in doses that meet their own individual requirements, for an adequate period of time, and at the lowest cost to them and their community (World Health Organization, 2012). Thus, the goal of rational drug use is to achieve the “5Rs”: “right drug,” “right dose,” “right route,” and “right time” for “right patient.” However, due to diseases with multiple similar treatment stages, various symptoms, and multiple pathogeneses and clinical experience and

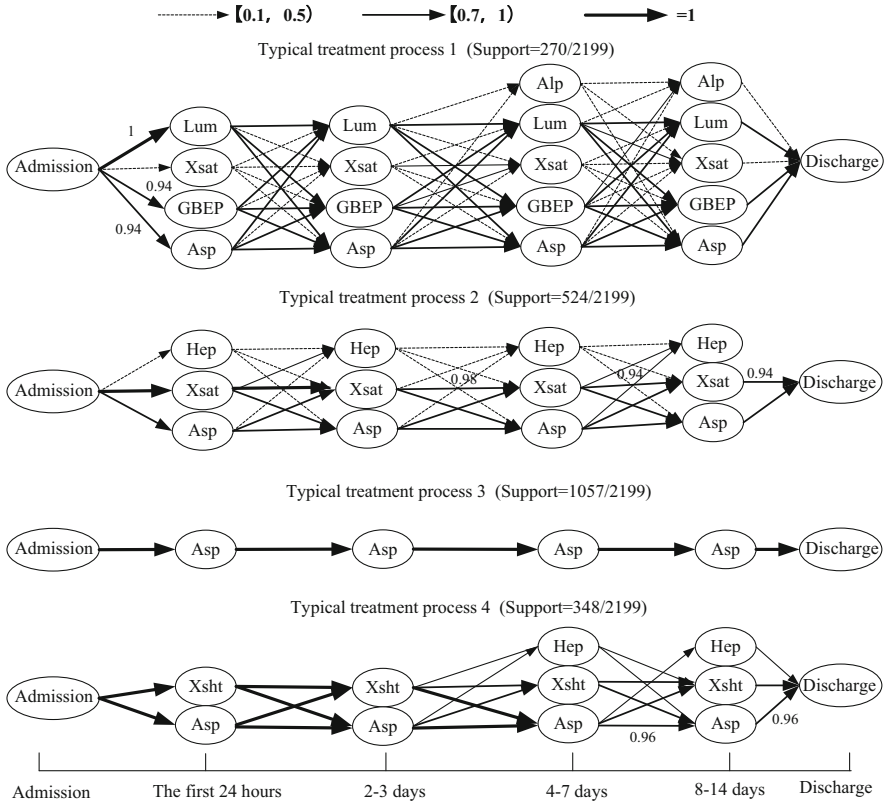


Fig. 2.29 The extraction result of the typical treatment process from dataset 3

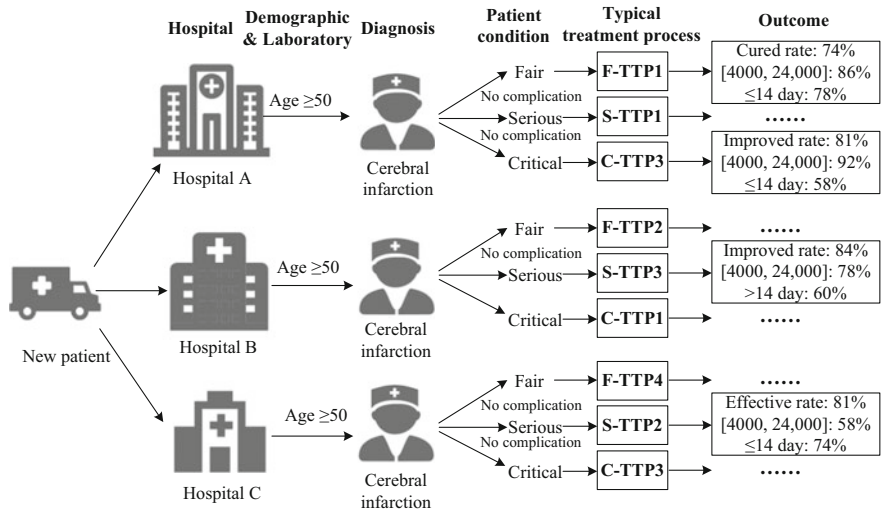


Fig. 2.30 Recommendation of typical treatment processes

knowledge with the characteristics of inadequate communication, experience exchange, and cooperation between young and senior doctors, it is difficult to achieve the “5Rs” goal in an accurate and efficient manner (Chen, Guo, et al., 2018; Chen, Sun, et al., 2018).

Therefore, Chen et al. (2020) analyzed the characteristics of doctor orders, formulate new patient representations and compute the corresponding patient similarity from three views (i.e., an improved doctor order content view patient similarity measure, a doctor order sequence view patient similarity measure, and a novel doctor order duration view patient similarity measure), and applied a multi-view Similarity Network Fusion (SNF) method to fuse three kinds of patient similarity for typical treatment pattern extraction. Figure 2.31 illustrates the fusion framework of a typical treatment pattern extraction in this chapter. The fusion framework mainly consists of four steps: (1) terms and definitions, (2) patient similarity measure methods, (3) the multi-view SNF method, and (4) the typical treatment pattern extraction method.

Real-world EMR data of cerebral infarction patients used in our experiment were collected from three Traditional Chinese Medicine (TCM) hospitals, which are

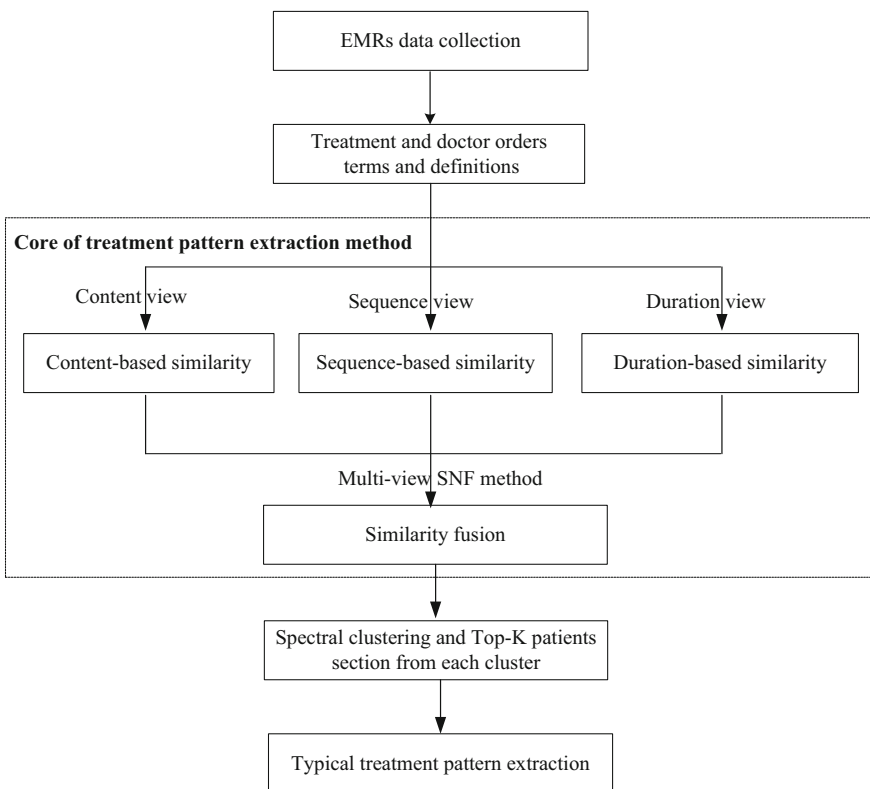


Fig. 2.31 The research framework of data-driven typical treatment pattern mining

located in three cities in China. After evaluating the performance of the multi-view SNF method, we first discussed the selection of the number of clusters and then use the proposed fusion framework to extract typical treatment patterns, including the distribution of typical drugs in different periods, delivery routes, doses per day, and repeated times in different periods from the content view, the transition of typical drugs in different periods from the sequence view, and the duration distribution of typical drugs from the duration view.

From the content view, Fig. 2.32 describes the distribution of typical drugs in four periods for TTP3. First, with the exception of Mannitol (78) in the fourth period (i.e., 8–14 days), most drugs are widely used in four periods, where Aspirin (133) has the largest support of 92%, followed by Xuesaitong (114), Ozagrel (36), and Heparin

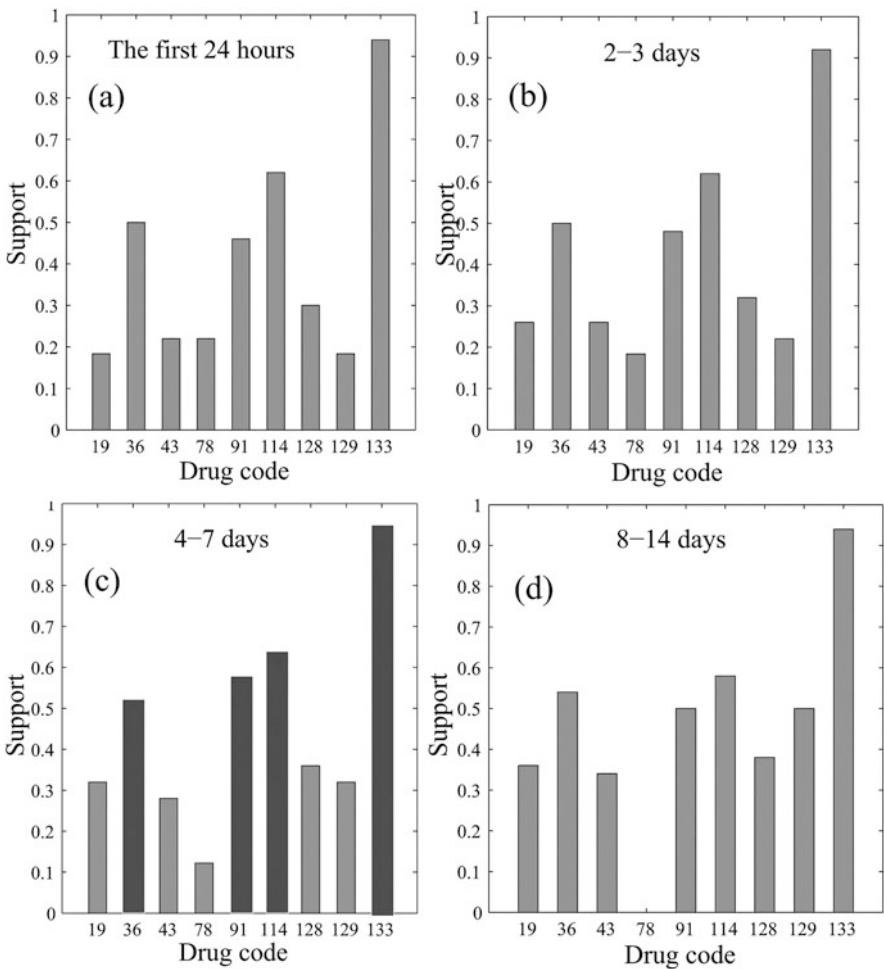


Fig. 2.32 The distribution of typical drugs in different periods for TTP3

(91). Second, Mannitol (78), as a hypertonic antihypertensive drug, is commonly used in clinical rescue, especially in the treatment of brain diseases with the characteristics of rapid and accurate antihypertensive effects to reduce intracranial pressure. Thus, along with the extension of the treatment period, the support gradually decreases from 22% to 0%. Finally, the support of Alprostadi (19), Yindanxinnatong capsules (128), and Ginkgo biloba extract powder (129) gradually increases from the first 24 h to 8–14 days post-infarction, while other drugs remain unchanged. Additionally, we selected Ozagrel (36), Heparin (91), Xuesaitong (114), and Aspirin (133) as four representative drugs in the third period (i.e., 4–7 days post-infarction) to further analyze how these typical drugs are used.

Figure 2.33 shows the usage manners of the four representative drugs used in the third period for TTP3, including their drug efficacy, delivery route, dose per day, and repeated times. Overall, each drug has multiple usage manners with different supports, and fewer drugs have the same usage manners. Specifically, for Ozagrel (36), the most widely used manner is “IV/80/4” with a support of 28%, followed by “IV/60/4” with a support of 10%, where “IV/80/4” indicates the delivery route is Intravenous Injection, the daily dose is 80 units, and the duration is 4 days in the third period. Heparin (91) is an important anticoagulant drug to treat cerebral infarction and has three distinct usage manners, where “ST/other” ranks first with

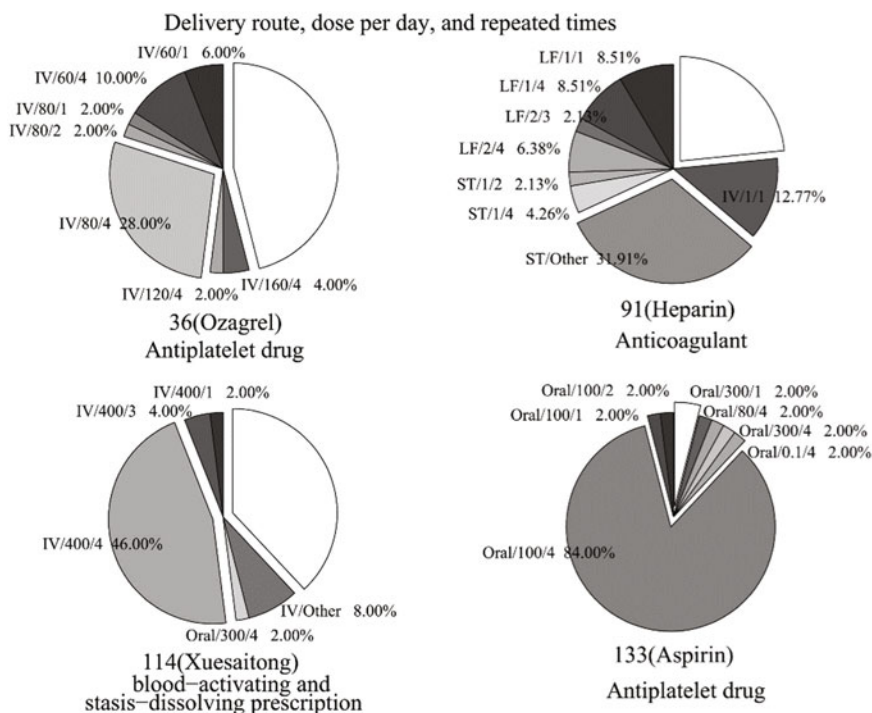


Fig. 2.33 The usage manners of four representative drugs used in the third period for TTP3

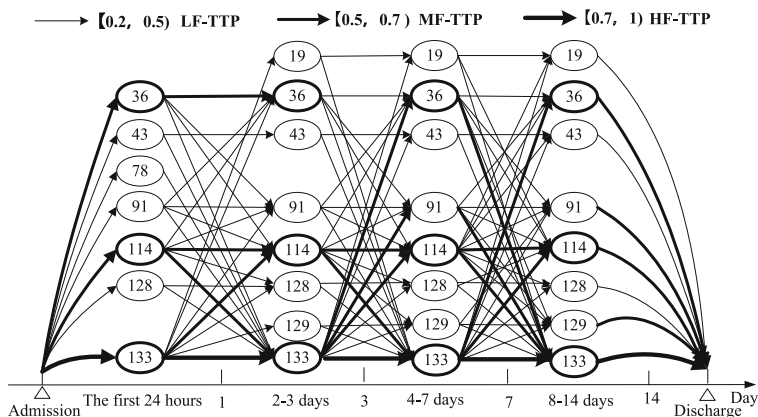


Fig. 2.34 The transition between typical drugs in four periods for TTP3

the support of 31.91%, followed by “IV/1/1,” “Lock Flush (LF)/1/4,” and “LF/1/1.” “ST/other” denotes the delivery route is Subcutaneous Injection, but the daily dose data are missing. For Xuesaitong (114) and Aspirin (133), the most popular usage manners are “IV/400/4” with a support of 46%, and “Oral/100/4” with the support of 84%, respectively.

From the sequence view, Fig. 2.34 shows the extracted transition patterns for TTP3, including an HF-TTP ({Admission, Aspirin (133), Aspirin (133), Aspirin (133), Aspirin (133), Discharge}), an MF-TTP ({Admission, {Ozagrel (36), Xuesaitong (114)}, {Ozagrel (36), Xuesaitong (114)}, {Ozagrel (36), Xuesaitong (114)}, Discharge}), and some LF-TTPs. In general, the HF-TTP and MF-TTP can be used as important guidance to build different levels of CPs.

From the duration view, Fig. 2.35 shows the duration distribution of the four representative drugs used in TTP3, the mean usage day and dispersion degree of these drugs are similar to distribution intervals [6.9, 7.4] and [15.3, 16.1], while the durations are greatly different, the shortest is Heparin (91) with 7.9 days, while the longest is Aspirin (133) with 12.35 days. Additionally, the start and end times of the four drugs are different. For example, Ozagrel (36) is started on approximately the second day and is ended on the twelfth day, Heparin (91) is started on approximately the third day and is ended the eleventh day, and Xuesaitong (114) is started on the seventh day and is ended on the fifteenth day, and Aspirin (133) is started on the first day and is ended on the fourteenth day.

2.4.5 The Examination of Typical Treatment Pattern Mining Approaches, Limitations, and Open Issues

As discussed in Sections 2.2 and 2.4, four kinds of typical treatment pattern mining approaches mainly include similarity measure method, clustering algorithm, and

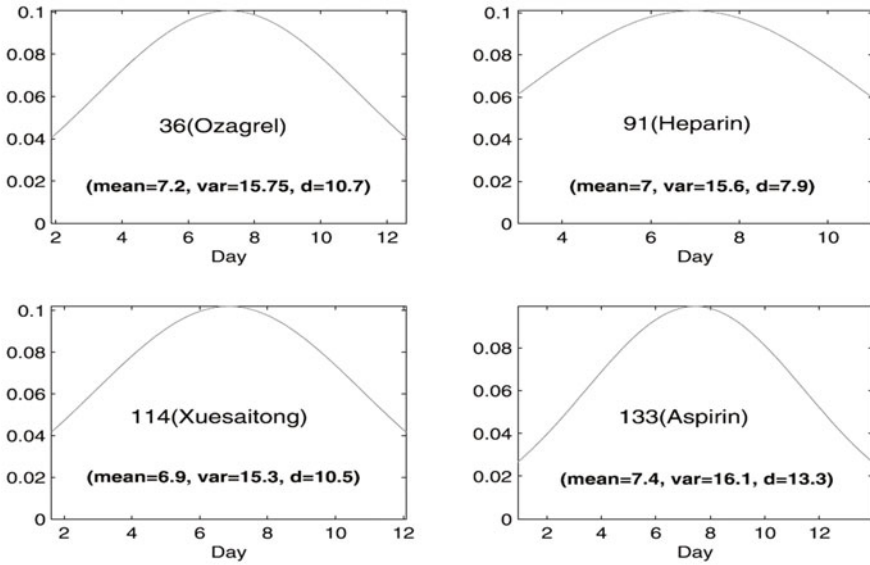


Fig. 2.35 The duration distribution of the four representative drugs used in TTP3

typical treatment pattern extraction method; thus it is an unsupervised learning approach and can be used the clustering results to examine the performance. Concretely, for the experimental setting, the labeled clinical dataset is essential, which requires patients with the same label have similar characteristics, such as patient condition, complication, treatment days, total payment, hospital code, and treatment efficacy. For evaluation criteria, clustering accuracy and normalized mutual information can be used to examine the clustering results (Chen, Sun, et al., 2018; Sun, Guo, et al., 2017). For the baselines of the similarity measure method, we can select the LDA with cosine distance, vector space model weighted by TF-IDF with cosine distance, and edit distance (Chen et al., 2016; Chen, Sun, et al., 2018; Guo et al., 2018). For the clustering algorithm, we can select AP clustering, K-center, and DPC. Chen, Sun, et al. (2018) have demonstrated that the typical treatment pattern mining approach we proposed achieved the highest clustering performance among different clustering algorithms.

In addition, there are still some limitations in our studies. Firstly, fixed intervals for treatment periods may not be the most optimal split due to the complex and varying length of treatment records (e.g., the four periods in the first two weeks), thus a new solution for future studies could involve splitting varying-length treatment records based on significant changes in prescription indications (Hoang & Ho, 2019). Secondly, in the experimental setting, some parameters need to be manually defined beforehand, such as the weights of different treatment periods, the threshold of typical drugs, the definition of the core area for a treatment cluster, and so on. Finally, the labeled clinical dataset is essential to examine the performance of typical treatment pattern mining approaches, while in our experiment, only a small

amount of the clinical dataset is manually annotated, which may be a lack of sufficient evidence to demonstrate the advantages of our approaches.

Furthermore, abnormal activities occur frequently in clinical practice; thus abnormal diagnosis and treatment patterns mining from mass EMRs is also a crucial issue for improving clinical diagnosis and treatment level, optimizing the existing clinical guidelines, and identifying healthcare insurance fraud incidents.

2.5 Conclusions

The advance of big data analytics in healthcare is accelerating the transformation of the medical paradigm. This chapter is an extension of our previous work (Guo & Chen, 2019), firstly discussed the research background of big data analytics in healthcare, summarized the research frameworks of big data analytics in healthcare, and analyzed two types of medical processes to highlight the important role of data-driven diagnosis-treatment pattern mining in clinical guidance. Then for three challenges, we investigated how to measure similarity between diagnosis and treatment records, how to extract typical diagnosis-treatment patterns from EMRs, and how to predict, evaluate, and recommend typical diagnosis-treatment patterns. Further, five clinical pieces of research have been provided to demonstrate the important role that data-driven diagnosis-typical treatment pattern mining can contribute to achieving the “5R” goal in UD identification and predication, rational drug use, and CP redesign and optimization. Finally, we also discussed the examination of typical diagnosis-treatment pattern mining approaches, limitations, and open issues.

Although big data analytics and artificial intelligence technology are promoting the automatization, informatization, and intellectualization of healthcare service, several challenges have been widely recognized as major barriers to the successful implementation of big data in healthcare. First and foremost, the security and privacy concerns surrounding big data in healthcare have become increasingly urgent in recent times, primarily due to the sensitive nature of diagnosis and treatment records. To address these concerns, one approach is to enact and enforce the laws and regulations of data sharing and exchange by the government such as the health insurance portability and accountability act (HIPAA) and the health information technology for economic and clinical health (HITECH) Act in the United States; another is to accelerate technological developments in data privacy protection by the technology of data masking, encryption, and de-identification.

In addition, with the increasing popularity of intelligent diagnosis and treatment machines in clinical practice, how to determine their ethics and the legal liability among clinicians, intelligent machines, and producers for medical accidents are becoming the subject of attention. Nowadays there exists a consensus that clinicians are the leader of human-machine relationships, and intelligent diagnosis and treatment machines cannot replace them completely, but assist them to make better clinical decisions. Thus, clinician intelligent diagnosis and treatment machine integration is an effective pathway to enhance the efficiency of healthcare service. In the

future, considering high integration and interdisciplinary cooperation of technologies, ethics, laws, and regulations, it is possible to embed ethics and laws into intelligent diagnosis and treatment machines and determine their status as liability subjects.

Acknowledgments This work was supported by the National Natural Science Foundation of China (71771034; 72101236), the Fundamental Funds for the Central Universities of China (DUT21YG108), the Henan Province Medical Science and Technology Research Plan (LHGJ20200279), and the Henan Province Youth Talent Promotion Project (2021HYTP052).

References

- Ainsworth, J., & Buchan, I. (2012). COCPIT: A tool for integrated care pathway variance analysis. *Studies in Health Technology and Informatics*, 180, 995–999.
- Auffray, C., Chen, Z., & Hood, L. (2009). Systems medicine: The future of medical genomics and healthcare. *Genome Medicine*, 1(1), 2–2.
- Bakker, M., & Tsui, K. L. (2017). Dynamic resource allocation for efficient patient scheduling: A data-driven approach. *Journal of Systems Science and Systems Engineering*, 26(4), 448–462.
- Bouarfa, L., & Dankelman, J. (2012). Workflow mining and outlier detection from clinical activity logs. *Journal of Biomedical Informatics*, 45(6), 1185–1190.
- Bricage, P. (2017). Use of chronolithotherapy for better individual healthcare and welfare. *Journal of Systems Science and Systems Engineering*, 26(3), 336–358.
- Chen, G. Q., Wu, G., Gu, Y. D., Lu, B. J., & Wei, Q. (2018). The challenges for big data driven research and applications in the context of managerial decision-making--paradigm shift and research directions. *Journal of Management Science in China*, 169(7), 6–15; In Chinese.
- Chen, H. C., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, 36(4), 1165–1188.
- Chen, J. D., Yuan, P. J., Zhou, X. J., & Tang, X. J. (2016). Performance comparison of TF*IDF, LDA and paragraph vector for document classification. In J. Chen, Y. Nakamori, W. Y. Yue, & X. J. Tang (Eds.), *Knowledge and systems sciences, KSS 2016, communications in computer and information science* (Vol. 660, pp. 225–235). Springer.
- Chen, J. F., Guo, C. H., Lu, M. L., & Ding, S. Y. (2022). Unifying diagnosis identification and prediction method embedding the disease ontology structure from electronic medical records. *Frontiers in Public Health*, 9, 793801.
- Chen, J. F., Guo, C. H., Sun, L. L., & Lu, M. L. (2018). Mining typical drug use patterns based on patient similarity from electronic medical records. In J. Chen, Y. Yamada, M. Ryoke, & X. J. Tang (Eds.), *Knowledge and systems sciences, KSS 2018, communications in computer and information science* (Vol. 949, pp. 71–86). Springer.
- Chen, J. F., Guo, C. H., Sun, L. L., & Lu, M. L. (2019). Mining typical treatment duration patterns for rational drug use from electronic medical records. *Journal of Systems Science and Systems Engineering*, 28(5), 602–620.
- Chen, J. F., Sun, L. L., Guo, C. H., Wei, W., & Xie, Y. M. (2018). A data-driven framework of typical treatment process extraction and evaluation. *Journal of Biomedical Informatics*, 83, 178–195.
- Chen, J. F., Sun, L. L., Guo, C. H., & Xie, Y. M. (2020). A fusion framework to extract typical treatment patterns from electronic medical records. *Artificial Intelligence in Medicine*, 103, 101782. <https://doi.org/10.1016/j.artmed.2019.101782>

- Chen, J. F., Wei, W., Guo, C. H., Tang, L., & Sun, L. L. (2017). Textual analysis and visualization of research trends in data mining for electronic health records. *Health Policy and Technology*, 6(4), 389–400.
- Chen, J. G., Li, K. L., Rong, H. G., Bilal, K., Yang, N., & Li, K. Q. (2018). A disease diagnosis and treatment recommendation system based on big data mining and cloud computing. *Information Sciences*, 435, 124–149.
- Cho, S. G., & Kim, S. B. (2017). Feature network-driven quadrant mapping for summarizing customer reviews. *Journal of Systems Science and Systems Engineering*, 26(5), 646–664.
- Dang, T. T., & Ho, T. B. (2017). Sequence-based measure for assessing drug-side effect causal relation from electronic medical records. In J. Chen, T. Theeramunkong, T. Supnithi, & X. J. Tang (Eds.), *Knowledge and systems sciences, KSS 2017, communications in computer and information science* (Vol. 780, pp. 53–65). Springer.
- Diao, X. L., Huo, Y. N., Zhao, S. A., Yuan, J., Cui, M., Wang, Y. X., Lian, X. D., & Zhao, W. (2021). Automated ICD coding for primary diagnosis via clinically interpretable machine learning. *International Journal of Medical Informatics*, 153, 104543. <https://doi.org/10.1016/j.ijmedinf.2021.104543>
- Esfandiari, N., Babavalian, M. R., Moghadam, A. M. E., & Tabar, V. K. (2014). Knowledge discovery in medicine: Current issue and future trend. *Expert Systems with Applications*, 41(9), 4434–4463.
- Groves, P., Kayyali, B., Knott, D., & Kuiken, S. V. (2013). The “big data” revolution in healthcare: Accelerating value and innovation. *McKinsey Quarterly*, 2(3), 1–19.
- Guo, C. H., & Chen, J. F. (2019). Big data analytics in healthcare: Data-driven methods for typical treatment pattern mining. *Journal of Systems Science and Systems Engineering*, 28(6), 694–714.
- Guo, C. H., Du, Z. L., & Kou, X. Y. (2018). Products ranking through aspect-based sentiment analysis of online heterogeneous reviews. *Journal of Systems Science and Systems Engineering*, 27(5), 542–558.
- Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann Publishers.
- Haque, A., Milstein, A., & Fei-Fei, L. (2020). Illuminating the dark spaces of healthcare with ambient intelligence. *Nature*, 585, 193–202.
- Herman, J. (1994). The unifying diagnosis. *Scandinavian Journal of Primary Health Care*, 12(2), 68–69.
- Hey, T., Tansley, S., & Tolle, K. (2009). *The fourth paradigm: Data-intensive scientific discovery*. Microsoft Research.
- Hirano, S., & Tsumoto, S. (2014). Mining typical order sequences from EHR for building clinical pathways. In W. C. Pend et al. (Eds.), *Trend and applications in knowledge discovery and data mining, LNAI 8643, PAKDD 2014* (pp. 39–49). Springer Charm.
- Hoang, K. H., & Ho, T. B. (2019). Learning and recommending treatments using electronic medical records. *Knowledge-Based Systems*, 181, 104788. <https://doi.org/10.1016/j.knosys.2019.05.031>
- Hopp, W. J., Li, J., & Wang, G. H. (2018). Big data and the precision medicine revolution. *Production and Operations Management*, 27(9), 1647–1664.
- Htun, H. H., & Sornlertlamvanich, V. (2017). Concept name similarity measure on SNOMED CT. In J. Chen, T. Theeramunkong, T. Supnithi, & X. J. Tang (Eds.), *Knowledge and systems sciences, KSS 2017, communications in computer and information science* (Vol. 780, pp. 76–90). Springer.
- Huang, Z. X., Dong, W., Bath, P., Ji, L., & Duan, H. L. (2015). On mining latent treatment patterns from electronic medical records. *Data Mining and Knowledge Discovery*, 29(4), 914–949.
- Huang, Z. X., Dong, W., Ji, L., Gan, C. X., Lu, X. D., & Duan, H. L. (2014). Discovery of clinical pathway patterns from event logs using probabilistic topic models. *Journal of Biomedical Informatics*, 47, 39–57.
- Huang, Z. X., Lu, X. D., Duan, H. L., & Fan, W. (2013). Summarizing clinical pathways from event logs. *Journal of Biomedical Informatics*, 46(1), 111–127.

- Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: Towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6), 395–405.
- Ji, G. J., Hu, L. M., & Tan, K. H. (2017). A study on decision-making of food supply chain based on big data. *Journal of Systems Science and Systems Engineering*, 26(2), 183–198.
- Jin, B., Yang, H. Y., Sun, L. L., Liu, C. R., Qu, Y., & Tong, J. N. (2018). A treatment engine by predicting next-period prescriptions. *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 1608–1616), 19–23 August, London.
- Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. W. H., et al. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3, 160035. <https://doi.org/10.1038/sdata.2016.35>
- Lakshmanan, G. T., Rozsnyai, S., & Wang, F. (2013). Investigating clinical care pathways correlated with outcomes. In F. Daniel, J. Wang, & B. Weber (Eds.), *Business process management* (Lecture notes in computer science) (Vol. 8094, pp. 323–338). Springer.
- Li, X., Mei, J., Liu, H. F., Yu, Y. Q., Xie, G. T., Hu, J. Y., & Wang, F. (2015). Analysis of care pathway variation patterns in patient records. *Studies in Health Technology & Informatics*, 210, 692–696.
- Liang, J. J., Goodsell, K., Grogan, M., & Ackerman, M. J. (2016). LMNA-mediated arrhythmogenic right ventricular cardiomyopathy and Charcot-Marie-tooth type 2B1: A patient-discovered unifying diagnosis. *Journal of Cardiovascular Electrophysiology*, 27(7), 868–871.
- Lynch, C. A. (2008). Big data: how do your data grow? *Nature*, 455(7209), 28–29.
- Malhi, G. S., Bell, E., Boyce, P., Mulder, R., & Porter, R. J. (2020). Unifying the diagnosis of mood disorders. *Australian & New Zealand Journal of Psychiatry*, 54(6), 561–565.
- Mans, R., Schonenberg, H., Leonardi, G., Panzarasa, S., Cavallini, A., Quaglini, S., & Van Der Aalst, W. (2008). Process mining techniques: An application to stroke care. *Studies in Health Technology and Informatics*, 136, 573–578.
- Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.
- Miller, K. (2012). Big data analytics in biomedical research. *Biomedical Computation Review*, 2, 14–21.
- MIT Critical Data. (2016). *Secondary analysis of electronic health records*. Springer.
- Naem, M., Jamal, T., Diaz-Martinez, J., Butt, S. A., Montesano, N., et al. (2022). Trends and future perspective challenges in big data. In J. S. Pan, V. E. Balas, & C. M. Chen (Eds.), *Advances in intelligent data analysis and applications, LNCS12080, IDA2020* (pp. 309–325). Springer.
- Niaksu, O. (2015). CRISP data mining methodology extension for medical domain. *Baltic Journal of Modern Computing*, 3(2), 92–109.
- Perer, A., Wang, F., & Hu, J. Y. (2015). Mining and exploring care pathways from electronic medical records with visual analytics. *Journal of Biomedical Informatics*, 56, 369–378.
- Rebuge, Á., & Ferreira, D. R. (2012). Business process analysis in healthcare environments: A methodology based on process mining. *Information Systems*, 37(2), 99–116.
- Sareen, J., Olafson, K., Kredentser, M. S., Bienvenu, O. J., Blouw, M., et al. (2020). The 5-year incidence of mental disorders in a population-based ICU survivor cohort. *Critical Care Medicine*, 48(8), e675–e683.
- Shi, Y. (2014). Big data history, current status, and challenges going forward. *The Bridge*, 44(4), 6–11.
- Shortliffe, E. H., & Cimino, J. J. (2006). *Biomedical informatics: Computer applications in health care and biomedicine* (3rd ed.). Springer.
- Sloan, E. A., Chiang, J., Villanueva-Meyer, J. E., Alexandrescu, S., et al. (2021). Intracranial mesenchymal tumor with FET-CREB fusion—A unifying diagnosis for the spectrum of intracranial myxoid mesenchymal tumors and angiomatoid fibrous histiocytoma-like neoplasms. *Brain Pathology*, 31(4), e12918.
- Sun, L. L., Chen, G. Q., Xiong, H., & Guo, C. H. (2017). Cluster analysis in data-driven management and decisions. *Journal of Management Science and Engineering*, 2(4), 227–251.

- Sun, L. L., Guo, C. H., Liu, C. R., & Xiong, H. (2017). Fast affinity propagation clustering based on incomplete similarity matrix. *Knowledge and Information Systems*, 51(3), 941–963.
- Sun, L. L., Jin, B., Yang, H. Y., Tong, J. N., Liu, C. R., & Xiong, H. (2019). Unsupervised EEG feature extraction based on echo state network. *Information Sciences*, 475, 1–17.
- Sun, L. L., Liu, C. R., Chen, G. Q., Guo, C. H., Xiong, H., & Xie, Y. M. (2021). Automatic treatment regimen design. *IEEE Transactions on Knowledge and Data Engineering*, 33(11), 3494–3506.
- Sun, L. L., Liu, C. R., Guo, C. H., Xiong, H., & Xie, Y. M. (2016). Data-driven automatic treatment regimen development and recommendation. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1865-1874), August 13–17, San Francisco.
- Tien, J. M., & Goldschmidt-Clermont, P. J. (2009). Healthcare: A complex service system. *Journal of Systems Science and Systems Engineering*, 18(3), 257–282.
- Topol, E. J. (2015). *The patient will see you now: The future of medicine is in your hands*. Basic Books.
- van Elten, H. J., Sülz, S., van Raaij, E. M., & Wehrens, R. (2022). Big data health care innovations: Performance dashboarding as a process of collective sensemaking. *Journal of Medical Internet Research*, 24(2), e30201.
- Wang, Y. Q., Qian, L. Q., Li, F. Z., & Zhang, L. (2018). A comparative study on shilling detection methods for trustworthy recommendations. *Journal of Systems Science and Systems Engineering*, 27(4), 458–478.
- World Health Organization. (2012). *The pursuit of responsible use of medicines: Sharing and learning from country experiences*. WHO/EMP/MAR/2012.3, Geneva, Switzerland.
- Wright, A. P., Wright, A. T., McCoy, A. B., & Sittig, D. F. (2015). The use of sequential pattern mining to predict next prescribed medications. *Journal of Biomedical Informatics*, 53, 73–80.
- Wu, X. D., Chen, H. H., Wu, G. Q., Liu, J., et al. (2015). Knowledge engineering with big data. *IEEE Intelligent Systems*, 30(5), 46–55.
- Wu, Y. F., Zeng, M., Fei, Z. H., Yu, Y., Wu, F. X., & Li, M. (2022). KAICD: A knowledge attention-based deep learning framework for automatic ICD coding. *Neurocomputing*, 469, 376–383.
- Xu, N., & Tang, X. J. (2018). Generating risk maps for evolution analysis of societal risk events. In J. Chen, Y. Yamada, M. Ryoake, & X. J. Tang (Eds.), *Knowledge and systems sciences, KSS 2018, communications in computer and information science* (Vol. 949, pp. 115–128). Springer.
- Yadav, P., Steinbach, M., Kumar, V., & Simon, G. (2018). Mining electronic health records (EHRs): A survey. *ACM Computing Surveys*, 50(6), 1–40.
- Yang, S., Dong, X., Sun, L. L., Zhou, Y. C., Farneth, R. A., Xiong, H., Burd, R. S., & Marsic, I. (2017). A data-driven process recommender framework. *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 2111–2120), August 13–17, Halifax NS, Canada.