Title	電子カルテを用いた治療の学習と推薦に関する研究
Author(s)	Hoang, Khanh Hung
Citation	
Issue Date	2019-09
Type	Thesis or Dissertation
Text version	ETD
URL	http://hdl.handle.net/10119/16179
Rights	
Description	Supervisor:Dam Hieu Chi, 先端科学技術研究科, 博士

A Study on Learning and Recommending Treatments Using Electronic Medical Records

Hoang Khanh Hung

Japan Advanced Institute of Science and Technology

Doctoral Dissertation

A Study on Learning and Recommending Treatments Using Electronic Medical Records

Hoang Khanh Hung

Supervisor: Assoc. Prof. Dam Hieu Chi

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
Knowledge science

September, 2019

Abstract

Data mining has shown to be a promising technique with many applications in various domains. In healthcare, accessible electronic medical records provide valuable resources for data mining tasks to address health-related issues. The two emerging tasks are learning treatment patterns and recommending treatments, which are expected to assist healthcare organizations and physicians to manage the use of medical resources and minimize accidental faults causing adverse drug reactions.

Although many data-driven models have been proposed for learning and recommending treatments, approaching these tasks is still very challenging due to several reasons. First, electronic medical records are heterogeneous, longitudinal and varying length objects. These characteristics pose challenges of data processing and representation, the important steps of the knowledge discovery process. Second, although solving healthcare issues typically requires a lot of domain knowledge, related studies have mainly developed black box models that neglect this factor. For example, few studies have focused on explaining the recommendation mechanism from the healthcare perspective or identifying treatment period intervals hidden in prescription records of acute disease patients. As a result, the lack of domain knowledge incorporation considerably weakens the interpretability of current studies.

This dissertation aims to propose a class of treatment learning and treatment recommendation methods to tackle the above challenges. Different from most of the current studies, our proposed methods take into account various patient information to maximize the capability of data utilization. To overcome the challenge of presenting mixed-type medical objects, we adopt a powerful data representation model named mixed-variate restricted Boltzmann machine for representing various patient information. We also address the challenge in handling longitudinal and varying length prescription records partially by a scoring algorithm that splits prescription records into periods where significant changes in prescription indication happen. In the treatment learning method, we propose an algorithm to fully reflect usage frequency of prescription drugs under a tree form. In the treatment recommendation methods, we propose a class of neighbor-based approaches to synthesize neighbor patients' treatments and suggest treatment for new patients.

The experimental evaluations show that the proposed treatment learning method can reveal many more different kinds of treatment patterns together with more interesting results connecting the curing relation of drugs and symptoms compared to traditional approaches using association analysis for treatment pattern discovery. In the case of treatment recommendation methods, we obtain competitive results with advanced recommendation systems for implicit feedback dataset in terms of precision and recall. More importantly, we point out that there are plenty of rooms for developing neighbor-based recommendation approaches that achieve similar precision and better interpretation compared to the black box models.

Keywords– data mining, treatment learning, treatment pattern, treatment recommendation, electronic medical records

Acknowledgements

Firstly, I would like to express my deepest gratitude to my main supervisor, Professor Ho Tu Bao, who kindly supported me during my Ph.D. journey. Professor Ho with his great research experience gave me a lot of informative advice not only on solving my research theme but also on formulating a good research problem that would help me can become an independent researcher. Although he retired last year, I can not thank him enough for his continued support so far. Without his encouragement and guidance, I would have never accomplished my research.

I also would like to extend my gratitude to Professor Dam Hieu Chi, who is my supervisor after Professor Ho's retirement. I feel grateful for his constructive advice in many aspects as well as his kindness in providing me a great research atmosphere to finish my study.

I also would like to express my appreciation to an adviser for my minor research, Professor Huynh Van Nam. He gave me useful discussion and guidance that not only broaden my research viewpoint but also assist me to solve many subtasks in my main theme.

I want to thank all members of Ho Laboratory and Dam Laboratory for friendly discussions on my research problem. We shared many things in daily life that relieved our stress from research work. Working and sharing with them are unforgettable memories in my life.

I wish to express my deep acknowledgements to the committee members consisting of Prof. Ho Tu Bao, Prof. Hiroshi Motoda, Prof. Naoshi Uchihira, Prof. Huynh Van Nam, Assoc. Prof. Dam Hieu Chi, who gave valuable comments to help me improve the quality of my dissertation.

I also would like to express my gratitude to MEXT for kindly providing me financial support during my Ph.D. In addition, I would like to thank Asian Office of Aerospace R&D for the additional financial support under research projects.

Lastly, I would like to give special thanks to all members of my family for their love, understanding and encouragement.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Disease diagnosis and treatment are essential aspects of healthcare. Disease diagnosis is a process of determining a preexisting set of categories agreed upon by the medical profession to designate a specific condition [Jutel, 2009] while a comprehensive treatment is a normalized care plan where therapeutic interventions and medicines for a particular disease are organized on a timeline [Healy et al., 1998, Ireson, 1997]. While the diagnostic prediction problem has been studied intensively over the past decades [Ma et al., 2017, Palaniappan and Awang, 2008, Pattekari and Parveen, 2012, Ordonez, 2006, Wilson and Evans, 1993, Soni et al., 2011], the treatment learning and recommendation problem is still in the early development stage [Gräßer et al., 2017, Mei et al., 2015, Wang et al., 2018, Jin et al., 2018].

Recently, addressing the second problem is becoming more urgent. Studies have pointed out that about 7% of patients in the United States experienced adverse drug reactions (ADRs) where 0.32% among them are fatal events that cost an estimate of 140 billion dollars every year [Lazarou et al., 1998]. Therefore, improving the quality of treatment plays a central role to minimize accidental faults leading to the ADRs.

Besides, medical resources allocated to hospitals [Ubel et al., 1996, Rushing, 1971, Blumstein, 1980] are often limited within the allocated budget.

In many situations, hospitals may need to replace known treatments by alternative ones that fit the available resources. Moreover, even under similar diagnostic codes, patients often suffer various symptoms that require being treated flexibly. Capturing common patterns among various treatments in practice turns out to be helpful to assist not only managers in managing their resources [Newdick et al., 2005] but also inexperience physicians in grasping treatment patterns often used in their organization.

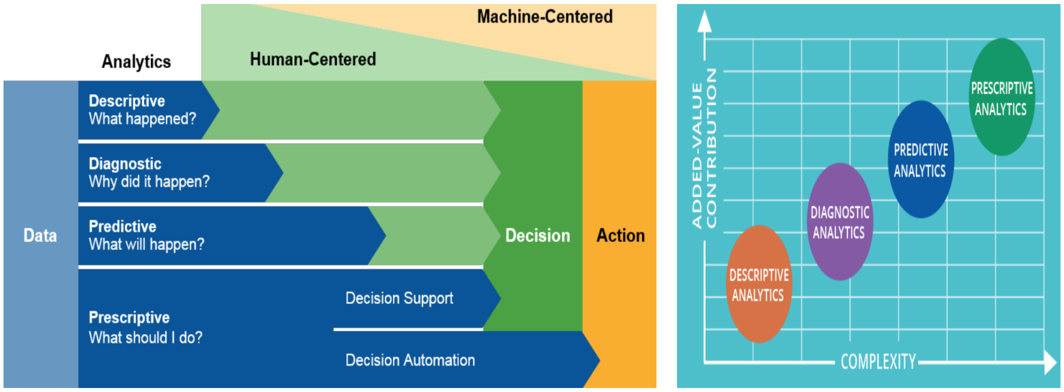
Principally, one can learn treatment for a particular disease through medical domain knowledge, for instance, a piece of written information available in the literature [Group et al., 2009, Association et al., 2008, Wada et al., 2013]. As there is a wide range of domain knowledge, the knowledge-driven approach may require much time and effort to absorb. In recent years, the fast development of electronic medical records has enabled addressing the problem via the data-driven approach where one can derive treatment patterns and recommend treatments for new patients automatically from a massive amount of patient medical records. The goal of this dissertation is to develop a series of domain interpretable data-driven methods for learning and recommending treatments. Once successfully developed, our methods are expected to assist physicians to address the above needs of discovering treatment patterns and improving treatment's quality.

1.2 Challenges

Problem viewpoint

Compared to the diagnostic prediction problem, the problem of learning treatment patterns and recommending treatments is more challenging for several reasons.

First, these problems are two types of data analytics that have different levels of complexity. According to the description of the four analytic tasks described in Figure 1.1, the diagnostic prediction problem can be categorized as the predictive analytics in which data-driven models try to predict patients' disease codes based on the observed symptoms and/or predefined knowledge about diseases. Compared to the diagnostic prediction problem, the treatment recommendation problem takes a higher level of problem complexity. In this problem, the models have to learn and prescribe what action to take given observed data, i.e. which drugs should be used given observed symptoms and obtained values in laboratory test exams, namely indicators. Therefore, the later problem can be categorized as the prescriptive analytics which may require deep domain incorporation to gain useful insights for decision-making.



(a) (b)

Figure 1.1: Four types of data analytics (a) and level of complexity (b) (source: Gartner, 2014)

Second, the two problems tackle two sets of patient features that are different in terms of complexity. In the diagnostic prediction problem, data-driven models typically take into account a set of observed signs/symptoms and measured indicators to predict patient disease codes. While the diagnostic prediction problem mainly deals with static features, the treatment learning and recommendation problem involves different kinds of patient records in addition to the above data, especially prescription records which are heterogeneous, temporal and varying in length. Obviously, data-driven models for the treatment recommendation problem have to deal with a much more complicated set of features than those of the diagnostic prediction problem.

Technical viewpoint

Recently, many data-driven models have been proposed to address the treatment learning and recommendation problem. Although some progress has been made, solving this problem is still very challenging from the technical viewpoint.

One of the challenges is maximizing the capability of data utilization in current data-driven models. Intuitively, patients who share many symptom features, indicator features, and demographic features are likely to be treated similarly. Unfortunately, most current studies have merely exploited limited patient information. The unavailability of rich-feature medical records could be attributed to the difficulty in collecting healthcare data [Char et al., 2018, Jensen et al., 2012]. Even if different kinds of patient records are made fully available, they usually exist in different data types that are not ready to feed to traditional machine learning methods.

In addition to the data utilization issue, it is also challenging to capture medical domain knowledge hidden in prescription records. For instance, it is well-known that treatments of patients, especially those with acute diseases, are usually divided into several stages named treatment periods, which may not be stated explicitly in varying-length prescription records. Since the derivation of treatment patterns highly depends on patients' prescription drugs in each period, identifying unseen treatment periods become a non-neglectable step in the treatment learning method.

Domain viewpoint

Although data mining has been applied successfully in many fields, the number of applications for healthcare, especially for the treatment issue, is still underdeveloped. This drawback could be attributed to the fact that current models are facing challenges not only in terms of accuracy but also in terms of interpretability. Since patients are objects that are directly affected by any single decision from physicians, it is very important for the models to provide domain experts informative insights and explainable mechanism of the derived knowledge. Unfortunately, many advanced methods developed recently have underestimated this issue with black box models which are difficult to explain in the healthcare perspective. We take the case of deep learning based models for treatment recommendation problem as an example. In such models, it is very difficult to answer which known patients whom a new patient's treatment is based on and how the recommended treatment is derived. This drawback prevents physicians from being c. As a result, the gap between research work and practical use has not been eliminated.

1.3 Tasks and Tentative Approaches

The objective of this dissertation is to pursue the following two tasks.

First, we aim to construct a treatment learning method which derives treatment patterns for a patient group. Our learning method address the above challenges in terms of domain viewpoint and technical viewpoint. Given a patient cohort, we divide it into clusters named subcohorts where treatment patterns over periods for each subcohort are discovered subsequently. To tackle the issue of representing mixed-type records, we adopt a representation model named mixed-variate restricted Boltzmann machine (MV.RBM) [Tran et al., 2014], which is robust in transforming mixed-type objects to their homogeneous representation. To address the issue of treatment period identification, we propose an algorithm which captures considerable changes in prescription drugs' indications. More interestingly, we construct prescription trees based on prescription drugs' frequencies to leverage treatment patterns for each subcohort.

Second, we aim to develop a class of neighbor-based recommendation methods that suggests top M drugs for new patients by taking into account treatment patterns extracted from learned prescription trees of subcohorts consisting of neighbor patients. Our proposed recommendation methods capture the intuition that a new patient's treatment can be learned from similar cases.

1.4 Problem Formulation

We consider a cohort of patients $\{p_1, p_2, ..., p_N\}$ having the same diagnosis codes. Each p_i is a heterogeneous object which consists of different data components describing information about basic demographics $Info_{p_i}$, laboratory examination data Lab_{p_i}, nursing notes $Note_{p_i}$, and prescription data $Presc_{p_i}$. It is noted that the components Lab_{p_i}, $Note_{p_i}$ and $Presc_{p_i}$ are longitudinal over $n^l_{p_i}$, $n^n_{p_i}$, $n^p_{p_i}$ timestamps, respectively. Each component is a set of features that could be detailed further. For instance, the $Info$ component of patient p_i can be decomposed as below.

$$Info_{p_i} = \{Info^{age}_{p_i},\ Info^{gender}_{p_i}, Info^{marriage}_{p_i},\ Info^{histIllness}_{p_i}, ...\}$$

It describes in detail the age, gender, martial status, history of illness, to name a few, of patient p_i. The component $Presc_{p_i}$ describes information about prescription drugs over $n^p_{p_i}$ timestamps. It can be decomposed as follows.

$$Presc_{p_i} = \{Presc^{tp_1}_{p_i},\ Presc^{tp_2}_{p_i}, ...,\ Presc^{tp_{n^p_{p_i}}}_{p_i}\}$$

where each $Presc^{tp_k} = \{dr^{tp_k}_1,\ dr^{tp_k}_2 ...\}$ is a set of drugs prescribed at timestamp tp_k. Each drug $dr =< name,\ startdate,\ enddate,\ dosage,\ route >$ is a compound object characterized by information about drug name, starting date, ending date of usage and the route describing how that drug was delivered. The component $Note_{p_i}$ contains nursing notes written in text format about p_i's treatment progress over $n^n_{p_i}$ timestamps.

$$Note_{p_i} = \{Note^{tn_1}_{p_i},\ Note^{tn_2}_{p_i}, ...,\ Note^{tn_{n^n_{p_i}}}_{p_i}\}$$

The component Lab_{p_i} describes different measurement values of patient condition in $n^l_{p_i}$ timestamps.

$$Lab_{p_i} = \{Lab^{tl_1}_{p_i},\ Lab^{tl_2}_{p_i}, ...,\ Lab^{tl_{n^l_{p_i}}}_{p_i}\}$$

where each $Lab^{tl_k} = \{in^{tl_k}_1,\ in^{tl_k}_2, ...\}$ is a set of indicator values, i.e lab exams that need for detecting a disease, at timestamp tl_k. Each $in =< name, value >$ is an indicator characterized by its name and value.

This research aims to address the following tasks.

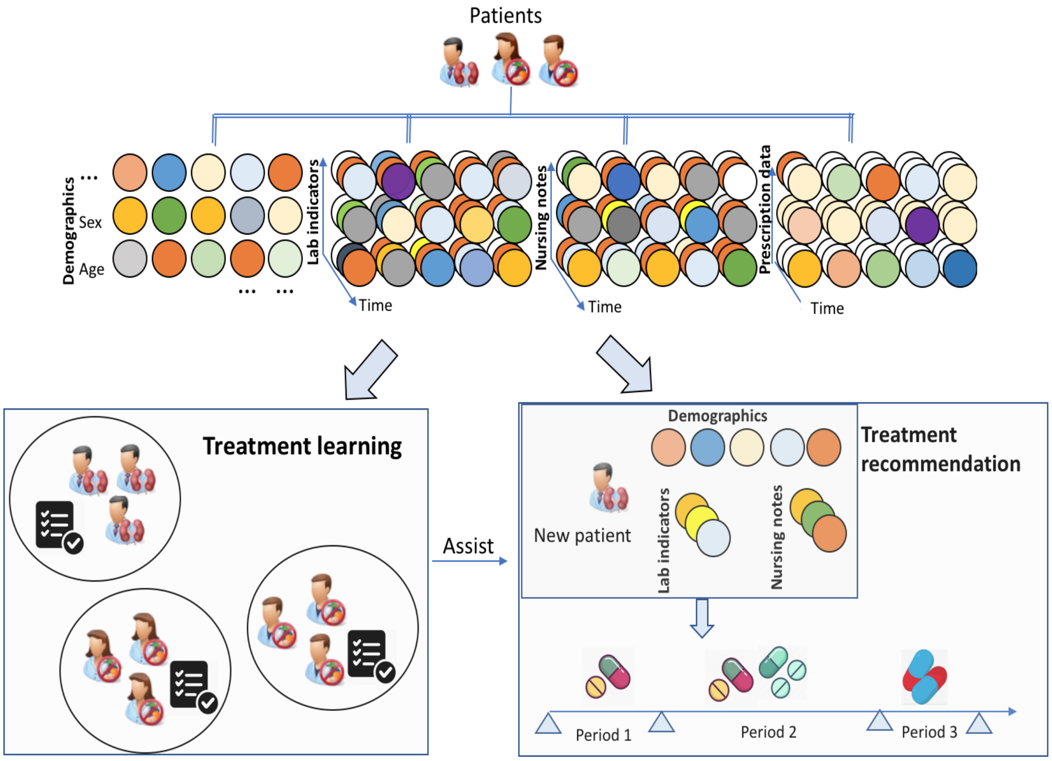


Figure 1.2: The two research problems conducted in this dissertation

1. Construct a treatment learning method that utilizes relevant features from the components $\{Info,\ Lab,\ Note,\ Presc\}$ to cluster a patient cohort into subcohorts and learn treatment patterns of each sub-cohort over n periods $\tau_1,\ \tau_2, ...,\ \tau_n$. We note that the n treatment periods are not given in advance and determining them is considered as a subtask of this method.

2. Construct a treatment recommendation method to recommend top M drugs that could be prescribed over n periods for new patients.

We denote the patients whose medical records are used in the treatment learning method as “training patients”. The terms “new patients” and “testing patients” are used interchangeably in this dissertation.

Figure 1.2 provides a general picture of two research problems that have been presented so far. Input data, the purpose of solving each research problem and their relation are simply illustrated.

1.5 Contribution

In short, we propose methodologies specially designed for the treatment learning and treatment recommendation problem. The main contributions of this thesis are listed as follows.

1. First, by adopting a MV.RBM for learning a homogeneous representation of mixed-type patient records, we encourage exploiting various relevant features that make data-driven models more robust in representing patient data.

2. Second, this work employs both knowledge-driven and data-driven approaches. The pieces of incorporated knowledge are prescription drugs' indications as well as their importance in treating diseases. The incorporated domain knowledge is shown to be flexible to split the longitudinal prescription data into reasonable intervals.

3. Third, we propose a framework to label prescription drugs' indications from external medical domain resources. Our framework is useful for not only identifying patients' treatment periods but also interpreting typical treatment patterns of each subcohort.

4. Fourth, our thesis also propose a new representation, namely prescription tree, to discover treatment patterns of each subcohort. Each path in a prescription tree summarizes the frequency of a sequence of prescription drugs. The tree is not only meaningful for physicians to capture frequent and infrequent prescription drugs but also helpful to assist the treatment recommendation method in finding typical prescription drugs of neighbor patients.

5. Fifth, we propose a class of recommendation methods which recommend prescription drugs for new patients based on neighbor patients'

treatments. Our methods emphasize the explainability of the recommendation mechanism under domain perspective. The experimental evaluation shows that even identifying neighbor patients is almost uncertain in many cases, our recommendation methods are still able to achieve competitive results compared to some well-known but hard domain-interpretable recommendation systems designed for implicit feedback datasets.

The first four contributions are reflected through Chapter 4 of this dissertation while the last one is expressed through Chapter 5 and Chapter 6.

1.6 Organization

The remaining chapters of this dissertation are organized as follows.

Chapter 2 shows a general picture of research studies on treatment learning and treatment recommendation problem. In this chapter, a series of related work on clinical pathway mining, a similar problem to the treatment learning problem is briefly summarized. We discuss the characteristics of each problem as well as the challenges of the later one. We also review typical approaches and their limitation in solving the treatment learning and recommendation problem, including the probabilistic-based approach, deep learning-based approach, reinforcement-based approach, and frequency-based approach.

Chapter 3 provides some background of this dissertation. We briefly introduce basic knowledge about the Unified Medical Language System (UMLS), a comprehensive biomedical ontology which will be used for standardized clinical terms in patient records. We also provide basic knowledge about cTAKES, a clinical processing tool which is widely used to extract UMLS terms in clinical notes. To normalize drug names and label drug indication, we present some insights about DrugBank database, a very large drug database describing various information of most prescription drugs. Lastly, this chapter introduces some background regarding the model, inference and training steps of restricted Boltzmann machine (RBM), a powerful data representation model for binary data. By providing some background on the RBM model, we hope readers can easily grasp the idea of MV.RBM, an extension of the RBM model for mixed-type data that will be presented in the subsequent chapter.

Chapter 4 presents our proposed methods for the treatment learning problem. We first introduce a summary of the work followed by the detailed steps of data processing, patient clustering, treatment period identification, prescription tree construction. We then show the experimental evaluation including analysis and possible interpretation of the outputs.

Chapter 5 presents different neighbor-based methods for the treatment recommendation problem. We introduce two different treatment learning aspects named symptom-based learning aspect and treatment-based learning aspect together with recommendation methods on each single learning aspect and both of them. Intensive experimental evaluation on the efficacy of the proposed methods compared to using treatment of the nearest neighbor patient and the baselines is given. We also show the behavior among the proposed recommendation methods in different cases to keep the evaluation objective. The chapter ends with a discussion of some possible hypotheses that can explain for experimental results.

Chapter 6 presents a weighting recommendation method which overcomes some drawback of the non-weighting recommendation method presented in the Chapter 5. We present the detailed steps of our methodology, the experimental evaluation to show the superiority of the weighting recommendation method compared to the non-weighting one and the baselines. Discussion on obtained results is also given in the end of the chapter.

Chapter 7 concludes the thesis and discusses promising research directions for upcoming studies.

Chapter 2

Research Context

2.1 Clinical Pathway Mining

In the early development of healthcare mining, prescription records were not published widely for research purpose. The most related studies at that time focused on mining clinical pathway, a close concept of treatment where research objects are clinical events such as examinations, treatments, prescriptions, nursing visits. Lin et al. [Lin et al., 2001] developed a graph mining technique to discover dependency patterns of clinical pathways for curing brain stroke. Haytham et at. [Elghazel et al., 2007] combined a b-color based framework with Markov model for clinical pathway clustering and prediction. Bouarfa et al. [Bouarfa and Dankelman, 2012] developed a tree-guide and global pair-wise multiple sequence alignment to detect consensus workflows and outliers from clinical activity logs. Chen et al. [Chen et al., 2015] proposed a model to learn and categorize workflows based on their duration for efficient workflow management.

We note that the treatment mining problem and the clinical pathway mining problem share some properties. Both address treatment data represented as a sequence of events, i.e., clinical procedures or medications, at different granularity levels. However, while the research objects of the first problem are a few clinical procedures, those of the second problem could be hundreds of prescription drugs plus additional information regarding prescription dosages and routes. The complicated prescription objects make the treatment mining problem more challenging to tackle compared to the clinical pathway mining problem. Another difference is the degree to which patients' health status is affected by solving each problem. It can be seen that solving the treatment mining problem where research objects are medications is likely to impact on patients' health status more considerably than solving the clinical pathway mining problem.

2.2 Probabilistic-based Approach

In recent years, probabilistic models have emerged as a promising technique for solving many data mining tasks. In the field of healthcare analytics, various probabilistic graphical models have been studied to derive common patterns of clinical pathways and treatments. Huang et al. [Huang et al., 2015] developed a novel topic model demonstrating the association of patients' conditions and their treatments. Lu et al. [Lu et al., 2016] fed different patient information such as as the diagnosis, contextual information and medications into a multiple channel LDA. Xu et al. [Xu et al., 2016] proposed exploiting billing and prescription data to identify pathway patterns. Park et al. [Park et al., 2017] suggested summarizing insurance data via a disease-medicine topic model. Yao et al. [Yao et al., 2018] collected literature resources and proposed a novel topic model to explain how Chinese prescription was generated.

The primary concern of the studies under the light of the probabilistic-based approach is that they often employed many hyper-parameters that difficult to train and may weaken the model interpretability as well.

2.3 Deep Learning-based Approach

Deep learning has recently received great attention from healthcare analytics researchers. Pham et al. [Pham et al., 2016] extended a LSTM model in dynamic context for future outcome prediction. More interestingly, their model was developed to address different tasks, including disease progression modelling, intervention suggestion, and risk stratification.

Le et al. [Le et al., 2018] developed an extended version of the memory network for the sequence to sequence modelling. The augmented network consists of encode-decode controllers which allow feeding patient history data and output corresponding treatments. More recently, an extension of LSTM models for multiple data types, namely multifaceted LSTM, has been suggested by Jin et al. [Jin et al., 2018].

Although deep learning has shown many promising results in different domains, deep learning for healthcare is still questionable for two reasons. First, it is not easy to integrate domain-knowledge into deep neuron networks. Second, most of them are black-box models that is difficult to capture basic intuition in the healthcare viewpoint.

2.4 Reinforcement Learning-based Approach

Reinforcement learning is one of the most well-known approaches to tackle treatment optimization dynamically. In studies using this approach, treatment is typically described as sequences of medicines while patient condition is presented as sequences of states. The goal is to find optimal sequence of treatment such that the reward function is optimized. Zhao et al. [Zhao et al., 2011] utilized a Q-learning method to find optimal medications for non-small cell lung cancer from clinical trials. In that work, the authors used a modified support vector regression to estimate the Q-function value. Liu et al. [Liu et al., 2017] exploited registry data and proposed a deep reinforcement learning method to optimize dynamic treatment regimens. In that work, a two-step learning which is a combination of supervised and deep reinforcement models was proposed to predict clinical procedures overtime. Nemati et al. [Nemati et al., 2016] worked on a case study of heparin dosage optimization using deep reinforcement and Markov models. More recently, Wang et al. [Wang et al., 2018] proposed exploiting patient records by supervised learning and reinforcement learning to optimize treatments over time.

Despite their ability to model complex sequences and interactions dynamically, reinforcement learning studies are often restricted with clinical trials where treatment outcomes are provided in advance. In case of prescription records, such information could be identified through progressing notes or results of laboratory exams. However, automatically inferring associated outcomes for every treatment is a non-trivial task as one may need profound knowledge to analyze sentiment meaning hidden in the clinical context. As a result, such limitation makes the deployment of studies under this approach seems to be impractical for real-world EMRs.

2.5 Frequency-based Approach

Intuitively, treatment patterns can be mined from the information about drug usage's frequency. For clinical pathway mining, Lin et al. [Lin et al., 2001] proposed the formation of dependency graphs which take the frequency of clinical procedures into account. The graphs then were used to identify frequent clinical pathways. Hirano et al. [Hirano and Tsumoto, 2014] built novels maps modelling the occurrence and transisition frequency to reveal important treatment events. In that work, the authors characterized each treatment sequence by a typicalness index and selected the highest typicalness values as candidate patterns. Sun et al. [Sun et al., 2016] studied a group-based recommendation method where patient clusters are grouped based on a new similarity metric between prescription records. Treatment patterns are discovered then using association rule mining. The work, however, did not integrate information such as patient symptoms, the primary causes of treatments, in the formation of discovered treatment patterns. Moreover, the treatment recommendation mechanism in that work required information about the patient outcome which may not be available or difficult to identify in real-world EMRs.

Chapter 3

Background

3.1 Unified Medical Language System

UMLS is a comprehensive resource consisting of databases and programs that aim to standardize biomedical terms and allow computer systems to interact and exchange information in the biomedicine and healthcare domain. It was created and has been updated by the US National Library of Medicine since 1986.

The databases in UMLS are known as Knowledge Sources. They can be used in different tasks such as process, query or retrieve biomedical and healthcare data. The data handled by these Knowledge Sources covers different contexts, for example, medical records, biomedical text in the literature or clinical guidelines. The programs in UMLS consists of various toolkits, which enable developers to use or customize the Knowledge Sources for different purposes.

The following sections present Metathesaurus and Semantic Network, the two components in UMLS Knowledge Sources. We also provide a brief introduction about Lexical Tool, a primary component of UMLS toolkits.

3.1.1 Metathesaurus

Concept and concept identifier

The Metathesaurus is a huge and multi-purpose terminology resource that consists of millions of concepts about biomedical and healthcare domain. It is built from different terminology databases used in the biomedical literature, health billing, patient care processes or public health statistics. These databases are named as "source vocabularies". Licenses of using purpose may be required to fully access some source vocabularies. It is noted that Metathesaurus is also a multi-language database. UMLS communities in many countries have built their own Metathesaurus and expanded them year by year.

The fundamental elements of Metathesaurus are UMLS concepts. Loosely speaking, each UMLS concept represents different biomedical terms and views describing the same meaning. In a given context, each UMLS concept belongs to one category named semantic type defined in the Semantic Network. There may exist useful relationships among the concepts that help to figure out the hierarchical level of them.

Since UMLS concepts are constructed from multiple source vocabularies, there are cases that the same term is used for different concepts. To this end, the Metathesaurus points out both meanings and clearly indicates the source vocabulary deriving each meaning. In case the same concept mentioned in different contexts, the Metathesaurus represents all the contexts. When there are conflicting relationships between two UMLS concepts, all relationships are also presented. In short, the Metathesaurus takes into account all source vocabularies to reflect different views of the same concept that could be useful for information extraction tasks.

In UMLS, a concept represents a specific meaning that can be described by different names. The core task of Metathesaurus is to identify the meaning of each name in each source vocabulary and link those having the same meaning by a concept. This task is facilitated by a group of experts who are supposed to have sufficient knowledge to group synonym terms from multi-source vocabularies with high accuracy.

The Metathesaurus uses different schemes to encode concepts and names. Each UMLS concept is characterized by a unique concept identifier named CUI. It is noted that each CUI alone has no specific meaning and keeps unchanged regardless that possible change of names associated with it. When two CUIs are identified as the same concept, one of them will be removed and all links of the removed one are switched to the retained CUI.

Concept name and string identifier

Each CUI represents a set of concept names describing the same thing. Each concept name is identified by a string unique identifier (SUI). Since each string may have name variants such as lower, upper cases, the Metathesaurus uses separated SUI for each variant concept name. Separated SUIs are also used for the same string that written in different languages.

It is worth noting that the mapping between concept (CUI) and concept name (SUI) are many-to-many mapping. Each UMLS concept typically is linked to many concept names representing the same thing. By contrast, a concept name may have multiple meanings and therefore, may be linked to different concept identifiers.

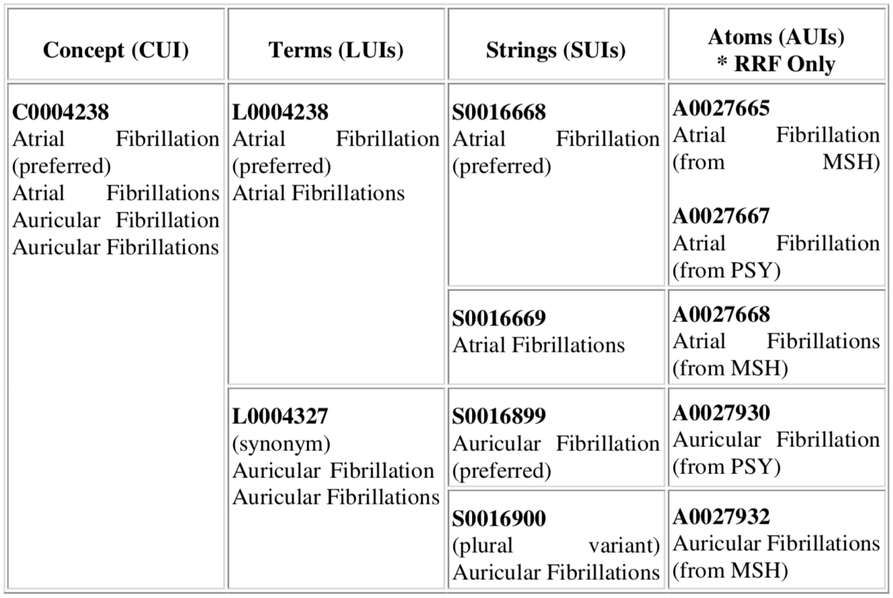
Atoms and atom identifiers

Atoms are the fundamental elements that lead to the construction of concepts and concept names in the Metathesaurus. Each appearance of a concept name in a source vocabulary is marked with a distinct atom identifier (AUI). When the same concept name appears multiple times in the same or different source vocabularies, a unique AUI is assigned for each appearance. These AUIs are linked to the same string unique identifier since they represent for the same instance of SUI. It is noted that while a concept name can be linked to multiple concept identifiers, each atom is only linked to a unique concept identifier since the context of its occurrence has already been determined.

Terms and lexical identifiers

For English vocabulary source, minor variants at the lexical level of similar concept names can be grouped into one unit named lexical unique identifier (LUI). This means that an LUI instance can be the representative for several similar SUIs.

Figure 3.1 illustrates different code schemes, i.e. concept, term, string, and atoms, of the same concept. In the provided illustration, the “Atrial Fibrillation” appears in two source vocabularies which are MSH and PSY. Each occurrence is assigned with a unique AUI and they are all context-aware instances of the same SUI. In the above source vocabularies, “Atrial Fibrillation” may be also mentioned in plural forms with different SUI and AUI. However, since they are simply minor lexical variations, associated SUI and AUI of those variations are also linked to the same LUI of the single form. Additionally, the term “Auricular Fibrillation” can be considered as a synonym of “Atrial Fibrillation” and therefore, is assigned to the same CUI.



Concept (CUI)	**Terms (LUIs)**	**Strings (SUIs)**	**Atoms (AUIs)** *** RRF Only**
C0004238 Atrial Fibrillation (preferred) Atrial Fibrillations Auricular Fibrillation Auricular Fibrillations	**L0004238** Atrial Fibrillation (preferred) Atrial Fibrillations	**S0016668** Atrial Fibrillation (preferred)	**A0027665** Atrial Fibrillation (from MSH) **A0027667** Atrial Fibrillation (from PSY)
		S0016669 Atrial Fibrillations	**A0027668** Atrial Fibrillations (from MSH)
	L0004327 (synonym) Auricular Fibrillation Auricular Fibrillations	**S0016899** Auricular Fibrillation (preferred)	**A0027930** Auricular Fibrillation (from PSY)
		S0016900 (plural variant) Auricular Fibrillations	**A0027932** Auricular Fibrillations (from MSH)

Figure 3.1: An example of different UMLS hierarchical code schemes describing the same concept (source: UMLS documentation)

3.1.2 Semantic network

In addition to the Metathesaurus, UMLS also consists of a Semantic Network which categorizes Metathesaurus CUI into semantic types and defines hierarchical structures or link among them. There are about 133 semantic types and 54 relationships. Nodes in the network represent semantic types and links express relationships that may exist among them. The semantic types cover a wide range of domains such as organisms, anatomical structures, biologic functions, chemicals, events, physical objects, and concepts or ideas.

Each Metathesaurus concept belongs to at least one semantic type. As the semantic relationships are organized in a hierarchical structure, a concept's semantic type is assigned by its most specific meaning. For example, the semantic type of the concept “Macaca” is determined as “Mammal” since there is no more specific semantic types other than “Mammal” among the child nodes of “Primate” appears in the hierarchy. In case a concept does not belong to any semantic types among child nodes, the Semantic Network assigns its semantic type to the most relevant parent node instead of creating a new semantic type for the considering concept. For instance, the “Manufactured Object” has two child nodes, which are “Medical Device” and “Research Device”. It is obvious that there are objects that do not belong to either “Medical Device” or “Research Device”. In such case, the network simply categorizes them to the semantic type “Manufactured Object”.

Two kinds of relationships existing among the Metathesaurus semantic types are “isa” relationships and “non-isa” relationships. Semantic types that belong to “isa” relationship are organized by a hierarchy structure. Figure 3.2 illustrates a portion of the hierarchy for the semantic type “Biologic Function” where each child node is linked to its parent node by an isa relation. The non-isa relations are represented by a non-hierarchical structure. Figure 3.3 gives a small portion of Semantic Network consisting of both “isa” and “non-isa” relations. Some major “non-isa” relations are “physically related to”, “temporally related to”, “conceptually related to”, “functionally related to”, “spatially related to”.

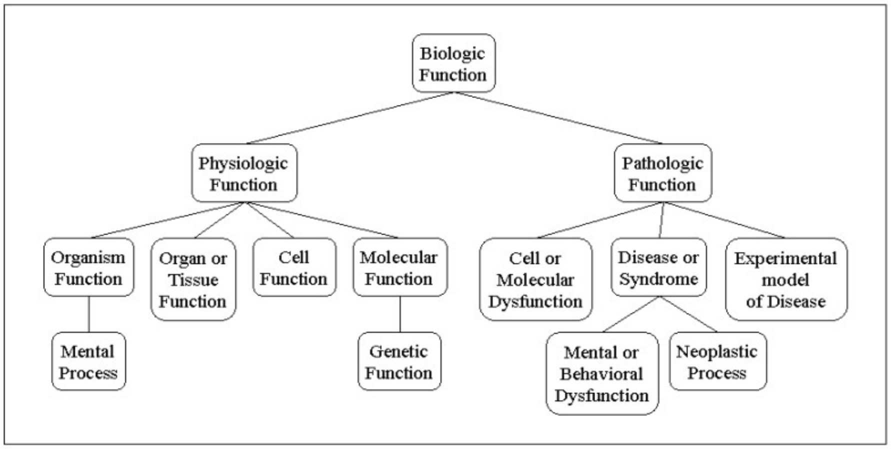


Figure 3.2: A portion of "isa" relationship for the semantic type "Biologic Function" (source: UMLS documentation)

Some rules of passing relationships between semantic types and UMLS concepts are explained as follows. When two semantic types are connected by a "non-isa" relationship, that relationship is also applied to the nodes linked with "isa relationship". For example, the Biological semantic type is linked to the "Organism" by the relationship "process of". Since "Organ" or "Tissue Function" is a "Biological Function" and "Animal" is an "Organism", it can be inferred that the relationship "process of" is also held between the semantic type "Organ" or "Tissue Function" and the semantic type "Animal".

In some cases, the inheritance of relationships are not logical and therefore it is blocked. For instance, under the inheritance mechanism, the semantic type "Mental Process" could be linked to the semantic type "Plant" via "the process of" relationship. However, this relationship is not true in reality since plants are unconscious beings. In such case, the Semantic Network sets explicit links of relationship to semantic types whose child nodes are blocked by that relationship. For example, the "Body System" and "Fully Formed Anatomical Structure" are linked by the relationship "conceptual part of". However, "conceptual part of" should not link "Body System" to "Cell" or "Tissue", the child nodes of "Fully Formed Anatomical Structure" since such

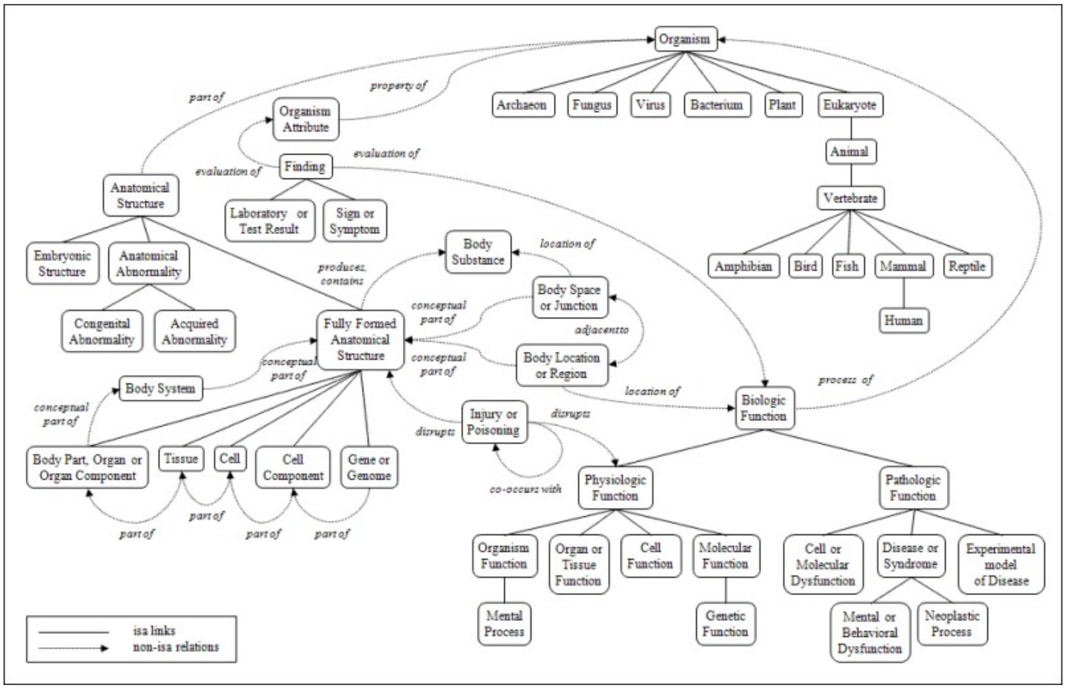


Figure 3.3: A portion of the UMLS Semantic Network that consists both hierarchical relationships of semantic types and non-hierarchical relationships among the semantic types (source: UMLS documentation)

relationship is not meaningful.

It is noted that the relationships between two semantic types may not hold between concepts that belong to those semantic types. For example, the semantic type “Sign” is linked to the “Organism Attribute” by the “evaluation of” relationship. Therefore, signs such as “fever” and “overweight” can be the evaluation of the “Organism Attribute” such as “body temperature” and “body weight” respectively. However, it is obvious that “overweight” is not evaluated by “body temperature” and “fever” is not assessed by “body weight”.

3.1.3 SPECIALIST lexicon

The lexicon consists of biomedical and common English dictionaries. Each term in the lexicon records information about the syntactic, morphological, and orthographic data. Such kinds of data are essential for the lexical tools to use the SPECIALIST NLP system to normalize text, index terms and find their lexical variants. The entries for the lexicon are summarized from different dictionary resources, for instances, Longman's Dictionary of Contemporary English, Collins COBUILD Dictionary, Dorland's Illustrated Medical Dictionary, The Oxford Advanced Learner's Dictionary, and Webster's Medical Desk Dictionary.

3.2 DrugBank Database

3.2.1 Overview

The DrugBank is a thorough database designed for searching information about drugs and their targets. DrugBank stores both detailed drug information regarding their chemistry, pharmacology, indication; and drug targets such as drug structure and pathway. Since DrugBank almost covers various drugs recorded in Wikipedia and their detailed information, it is considered as an encyclopedia rather than a drug database. The DrugBank serves different stakeholders including researchers, pharmacists, chemists who exploit a huge source of drugs and drug targets for different purposes, for instance, drug re-positioning.

The newest version of DrugBank (2018) consists of nearly 12000 drugs of which more than 2500 approved ones are low-weight molecules, nearly 1200 approved drugs are biotech ones and the rest is almost in the investigation phase. More interestingly, those drugs are linked to 5131 protein sequences which provide further information about drug target, enzyme, transporter to name a few. Each drug is described by 200 data fields where half of them is dedicated to the drug/chemistry information and the rest describes drug targets or their protein sequence information.

3.2.2 Construction

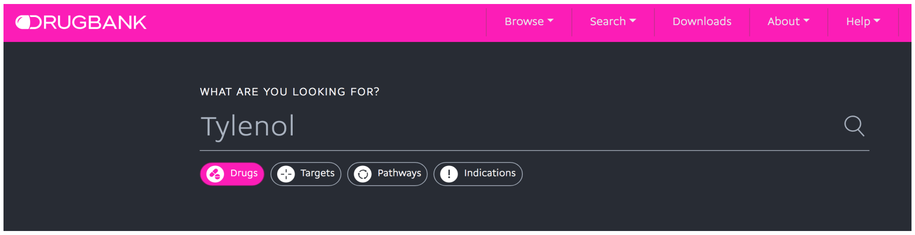
DrugBank database is constructed from multiple sources, either electronic or paper versions. Two kinds of collected data are molecular biology content recorded in Swiss-Prot or UniProt database and chemical-related content found in textbooks. There are dozens of textbooks, hundreds of papers and electronic sources contributing to the construction of DrugBank. Many of these sources are paper versions which require deep domain knowledge to consume. This makes the construction of DrugBank challenging.

The DrugBank teams include expert pharmacists, physicians, bioinformaticians who work together to create drug entries and their detailed information. The drugs were selected by the following criteria: its molecule must be non-redundant, include more than one atom with known chemical structure and be extracted from well-known data sources. Drug entries in DrugBank databases are divided into two major drug-groups, one including FDA-approved drugs which are small molecules, biotech, nutraceuticals or micronutrients and metabolites; and the other group is experimental drugs which are unapproved or illicit.

DrugBank is maintained and updated every year to make sure it is an up-to-date and correct database. Each drug entry is created by one member and supervised by the other member of the team. Additional checks are routinely conducted by senior physicians, pharmacists and biochemists.

3.2.3 DrugBank in use

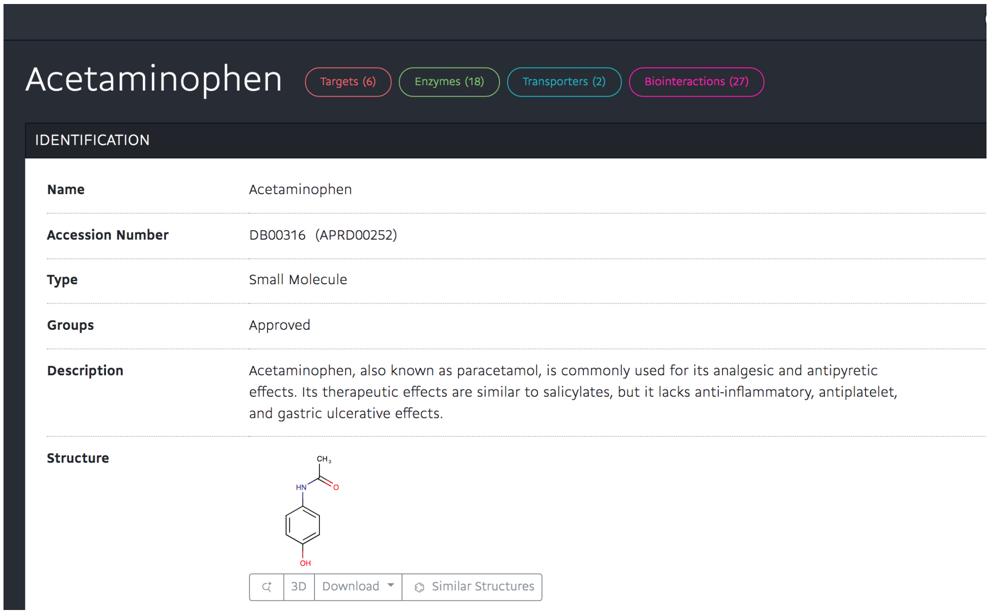
Figure 3.4 presents the searching interface of DrugBank database. It allows users to search by drugs, targets, pathways or indications. Figure 3.5 takes an example of searching results obtained when querying information about the drug acetaminophen. Basic information of the drug, for example, the drug code, drug type, groups, description, chemical structure of drug molecule to name a few is presented clearly. Figure 3.6 shows pharmacology information of the drug acetaminophen. It presents information about the indication of the searching drug, associated conditions which can be used acetaminophen, its pharmacodynamics information. In addition, DrugBank also allows users to search for name variations of the same drug. Figure 3.7 shows synonym drug names of the drug acetaminophen.



WHAT ARE YOU LOOKING FOR?

Tylenol

Figure 3.4: The searching interface of DrugBank database

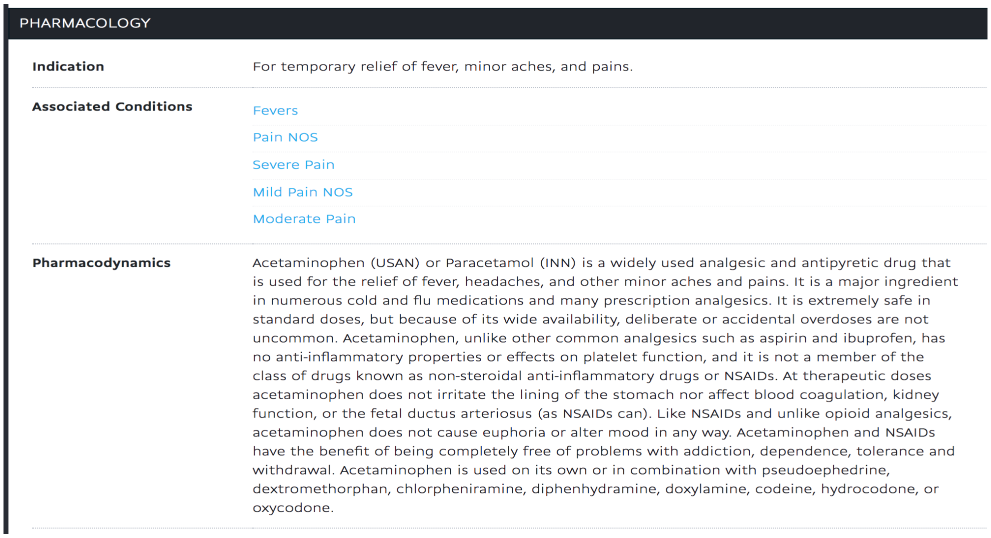


Acetaminophen

IDENTIFICATION

Name	Acetaminophen
Accession Number	DB00316 (APRD00252)
Type	Small Molecule
Groups	Approved
Description	Acetaminophen, also known as paracetamol, is commonly used for its analgesic and antipyretic effects. Its therapeutic effects are similar to salicylates, but it lacks anti-inflammatory, antiplatelet, and gastric ulcerative effects.
Structure	

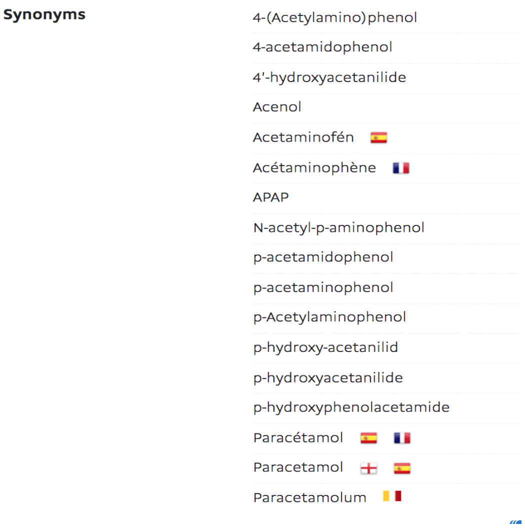
Figure 3.5: A part of searching results showing the basic info of the drug acetaminophen



PHARMACOLOGY

Indication	For temporary relief of fever, minor aches, and pains.
Associated Conditions	Fevers Pain NOS Severe Pain Mild Pain NOS Moderate Pain
Pharmacodynamics	Acetaminophen (USAN) or Paracetamol (INN) is a widely used analgesic and antipyretic drug that is used for the relief of fever, headaches, and other minor aches and pains. It is a major ingredient in numerous cold and flu medications and many prescription analgesics. It is extremely safe in standard doses, but because of its wide availability, deliberate or accidental overdoses are not uncommon. Acetaminophen, unlike other common analgesics such as aspirin and ibuprofen, has no anti-inflammatory properties or effects on platelet function, and it is not a member of the class of drugs known as non-steroidal anti-inflammatory drugs or NSAIDs. At therapeutic doses acetaminophen does not irritate the lining of the stomach nor affect blood coagulation, kidney function, or the fetal ductus arteriosus (as NSAIDs can). Like NSAIDs and unlike opioid analgesics, acetaminophen does not cause euphoria or alter mood in any way. Acetaminophen and NSAIDs have the benefit of being completely free of problems with addiction, dependence, tolerance and withdrawal. Acetaminophen is used on its own or in combination with pseudoephedrine, dextromethorphan, chlorpheniramine, diphenhydramine, doxylamine, codeine, hydrocodone, or oxycodone.

Figure 3.6: The resulting section describing the indication of the drug acetaminophen



Synonyms	4-(Acetylamino)phenol 4-acetamidophenol 4'-hydroxyacetanilide Acenol Acetaminofén Acétaminophène APAP N-acetyl-p-aminophenol p-acetamidophenol p-acetaminophenol p-Acetylaminophenol p-hydroxy-acetanilid p-hydroxyacetanilide p-hydroxyphenolacetamide Paracétamol Paracetamol Paracetamolum

Figure 3.7: The resulting section pointing out the synonym names of the drug acetaminophen

3.3 cTAKES

3.3.1 The role of clinical processing tools

Electronic medical records have been shown to be promising resources for clinical and healthcare research. However, the vast amount of unstructured and heterogeneous EMRs makes leveraging useful information become challenging. Researchers may employ many domain experts to extract clinical terms in unstructured clinical notes that are meaningful for characterizing patients. This manual approach, however, often requires much time and effort for handling large EMRs corpus. To avoid this limitation, one can rely on natural language processing tools which have been widely developed in the literature. However, most of the tools are designed for general text and they have not been customized for the biomedical domain. Therefore, annotating clinical terms and their semantic meaning automatically becomes an essential need to exploit EMR resources effectively.

Figure 3.8 illustrates an example of expected outputs using clinical processing tools. They should be able to recognize clinical words or phrases, and map them to concepts of standardized ontologies, for example, CUI in UMLS Metathesaurus.

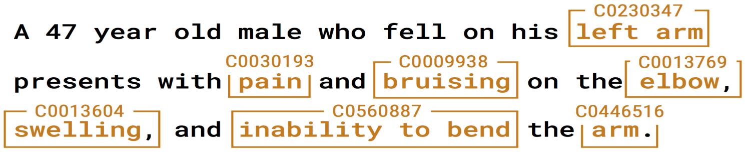


Figure 3.8: An example of expected outputs produced by clinical processing tools

3.3.2 cTAKES overview and components

cTAKES stands for Clinical Text Analysis and Knowledge Extraction System [Savova et al., 2010]. It is a pipelined NLP program specifically developed for clinical narratives. cTAKES components are trained with a portion of clinical notes recorded Mayo Clinic EMRs to provide semantic annotations that are useful for clinical support systems and clinical research. The main components of cTAKES are listed as follows.

Assertion plays a role in examining the implication of annotated words in a given context. For example, occurrences of the instance "diabetes" are mostly thought as patients with diabetes. However, the assertion component allows distinguishing the mention is negated, general context, patient history, or uncertain. More interestingly, it can identify the subject of the instance is the patient himself or someone else.

Chunker is often regarded as a shallow parser that labels phrases to noun phrases, verb phrases, etc. It consists of three main tasks: building a model, tagging text data with a trained model and adjusting the offset of certain chunks.

Context dependent tokenizer uses surrounding information to annotate one or more tokens. For example, the token created by a dash in between two number (e.g 3-4). This component is served for annotations about the date, fraction, measurement, range, roman number, times.

Clinical documents pipeline allows processing a clinical document in a pipeline mode that consists of detecting sentence boundary, tagging part of speech, chunking, recognizing name entities, detecting context, detecting negation. It takes input files as either plain text or CDA format (clinical document architecture format).

Drug named entity recognition aims to annotate drug name mentions and related attributes such as dosage, route. It receives input in either plain text or CDA format.

Named Entity Recognition provides dictionary mapping (lookup algorithm) that maps each named entities to one of the following UMLS semantic types: signs/symptoms, diseases/disorders, medications, anatomical sites, procedures.

Core contains several analysis engines that mainly perform sentence segmentation and tokenizer

POS tagger borrows wrappers from UIMA (unstructured information management architecture) in addition to the popular OpenNLP part-of-speech tagger so that it can work well with clinical context.

Lexical variant generator (LVG) covers the NLM Lexical tools. It generates variants of words, e.g capitalization words, plurals forms so that it can be looked up by the dictionary. We note that it is an optional function for dictionary lookup.

Dependency parser and semantic role labeler provides syntactic information about terms and assigns predicate-argument structure of the sentence (who, what, whom and where).

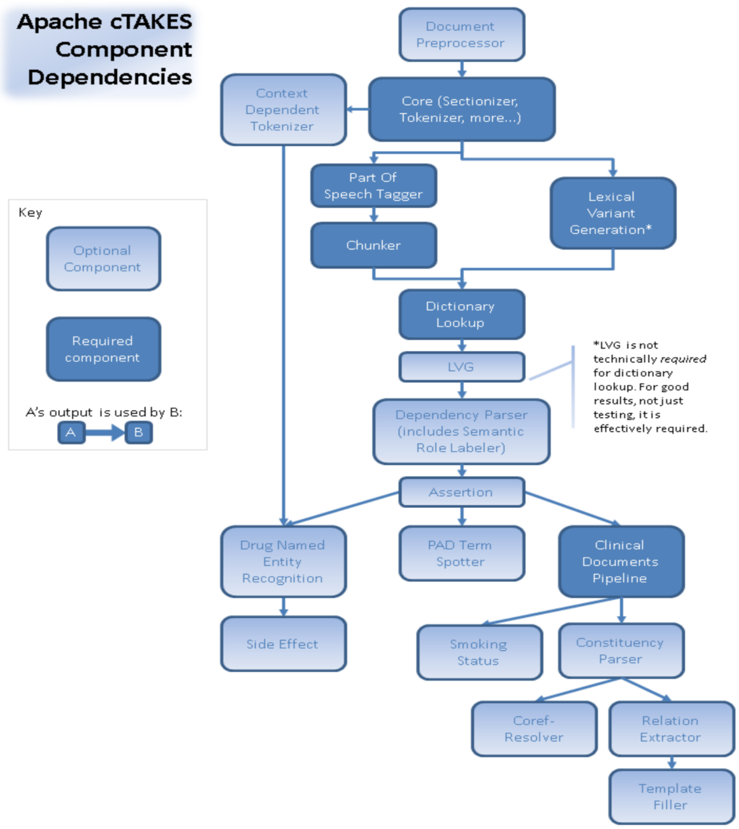
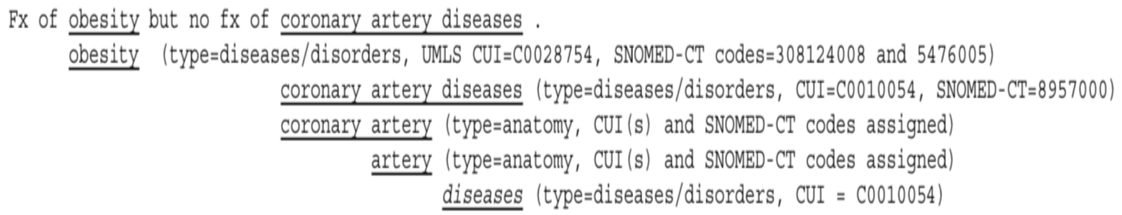

Apache cTAKES Component Dependencies
Document Preprocessor
Context Dependent Tokenizer
Core (Sectionizer, Tokenizer, more...)
Part Of Speech Tagger
Lexical Variant Generation*
Chunker
Dictionary Lookup
LVG
*LVG is not technically required for dictionary lookup. For good results, not just testing, it is effectively required.
Dependency Parser (includes Semantic Role Labeler)
Assertion
Drug Named Entity Recognition
PAD Term Spotter
Clinical Documents Pipeline
Side Effect
Smoking Status
Constituency Parser
Coref-Resolver
Relation Extractor
Template Filler
Key
Optional Component
Required component
A's output is used by B:
A
B

Figure 3.9: Component dependencies in the cTAKES architecture (source: cTAKES documentation)

Figure 3.9 shows the component dependencies in cTAKES architecture. The filled blue boxes indicate the required components while the transparent ones are optional components. Each component could take outputs of one or



```
Fx of obesity but no fx of coronary artery diseases .
      obesity  (type=diseases/disorders, UMLS CUI=C0028754, SNOMED-CT codes=308124008 and 5476005)
                     coronary artery diseases (type=diseases/disorders, CUI=C0010054, SNOMED-CT=8957000)
                     coronary artery (type=anatomy, CUI(s) and SNOMED-CT codes assigned)
                              artery (type=anatomy, CUI(s) and SNOMED-CT codes assigned)
                                     diseases (type=diseases/disorders, CUI = C0010054)
```

Figure 3.10: An example outputs of named entities recognition and concept mapping resulted by cTAKES

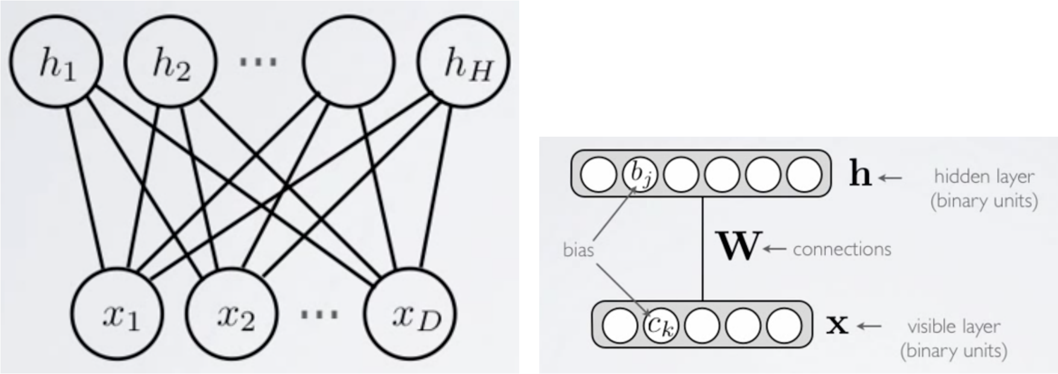
several other components directly linked to it. Figure 3.10 gives an example of how the named entities are recognized by using cTAKES.

3.4 Restricted Boltzmann Machine

3.4.1 Model

The restricted Boltzmann machine [Fischer and Igel, 2012] is a non-directed graphical model that defines the distribution over some input vector $\mathbf{x}$ and models the distribution of input vectors in training data via a binary hidden unit layer $\mathbf{h}$. It consists of D visible units representing the input vector, H hidden units and the interactions between pairs of visible and hidden units. Each unit takes a binary value 0 or 1. We denote $\theta = (\mathbf{W}, \mathbf{b}, \mathbf{c})$ as the model parameters where $\mathbf{W} = (w_{jk})$ consisting of the interactaction weights of hidden node j and visible node k, the bias of visible units $\mathbf{c} = (c_k)$ and the bias of hidden units $\mathbf{b} = (b_j)$. Figure 3.11 illustrates the graphical representation of the RBM model and its parameters.

The RBM model defines an energy-based function that characterizes the joint distribution of hidden and visible units. The formula of the energy function is given as follows.



(a)

(b)

Figure 3.11: The graphical representation of the RBM model (a). The full representation where the bottom nodes are input units and the top nodes are hidden units (b). Model parameters including bias terms c_k and b_j of each visible unit and hidden unit, respectively, and the weighting matrix $\mathbf{W}$

$$\begin{aligned} E(\mathbf{x}, \mathbf{h}) &= -\mathbf{h}^\top \mathbf{W} \mathbf{x} - \mathbf{c}^\top \mathbf{x} - \mathbf{b}^\top \mathbf{h} \\ &= -\sum_j \sum_k W_{j,k} h_j x_k - \sum_k c_k x_k - \sum_j b_j h_j \end{aligned}$$

It can be seen that the above function is linear in terms of either $\mathbf{h}$ or $\mathbf{x}$. It involves the hidden unit vector, the visible unit vector, the weighting matrix and the bias vectors of the visible units and hidden units. The fully joint distribution of hidden and visible units is defined as below.

$$p(\mathbf{x}, \mathbf{h}) = \exp(-E(\mathbf{x}, \mathbf{h}))/Z$$

where Z is a normalization constant:

$$Z = \sum_{\mathbf{x}} \sum_{\mathbf{h}} \exp(-E(\mathbf{x}, \mathbf{h}))$$

The joint distribution $p(\mathbf{x}, \mathbf{h})$ can be expanded as follows.

$$\begin{aligned}
p(\mathbf{x}, \mathbf{h}) &= \exp(-E(\mathbf{x}, \mathbf{h}))/Z \\
&= \exp(\mathbf{h}^\top \mathbf{W}\mathbf{x} + \mathbf{c}^\top \mathbf{x} + \mathbf{b}^\top \mathbf{h})/Z \\
&= \exp(\mathbf{h}^\top \mathbf{W}\mathbf{x}) \exp(\mathbf{c}^\top \mathbf{x}) \exp(\mathbf{b}^\top \mathbf{h})/Z \\
&= \frac{1}{Z} \prod_j \prod_k \exp(W_{j,k} h_j x_k) \prod_k \exp(c_k x_k) \prod_j (b_j h_j)
\end{aligned}$$

The bipartite structure with no intra-layer connections of the RBM model makes the hidden units mutually independent given the visible units and vice verse. This property allows factorizing the following conditional probabilities.

$$\begin{aligned}
p(\mathbf{h}|\mathbf{x}) &= \prod_j p(h_j|\mathbf{x}) \\
\text{where } p(h_j = 1|\mathbf{x}) &= \frac{1}{1 + \exp(-(b_j + \mathbf{W}_{j\cdot}\mathbf{x}))} = sigm(b_j + \mathbf{W}_{j\cdot}\mathbf{x}) \\
\text{and } p(\mathbf{x}|\mathbf{h}) &= \prod_k p(x_k|\mathbf{h}) \\
\text{where } p(x_k = 1|\mathbf{h}) &= \frac{1}{1 + \exp(-(c_k + \mathbf{h}^\top \mathbf{W}_{\cdot k}))} = sigm(c_k + \mathbf{h}^\top \mathbf{W}_{\cdot k})
\end{aligned}$$

The following derivation gives a detailed proof of the factorizable property of the conditional probability of the hidden unit vector given the visible unit vector. The remaining conditional probability can be proved similarly.

$$
\begin{aligned}
p(\mathbf{h}|\mathbf{x}) &= p(\mathbf{x},\mathbf{h}) / \sum_{\mathbf{h}'} p(\mathbf{x},\mathbf{h}') \\
&= \frac{\exp(\mathbf{h}^\top \mathbf{W}\mathbf{x} + \mathbf{c}^\top \mathbf{x} + \mathbf{b}^\top \mathbf{h})/Z}{\sum_{\mathbf{h}' \in \{0,1\}^H} \exp(\mathbf{h}'^\top \mathbf{W}\mathbf{x} + \mathbf{c}^\top \mathbf{x} + \mathbf{b}^\top \mathbf{h}')/Z} \\
&= \frac{\exp(\sum_j h_j \mathbf{W}_{j\cdot}\mathbf{x} + b_j h_j)}{\sum_{h_1' \in \{0,1\}} \cdots \sum_{h_H' \in \{0,1\}} \exp(\sum_j h_j' \mathbf{W}_{j\cdot}\mathbf{x} + b_j h_j')} \\
&= \frac{\prod_j \exp(h_j \mathbf{W}_{j\cdot}\mathbf{x} + b_j h_j)}{\sum_{h_1' \in \{0,1\}} \cdots \sum_{h_H' \in \{0,1\}} \prod_j \exp(h_j' \mathbf{W}_{j\cdot}\mathbf{x} + b_j h_j')} \\
&= \frac{\prod_j \exp(h_j \mathbf{W}_{j\cdot}\mathbf{x} + b_j h_j)}{\left(\sum_{h_1' \in \{0,1\}} \exp(h_1' \mathbf{W}_{1\cdot}\mathbf{x} + b_1 h_1')\right) \cdots \left(\sum_{h_H' \in \{0,1\}} \exp(h_H' \mathbf{W}_{H\cdot}\mathbf{x} + b_H h_H')\right)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W}_{j\cdot}\mathbf{x} + b_j h_j)}{\prod_j \left(\sum_{h_j' \in \{0,1\}} \exp(h_j' \mathbf{W}_{j\cdot}\mathbf{x} + b_j h_j')\right)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W}_{j\cdot}\mathbf{x} + b_j h_j)}{\prod_j (1 + \exp(b_j + \mathbf{W}_{j\cdot}\mathbf{x})} \\
&= \prod_j \frac{\exp(h_j \mathbf{W}_{j\cdot}\mathbf{x} + b_j h_j)}{1 + \exp(b_j + \mathbf{W}_{j\cdot}\mathbf{x})} \\
&= \prod_j p(h_j|\mathbf{x})
\end{aligned}
$$

It is noted that to get the last equation, the following derivation is applied.

$$
\begin{aligned}
p(h_j = 1|\mathbf{x}) &= \frac{\exp(b_j + \mathbf{W}_{j\cdot}\mathbf{x})}{1 + \exp(b_j + \mathbf{W}_{j\cdot}\mathbf{x})} \\
&= \frac{1}{1 + \exp(-b_j - \mathbf{W}_{j\cdot}\mathbf{x})} \\
&= sigm(b_j + \mathbf{W}_{j\cdot}\mathbf{x})
\end{aligned}
$$

The marginal distribution $p(\mathbf{x})$ can be expanded as follows.

$$
\begin{aligned}
p(\mathbf{x}) &= \sum_{\mathbf{h}\in\{0,1\}}^{H} p(\mathbf{x},\mathbf{h}) \\
&= \sum_{\mathbf{h}\in\{0,1\}}^{H} \exp(-E(\mathbf{x},\mathbf{h}))/Z \\
&= \sum_{\mathbf{h}\in\{0,1\}^{H}} \exp(\mathbf{h}^{\top}\mathbf{W}\mathbf{x} + \mathbf{c}^{\top}\mathbf{x} + \mathbf{b}^{\top}\mathbf{h})/Z \\
&= \exp(\mathbf{c}^{\top}\mathbf{x}) \sum_{h_1\in\{0,1\}} \cdots \sum_{h_H\in\{0,1\}} \exp\Big(\sum_j h_j \mathbf{W}_{j}.\mathbf{x} + b_j h_j\Big)/Z \\
&= \exp(\mathbf{c}^{\top}\mathbf{x})\Big(\sum_{h_1\in\{0,1\}} \exp(h_1 \mathbf{W}_{\mathbf{1}}.\mathbf{x} + b_1 h_1)\Big) \cdots \Big(\sum_{h_H\in\{0,1\}} \exp(h_H \mathbf{W}_{\mathbf{H}}.\mathbf{x} + b_H h_H)\Big)/Z \\
&= \exp(\mathbf{c}^{\top}\mathbf{x})(1+\exp(b_1 + \mathbf{W}_{1}.\mathbf{x})) \cdots (1+\exp(b_H + \mathbf{W}_{H}.\mathbf{x}))/Z \\
&= \exp(\mathbf{c}^{\top}\mathbf{x}) \exp(\log((1+\exp(b_1 + \mathbf{W}_{1}.\mathbf{x})))) \cdots \exp(\log((1+\exp(b_H + \mathbf{W}_{H}.\mathbf{x}))))/Z \\
&= \exp\Big(\mathbf{c}^{\top}\mathbf{x} + \sum_{j=1}^{H} \log(1+\exp(b_j + \mathbf{W}_{j}.\mathbf{x}))\Big)/Z
\end{aligned}
$$

In the second last equation, the marginal distribution can be considered as the product of experts model. It has been proven that the RBM can approximate any distribution over $\{0,1\}^D$ arbitrarily well (in the sense of the KL divergence) with $k+1$ hidden units where k is the number of input vectors whose probability is not 0 [Le Roux and Bengio, 2008]. This property shows the powerful representation of the RBM model.

It can be understood as the number of dimensions whose are not zero vectors in the input matrix. I keep the same words like in the reference paper

3.4.2 Training and inference

To train the RBM, we minimize the average negative log-likelihood (NLL)

$$\frac{1}{T}\sum_t l(f(\mathbf{x}^{(t)})) = \frac{1}{T}\sum_t -\log(p(\mathbf{x}^{(t)}))$$

Taking the partial derivation of $\log(p(\mathbf{x}^{(t)})$, one can show that:

$$\frac{\partial(-\log(p(\mathbf{x}^{(t)}))}{\partial\theta} = \mathbb{E}_{\mathbf{h}}\left[\frac{\partial E(\mathbf{x}^{(t)},\mathbf{h})}{\partial\theta}\middle|\mathbf{x}^{(t)}\right] - \mathbb{E}_{\mathbf{x},\mathbf{h}}\left[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial\theta}\right] \tag{3.1}$$

In the above equation, computing the first expectation is named as a positive phase and computing the second expectation is named as a negative phase. Since the second expectation takes exponential number configuration of pairs $(\mathbf{h},\mathbf{v})$, computing it is intractable in practice. Instead, we can apply the contrastive divergence [Hinton, 2012] to replace the expectation by a point estimate at $\tilde{\mathbf{x}}$. We repeat the following Gibbs sampling procedure k times to sample a negative $\tilde{\mathbf{x}}$ as illustrated in Figure 3.12.

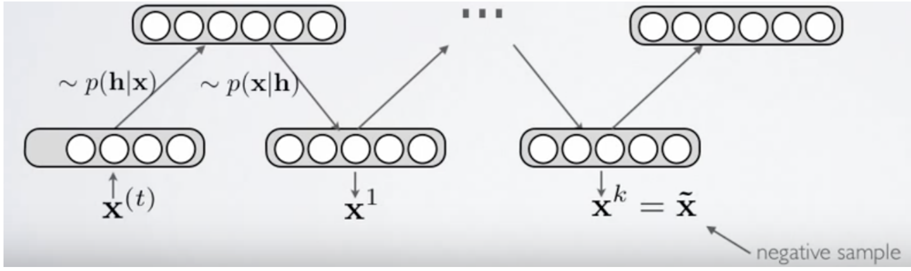

$\sim p(\mathbf{h}|\mathbf{x})$
$\sim p(\mathbf{x}|\mathbf{h})$
...
$\mathbf{x}^{(t)}$
$\mathbf{x}^{1}$
$\mathbf{x}^{k} = \tilde{\mathbf{x}}$
negative sample

Figure 3.12: Gibbs sampling procedure of $\tilde{\mathbf{x}}$

$$\begin{aligned}
&\mathbf{x}^{\mathbf{0}} = \mathbf{x}^{(t)}\\
&\mathbf{h}^{\mathbf{k}} \sim p(\mathbf{h}|\mathbf{x} = \mathbf{x}^{\mathbf{k}}) \text{ for } k \geq 0\\
&\mathbf{x}^{\mathbf{k}} \sim p(\mathbf{x}|\mathbf{h} = \mathbf{h}^{\mathbf{k-1}}) \text{ for } k \geq 1
\end{aligned}$$

With $\tilde{\mathbf{x}}$, the expectation terms in the positive phase and negative phase are approximated as follows.

$$\begin{aligned}
\mathbb{E}_{\mathbf{h}}\left[\frac{\partial E(\mathbf{x}^{(t)},\mathbf{h})}{\partial\theta}\middle|\mathbf{x}^{(t)}\right] &= \frac{\partial E(\mathbf{x}^{(t)},\tilde{\mathbf{h}}^{(t)})}{\partial\theta}\\
\mathbb{E}_{\mathbf{x},\mathbf{h}}\left[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial\theta}\right] &= \frac{\partial E(\tilde{\mathbf{x}},\tilde{\mathbf{h}})}{\partial\theta}
\end{aligned}$$

The derivation of $\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial \theta}$ for $\theta = W_{jk}$ can be derived as below.

$$\begin{aligned}\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial W_{jk}} &= \frac{\partial}{\partial W_{jk}}\Big(-\sum_{jk} W_{jk}h_j x_j - \sum_k c_k x_k - \sum_j b_j h_j \Big) \\ &= -\frac{\partial}{\partial W_{jk}} \sum_{jk} W_{jk} h_j x_k \\ &= -h_j x_k \end{aligned}$$

Putting the above obtained derivation in matrix form, we get the gradient of the energy function with respect to the full matrix $\mathbf{W}$ as following vectorization form.

$$\nabla_{\mathbf{W}} E(\mathbf{x},\mathbf{h}) = -\mathbf{h}\mathbf{x}^\top$$

The expectation term $\mathbb{E}_{\mathbf{h}}\left[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial \theta}\Big|\mathbf{x}\right]$ for $\theta = W_{jk}$ then can be derived as below.

$$\mathbb{E}_{\mathbf{h}}\left[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial \theta}\Big|\mathbf{x}\right] = \mathbb{E}_{\mathbf{h}}\left[-h_j x_k \Big| \mathbf{x}\right] = \sum_{h_j \in \{0,1\}} -h_j x_k p(h_j|\mathbf{x}) = -x_k p(h_j = 1|\mathbf{x})$$

Be definition, it is noted that

$$\mathbf{h}(\mathbf{x}) \overset{def}{=} \begin{pmatrix} p(h_1 = 1|\mathbf{x}) \\ \cdots\cdots\cdots\cdots \\ p(h_H = 1|\mathbf{x}) \end{pmatrix} = sigm(\mathbf{b} + \mathbf{W}\mathbf{x})$$

Therefore, we get

$$\mathbb{E}_{\mathbf{h}}[\nabla_{\mathbf{W}} E(\mathbf{x},\mathbf{h})|\mathbf{x}] = -\mathbf{h}(\mathbf{x})\mathbf{x}^\top$$

Given $\mathbf{x}^{(t)}$ and $\tilde{\mathbf{x}}$, the learning rule for $\theta = \mathbf{W}$ becomes

$$\begin{aligned}
\mathbf{W} \Longleftarrow & \mathbf{W} - \alpha\Big(\nabla_{\mathbf{W}}\big(-\log p(\mathbf{x}^{(t)})\big)\Big)\\
\Longleftarrow & \mathbf{W} - \alpha\Big(\mathbb{E}_{\mathbf{h}}\left[\nabla_{\mathbf{W}} E(\mathbf{x}^{(t)},\mathbf{h})|\mathbf{x}^{(t)}\right] - \mathbb{E}_{\mathbf{x},\mathbf{h}}[\nabla_{\mathbf{W}} E(\mathbf{x},\mathbf{h})]\Big)\\
\Longleftarrow & \mathbf{W} - \alpha\Big(\mathbb{E}_{\mathbf{h}}\left[\nabla_{\mathbf{W}} E(\mathbf{x}^{(t)},\mathbf{h})|\mathbf{x}^{(t)}\right] - \mathbb{E}_{\mathbf{h}}[\nabla_{\mathbf{W}} E(\tilde{\mathbf{x}},\mathbf{h})|\tilde{\mathbf{x}}]\Big)\\
\Longleftarrow & \mathbf{W} + \alpha\Big(\mathbf{h}(\mathbf{x}^{(t)})\mathbf{x}^{(t)^\top} - \mathbf{h}(\tilde{\mathbf{x}})\tilde{\mathbf{x}}^\top\Big)
\end{aligned}$$

Algorithm 1 summarizes the pseudo-code for learning the parameters of the RBM model. It is noted that the bigger k, the less biased the estimate of the gradient. In practice, using $k = 1$ works well for generating negative samples.

Algorithm 1: Constrastive divergence algorithm for learning RBM parameters

```
1 for each training example $\mathbf{x}^{(t)}$ do
2     Generate a negative sample $\tilde{x}$ using k-steps of Gibbs sampling, starting at $x^{(t)}$ ;
3     Update parameters
          $\mathbf{W} \Longleftarrow \mathbf{W} + \alpha\Big(\mathbf{h}(\mathbf{x}^{(t)})\mathbf{x}^{(t)^\top} - \mathbf{h}(\tilde{\mathbf{x}})\tilde{\mathbf{x}}^\top\Big)$
          $\mathbf{b} \Longleftarrow \mathbf{b} + \alpha\Big(\mathbf{h}(\mathbf{x}^{(t)}) - \mathbf{h}(\tilde{\mathbf{x}})\Big)$
          $\mathbf{c} \Longleftarrow \mathbf{c} + \alpha\Big(\mathbf{x}^{(t)} - \tilde{\mathbf{x}}\Big)$
4 Go back to 1. until stopping criteria
```

Chapter 4

Treatment Learning Method

4.1 Introduction

Treatment is one of the most crucial procedures that significantly affect patient health status. Understanding treatment behavior has become a growing requirement for different purposes such as designing treatment plans, optimizing treatments, supporting the clinical decision, detecting the medical deviation and so on. More importantly, grasping treatment patterns often used in healthcare organizations could help inexperienced doctors to sharpen their knowledge about disease curing and avoid making prescriptions that can cause adverse drug reaction.

Treatment for a particular disease can be learned from personal experience or clinical guidelines. However, this approach requires much time and effort to absorb various treatment protocols. In addition, patient conditions are diverse in terms of demographics and symptoms. As a result, detailed treatments for any patient might not be found easily in the literature.

In recent years, along with the increasing use of electronic medical records, learning treatment patterns has become an attractive task to data scientists. The two most common data-driven approaches, which have been employed so far are probabilistic models and association analysis. However, most of the current research works have used a limited subset of patient features due to the data collection issue and the challenge of representing complex EMRs objects. Moreover, while the resulting treatment patterns may be affected significantly by prescription drugs given in a treatment period, identifying hidden treatment periods in longitudinal prescription records are still underestimated with fixed treatment period intervals.

In addition to the underestimate in identifying treatment periods, current research studies are also limited in discovering different kinds of treatment patterns. While the treatment learning task generally aims to discover frequent patterns or notable patterns in treating groups of patients who suffer similar symptoms, most of the related works have focused on deriving a few treatment patterns from all patients within a patient group or a few typical patients in that group. An interpretable approach is to employ association mining techniques to discover the most frequent treatment patterns. Given a patient group, n treatment patterns of that group are usually represented as follows.

Treatment pattern 1: $\{drug_{11}, drug_{12}, ...\}$, support: $s_1\%$

Treatment pattern 2: $\{drug_{21}, drug_{22}, ...\}$, support: $s_2\%$

...

Treatment pattern n: $\{drug_{n1}, drug_{n2}, ...\}$, support: $s_n\%$

This representation suffers from several major drawbacks. First, it is hard to figure out different kinds of drug combinations. It is well known that even under similar symptoms, patients are often treated in diverse ways. The traditional representation merely reflects the co-occurrence of a set of frequent prescription drugs. However, in addition to set of frequently prescribed drugs, physicians may also care about other kinds of pattern drugs, for example whether a drug is prescribed in conjunction with some other drugs. Such simple but essential inquiry could not be easily answered via the flat representation of treatment patterns. Second, association analysis approach typically requires users to set a threshold that is sensitive. A low threshold could result in an explosion of numbers of treatment patterns that increase the computational complexity. On the other hand, a high threshold could make the solution miss many rare patterns.

In this chapter, we propose a method that addresses the above drawbacks. To tackle the challenge of representing mixed type electronic medical records, our method employs a MV.RBM model [Tran et al., 2014] to transfer heterogeneous objects into homogeneous ones. For the second drawback, we propose an indication-based algorithm which automatically splits treatment periods into milestones. To capture more different kinds of treatment patterns in a patient group, we present prescription drugs' frequency along with a tree's nodes so that it can reflect as much as possible the association among prescription drugs. The following section describes the proposed method in detail.

4.2 Method

Our methodology for the treatment learning problem considers a set of training patients who have the same diagnostic codes. It captures the intuition that patients who share latent features underlying their health condition and profiles are likely to belong to the same subcohort. Under the above assumption, patients in each subcohort are supposed to be treated by similar care plans. Figure 4.1 describes an overview of the proposed treatment learning method. It consists of two major tasks: clustering patients into subcohorts and learning typical treatment patterns of each subcohort. We present all relevant steps of the treatment learning method in the following subsections.

4.2.1 Data preprocessing

In this work, training data can be divided into two different sets. One consists of the components $\{Lab,\ Info,\ Note\}$ named non-treatment-based or symptom-based records to serve for the subcohort construction and the other consists of the component $Presc$ named treatment-based records to serve for the treatment pattern identification of each subcohort. We note that non-treatment-based data contains both static ($Info$) and longitudinal data ($\{Lab,\ Note\}$). As the ultimate goal of our work is to recommend prescription drugs for new patients from the beginning dates after admitted to hospitals, only initial values, i.e. the data recorded at timestamp tl_1, tn_1, of time-dependent non-treatment-based data are considered. This data extrac-

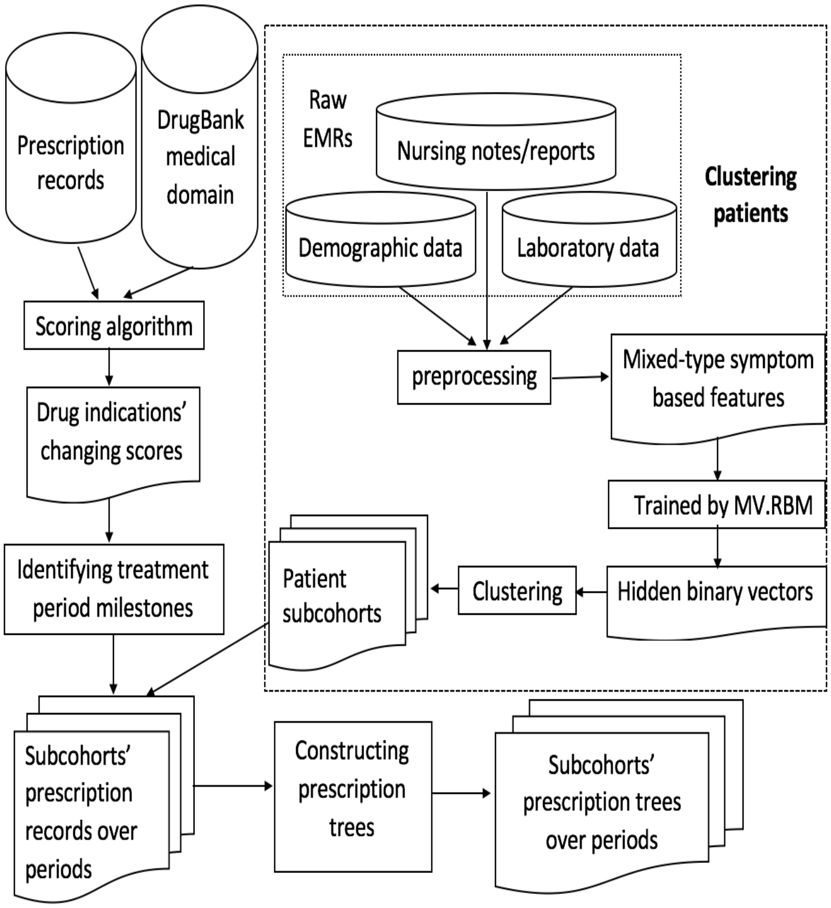

Prescription records
DrugBank medical domain
Scoring algorithm
Drug indications' changing scores
Identifying treatment period milestones
Raw EMRs
Nursing notes/reports
Demographic data
Laboratory data
Clustering patients
preprocessing
Mixed-type symptom based features
Trained by MV.RBM
Hidden binary vectors
Clustering
Patient subcohorts
Subcohorts' prescription records over periods
Constructing prescription trees
Subcohorts' prescription trees over periods

Figure 4.1: An overview of the treatment learning method

tion approach is based on the assumption that patients with similar initial signs, symptoms, or laboratory indicators can be treated similarly.

Extracted training data is normalized in different ways. Categorical variables are represented as one-hot vectors while numerical ones are scaled to zero mean unit variance. It is noted that to facilitate the task of clinical processing, we employ cTAKES to extract initial features from nursing notes. As described in the background chapter, cTAKES is a clinical processing toolkit that allows not only recognizing clinical concepts effectively but also determining their linked semantic types. In our method, we focus on exploiting meaningful clinical features to represent patients. For this reason, the UMLS clinical concepts describing signs/symptoms or diseases are extracted.

4.2.2 Data representation and patient clustering

Data mining methods are typically designed for homogeneous input. Unfortunately, the preprocessed patient records in the previous step exist in different data types. Therefore, it is necessary to transform such heterogeneous data to homogeneous space so that input data is ready-to-use for clustering methods.

MV.RBM, an extended version of RBM, is a robust representation model which can fulfill the above objective. Input units of MV.RBM allow to feed not only binary values but also other data types such as numerical or categorical ones. Let $\mathbf{v} = (v_1, v_2, .., v_N)$ be the set of N visible units and $\mathbf{h} = (h_1, h_2, .., h_K)$ be the set of K hidden units. The energy function of MV.RBM is added extra terms which are dedicated to different data types. Its formula is provided below.

$$E(\mathbf{v}, \mathbf{h}) = -(\sum_i G_i(v_i) + \sum_k b_k h_k + \sum_{ik} H_{ik}(v_i) h_k)$$

where $\mathbf{b} = (b_1, b_2, .., b_N)$ is the bias values of hidden units, $G_i(v_i)$ and $H_{ik}(v_i)$ are specific-type functions. Due to the conditional independence of nodes within each layer, one can factorize the joint distribution of visible units given hidden units as follows.

$$P(\mathbf{v}|\mathbf{h}) = \prod_{i=1}^{N} P_i(v_i|\mathbf{h});\ P(\mathbf{h}|\mathbf{v}) = \prod_{k=1}^{K} P(h_k|\mathbf{v})$$

$$P_i(v_i|\mathbf{h}) = \frac{1}{Z(\mathbf{h})} \exp(G_i(v_i) + \sum_k H_{ik}(v_i) h_k)$$

$$P(h_k^1|\mathbf{v}) = \frac{1}{1 + \exp(-w_k - \sum_i H_{ik}(v_i))}$$

where h_k^1 indicates the assignment $h_k = 1$, and $Z(\mathbf{h})$ is a normalization constant. The functions $G_i(v_i)$, $H_{ik}(v_i)$ and corresponding $P_i(v_i|\mathbf{h})$ for each data type are described in Table 4.1. Detailed steps for learning model parameters can be referred in [Tran et al., 2014].

	$G_i(v_i)$	$H_{ik}(v_i)$	$P_i(v_i \mid \mathbf{h})$
Binary	$a_i v_i$	$w_{ik} v_i$	$\frac{\exp(a_i v_i + \sum_k w_{ik} h_k v_i)}{1+\exp(a_i + \sum_k w_{ik} h_k)}$
Gaussian	$-v_i^2/2\sigma^2 + a_i v_i$	$w_{ik} v_i$	$\mathcal{N}(\sigma_i^2(a_i + \sum_k w_{ik} h_k),\, \sigma_i)$
Categorical	$\sum_m a_{im} \delta_m[v_i]$	$\sum_{m,k} a_{imk} \delta_m[v_i]$	$\frac{\exp(\sum_m a_{im} \delta_m[v_i]) + \sum_{m,k} w_{imk} \delta_m[v_i] h_k)}{\sum_l \exp(a_{il} + \sum_k w_{ilk} h_k)}$

Table 4.1: The type specific functions [Tran et al., 2014]. a_i, a_{im} are input bias parameters, w_{ik}, w_{imk} are input-hidden weighting parameters. Those with extra subscript m are dedicated for categorical features

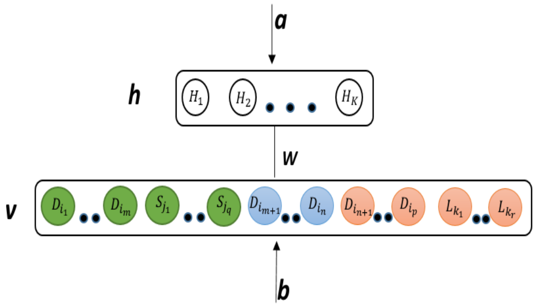


Figure 4.2: A MV.RBM for patient records. The green, blue and orange circles represent for binary, categorical and continuous input units. The circles with labels D, S, L indicate demographic, sign/symptom and laboratory indicator features, respectively

To feed MV.RBM, input features are supposed to be mutually independent conditioned on the hidden unit values. Figure 4.2 illustrates how non-treatment features are fed. For illustration purpose, we assume demographics consist of all kinds of features. Indicator ones are supposed to take numerical values while sign/symptom features extracted form clinical notes are represented as one-hot vectors. After training the MV.RBM, we utilize the hidden states as transformed features of encoded input vectors. The latent representation vectors then can be taken as input values of clustering algorithms for binary data. In our work, we select hierarchical clustering method and decide to use the complete linkage which is claimed to work well with symmetric distance, e.g. Hamming distance, according to a survey by [Tamasauskas et al., 2012].

4.2.3 Drugs' normalization and indication labeling

This section describes our approach for preprocessing treatment-based data. It is a common fact that although there are many prescription drugs sharing similar ingredient, they are often prescribed under different brand names. Therefore, tt is crucial to normalize prescription drugs before doing further steps. We again tackle this issue with the help of cTAKES. For every prescription drug, we identify its candidate UMLS clinical terms and select the one of which semantic type is about medication. In case a few medications are suggested for the same prescription drug, we revise the mapping with the help of domain experts to filter the most appropriate UMLS term linked with the considering medicine.

Besides the necessity of performing drug name normalization, categorizing prescription drugs into indication groups is important to facilitate further tasks of this thesis. In the next chapters, we will show that drugs with attached indication information are helpful to measure the change of indication in prescription records and to interpret derived patterns of subcohorts. To label drug indication, we identify which diseases or symptoms are healed by each normalized drug dr. Following advice from a domain expert, we divide indication groups into three categories listed as follows.

- **Primary group** consists of drugs that treat one or more diagnostic diseases or highly relevant diseases.
- **Sign/symptom group** consists of drugs that treat primary symptoms characterizing the concerning diseases.
- **Risk factor group** consists of drugs that heal any factors causing the concerning diseases.

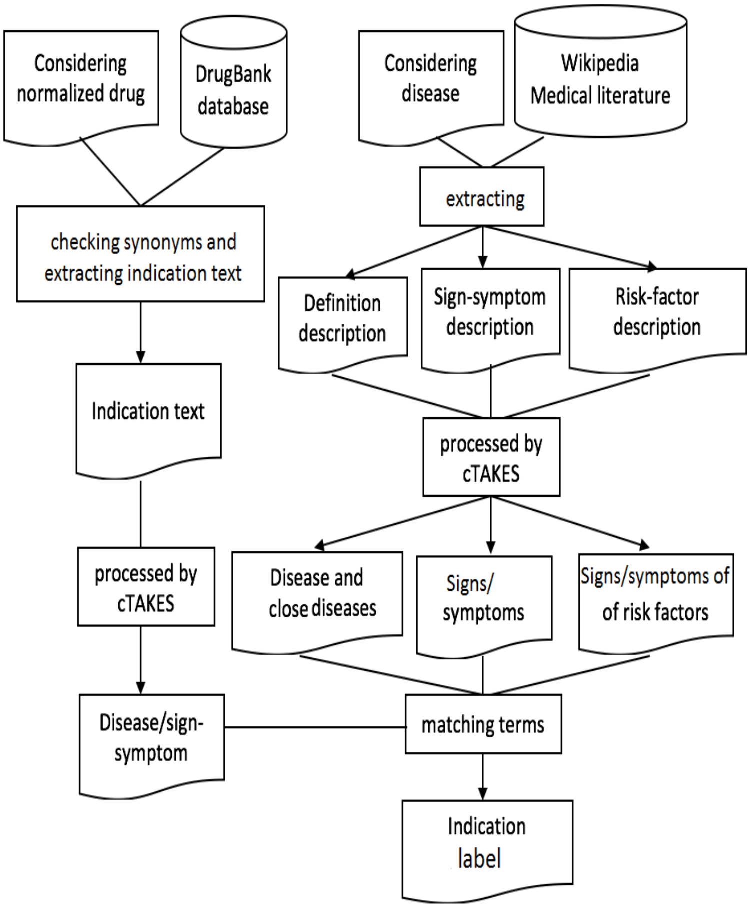


Figure 4.3: An overview of the indication assignment freamework

Figure 4.3 demonstrates our idea to label indication for prescription drugs. Pieces of text describing the definition, typical symptoms and risk factors of considering diseases are crawled from Wikipedia or domain sources. Symptom/disease terminologies from those pieces of text are then extracted with the help of cTAKES. The obtained outputs are UMLS CUI whose semantic types are signs/symptoms or diseases. For normalized drugs, we look for their indication description given in the DrugBank database. We carefully consider all synonym drugs to ensure that the indication of as many normalized drug can be found in the database.

The labeling mechanism works as follows. Given a normalized drug, we process its associated indication text by cTAKES and extract UMLS terms of which semantic is about diseases or symptoms. We label prescription drugs' indication to the primary group, sign/symptom group and risk factor group in that priority if its indication text and the text section describing the definition, the symptom, the risk factor of the considering diseases share any UMLS terms about diseases or symptoms.

We note that our indication labeling mechanism also allows identifying in detail the diseases or symptoms treated by every normalized drug. This feature is useful in interpreting the treatment patterns of each subcohort which will be presented in the later part of this dissertation.

4.2.4 Treatment period identification

Treatments, especially those for acute diseases, are often made through periods. It is apparent that learned patterns considerably depend on the number of drugs given to a patient group within each period. However, in real-world EMRs, there is no clear boundary between periods in prescription data. An underestimate of the role of treatment period identification therefore could affect the learning process of treatment learning methods. In literature, [Sun et al., 2016, Chen et al., 2018] split varying-length prescription into fixed intervals. Our work suggests a more flexible way to identify treatment periods by incorporating medical domain knowledge. In particular, we take into account prescription drug indication to measure the indication change over time among prescription records of every patient. For every timestamp of treatment, we calculate an accumulated score that takes prescription drugs with are new, recent stopped or redelivered with dosage changed into consideration. Since every normalized drug is attached with indication labels via the indication labeling framework, the frequency of above concerning drugs which belong to each drug group can be counted easily. The accumulated score then sums up these quantities. It is noted that each quantity is weighted by the importance of its associated drug, depending on its drug group in curing considering diseases. In our work we ask for expert advice to assign the values of the weight in decreasing order of the primary group, sign/symptom group and risk factor group.

Algorithm 2: Scoring prescription records

```
Data: $\Theta$, $T$
Result: return $scores$ as a list of accumulated scores
Initialize $U$ as an empty set ;    ▷ set of recently delivered drugs
Initialize $scores$ as an empty list ;
$aScore := 0$ ;    ▷ the accumulated score
for each $t \in T$ do
    $D := \{dr \mid \forall dr \in \Theta \wedge dr.startdate == t\}$ ; ▷ delivered drugs on date d
    $N := \{dr \mid \forall dr \in D \wedge dr.name \notin U.name\}$ ; ▷ newly delivered drugs
    $DC := \{dr \mid \forall dr \in D, \exists dr' \in U$ such that $dr.name == dr'.name \wedge dr.dosage <> dr'.dosage\}$ ;    ▷ dosage changed drugs
    $S := \{dr \mid \forall dr \in U \wedge dr.name \notin D.name \wedge dr.enddate < t\}$ ; ▷ recently stopped using drugs
    for each $dr' \in U$ do
        if $\exists dr'' \in D$ such that $dr'.name == dr''.name$ then
            $dr' := dr''$ ;    ▷ update U with redelivered drugs
    $U := (U \setminus S) \cup N$ ;    ▷ update U with newly delivered drugs
    $CD := N \cup DC \cup S$ ; ▷ considering drugs for calculating scores
    $CPD := CD.name \cap PD$;    ▷ considering primary drugs
    $CSD := CD.name \cap SD$;    ▷ considering sign/symptom drugs
    $CRD := CD.name \cap RD$;    ▷ considering risk factor drugs
    $aScore = aScore + |CPD| \times w_{main} + |CSD| \times w_{symp} + |CRD| \times w_{risk}$ ;
    Add $aScore$ to $scores$
```

The algorithm for treatment period identification is described formally as follows. Given a patient p, let $dr_j^{p,t} =< name, startdate, enddate, dosage >$ characterize every drug dr_j prescribed at specific timestamp t by its normalized drug name, starting date, ending date of usage, and dosage. Let $\Theta^p = \{dr_j^{p,t}.name\}$ be the set of drugs given to the patient, $T^p = \{dr_j^{p,t}.startdate\}$ be the ordered set of prescribed dates, and PD, SD, RD be the sets of primary drugs, sign/symptom drugs and risk factor drugs. Algorithm 2 describes in detail how the accumulated score in prescription drugs' indications for a patient p at the timestamp t is calculated. For simplicity, the superscripts p, t, j are removed.

4.2.5 Prescription tree construction

We have demonstrated our domain-embedded algorithm for the treatment period identification. This section presents a new algorithm to derive treatment patterns for a given patient over a period. It is worth noting that the drugs without indication labels are excluded as input for the construction of prescription trees. We suppose that physicians are sufficient knowledge to decide the detailed dosage, delivered time order or other details of the treatment.

Algorithm 3: Procedure for the construction of a prescription tree

```
Tree($d, \nu, \Gamma, \Omega_{\delta_\nu}, \Lambda$)
if $\Omega_{\delta_\nu}$ is $\emptyset$ or $d == \Upsilon$ then
    return
$k := \arg\max_i \sum_{j=1}^{n} a_{ij}$ ;
$\Gamma[\delta_\nu, k] :=$ "$\nwarrow$";
$\delta_k := \delta_\nu \cup k$ ;
$\Omega_{\delta_k} := \Omega^{k}_{\delta_\nu}$;
$\Omega_{\delta_{\nu+}} := \Omega^{-k}_{\delta_\nu}$;
$\Lambda[\delta_k] := \{j \text{ s.t } \omega^{kj}_{\delta_\nu} = 1\}$ ;
if $|\Lambda[\delta_k]| < \epsilon$ then
    Tree($d, \nu, \Gamma, \Omega_{\delta_k}, \Lambda$);
else
    Tree($d + 1, k, \Gamma, \Omega_{\delta_k}, \Lambda$);
    Tree($d, \nu, \Gamma, \Omega_{\delta_{\nu+}}, \Lambda$);
return ;
```

In the literature, treatment patterns are often discovered as a set of frequent prescription drugs from core patients of each subcohort [Sun et al., 2016]. This approach, however, often requires a minimum support threshold which is subjective and sensitive. Moreover, it is uneasy for physicians to figure out the association level of drugs of which support is not greater than the threshold. In the context of the healthcare domain, this limitation becomes non-negligible since physicians may also need to pay attention to infrequent ones in addition to looking for frequent prescription drugs, to avoid mistakes in making treatment decisions.

To this end, we suggest organizing patterns in a tree where each node represents a prescription drug. For each node in the tree, we query for prescription data of patients who were cured by the drugs on nodes from the root node until the current node. The next unlabeled child node is labeled with the most frequent prescription drug apart from those linked with the parent nodes. Determining the drug label of the next child node follows the same mechanism, but we exclude prescription records of patients who were treated by labeled nodes on the same level. We continue this procedure recursively until the number of patients who were treated with drugs from the root until the considering node is fewer than some threshold. For each node, we also save the ID of the patients who were treated by the set of drugs from the root until that node. It is of interest to note that each patient p is treated by nodes on a unique path named treatment path in the tree. This property is utilized for the treatment recommendation task presented in the subsequent section. We denote the notations for the prescription tree construction algorithm as follows.

- d: the current depth of the constructing prescription tree.
- ν: the constructing node.
- Γ: the constructing prescription tree.
- δ_ν: the treatment path from the root until ν.
- $\delta_{\nu+}$: the treatment path from the root until the next unlabeled child node of ν.

- Ω_{δ_ν}: the current patient-drug interaction matrix corresponding to treatment path δ_ν. Table 4.2 illustrates the initial interaction matrix Ω_{δ_ϕ} of the root node. We suppose there are N patients and Q distinct drugs with indication labels in the considering subcohort. We have:

$$\begin{cases} \omega_{\delta_\phi}^{kj} = 1; \text{ if patient } p_j \text{ was treated with drug } dr_k \\ \omega_{\delta_\phi}^{kj} = 0; \text{ if patient } p_j \text{ was not treated with drug } dr_k \end{cases}$$

	p_1	p_2	...	p_j	...	p_N
dr_1	ω^{11}	ω^{12}	...	ω^{1j}	...	ω^{1N}
dr_2	ω^{21}	ω^{22}	...	ω^{2j}	...	ω^{2N}
...	...	...	...	...	...	...
dr_k	ω^{k1}	ω^{k2}	...	ω^{kj}	...	ω^{kN}
...	...	...	...	...	...	...
dr_Q	ω^{Q1}	ω^{Q2}	...	ω^{Qj}	...	ω^{QN}

Table 4.2: The initial drug-patient interaction matrix

- $\Omega^k = (\omega^{ij})$: the interaction matrix of patients who were treated with drug k, where

$$\begin{cases} i \text{ s.t } i \neq k \\ j \text{ s.t } a_{kj} = 1 \end{cases}$$

- $\Omega^{-k} = (\omega^{ij})$: the interaction matrix of patients who were not treated with drug k, where

$$\begin{cases} i \text{ s.t } i \neq k \\ j \text{ s.t } a_{kj} = 0 \end{cases}$$

- $\Lambda[\delta_\nu]$: the patient IDs of patients who were treated by drugs on δ_ν.

- ϵ: the threshold to stop constructing prescription tree at the considering node.

- Υ: the highest depth of the prescription tree.

  Algorithm 3 demonstrates the detailed algorithm of a prescription tree's construction for a patient subcohort over a specific period.

4.3 Experimental Result

4.3.1 Dataset

Our experimental evaluation was performed on MIMIC III [Johnson et al., 2016], an open EMRs database developed by MIT Lab. It includes patient records of over 58000 admissions of patients with ICU stay recorded continually in Beth Israel Deaconess Medical Center from 2010 to 2012.

Although our method is designed for a set of patients who have the same diagnostic codes, it is uneasy to find a large dataset in MIMIC III since patients are usually diagnosed with non-identical series of ICD-9 codes. Therefore, we consider patients who have the same first diagnostic code as a cohort. Table 4.3 reports top five single admission cohorts in the MIMIC III database.

Primary ICD 9	Name	Number of patients
41401	Coronary atherosclerosis of native coronary artery	3430
0389	Unspecified septicemia	1805
41071	Myocardial infarction	1654
4241	Aortic valve disorders	1122
51881	Acute respiratory failure	945

Table 4.3: Top 5 primary ICD codes among patients with singe admission

We extracted patient records of the second, the third and the fifth cohort for our experimental evaluation as they are groups of patients who primary entered hospitals with acute diseases. The names of three cohorts are shortly named as myocardial infarction cohort, septicemia cohort, and respiratory cohort. We extracted only patients who were prescribed not fewer than three times. Those with no prescription data or nursing note data are not included in our experiments. We used cTAKES to extract initial sign/symptom features. Clinical features which appear less than 5% or greater than 95% in each cohort were excluded. As not all patients have laboratory tests on all indicators, we filled out the indicator features with unavailable values to 0. For demographic data, we only collected the features that probably affect patient health status for example gender, age, martial status. For prescription data, we normalized drug names by selecting the most appropriate UMLS term for each prescription drug. We note that some UMLS medication terms have equivalent terms in the DrugBank database but it is hard to find their indication text if using the UMLS medication terms. For such cases, we used the DrugBank terms instead. It is worthy to note that we only recommend prescription drugs with indication labels as they are highly relevant to the primary diagnosis code.

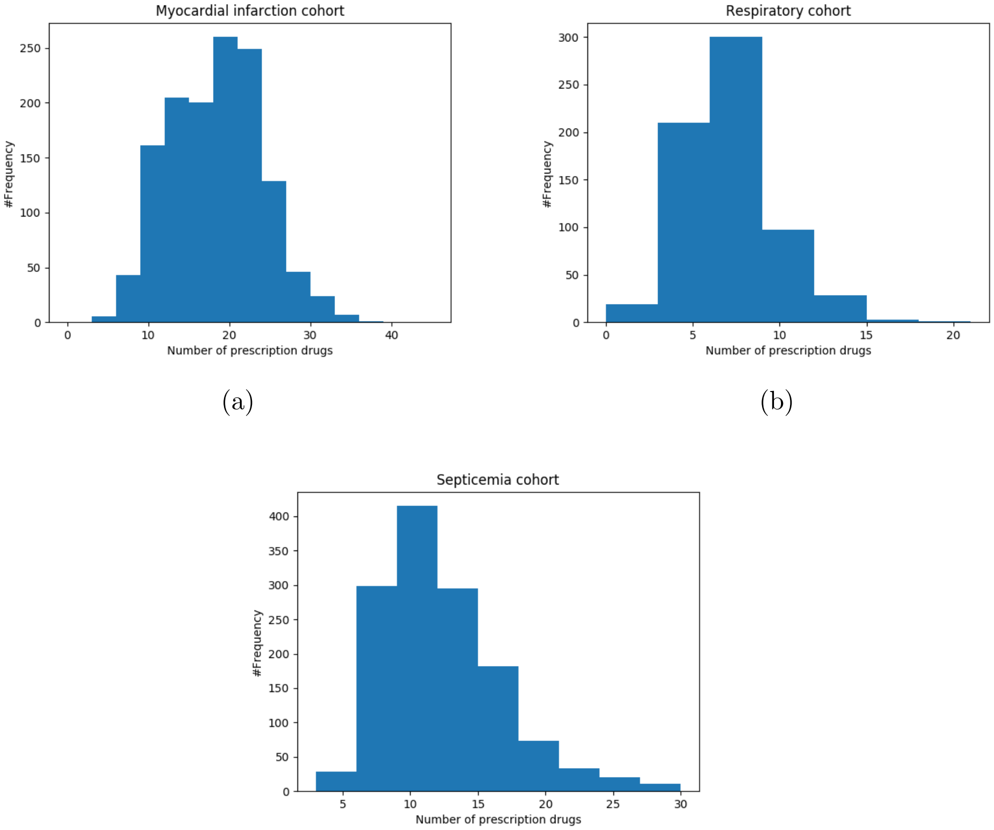
Table 4.4 shows some statistics of the three datasets before and after preprocessing. There are more than 500 prescription drugs delivered to each cohort. Figure 4.4 provides the histogram of the number of prescription drugs with labeled indication delivered to each patient in the three cohorts. Most patients in these cohorts were treated with more than 10 labeled drugs.

	Myocardial infarction	Septicemia	Respiratory
Number of patients	1654	1805	945
Number of processed patients	1330	1359	658
Number of prescription drugs	1038	1238	1047
Number of normalized drugs	558	630	537
Number of drugs with relevant indication labels	244	190	100

Table 4.4: Statistic about datasets used in our experimental evaluation

4.3.2 Parameter setting

This section describes our parameter selection. To train input data with a MV.RBM, we tried several values of the number of hidden units. We set the



(a)

(b)

(c)

Figure 4.4: Histogram of the number of prescription drugs per patient in the three cohorts

number of hidden units in the trained MV.RBM models to 100 since the error rate does not reduce significantly for greater size. Figure 4.5 shows the resulting dendrogram of the myocardial infarction, respiratory and septicemia cohort, respectively. In these cohorts, training patients are grouped more separately when the distance is above 0.6. However, cutting the dendrogram patients at such distance will lead to large size subcohorts. As a result, there is a high chance that treatment in each subcohort may vary significantly. Therefore, we decide the distance parameter is 0.4 so that small-size subcohorts are returned.

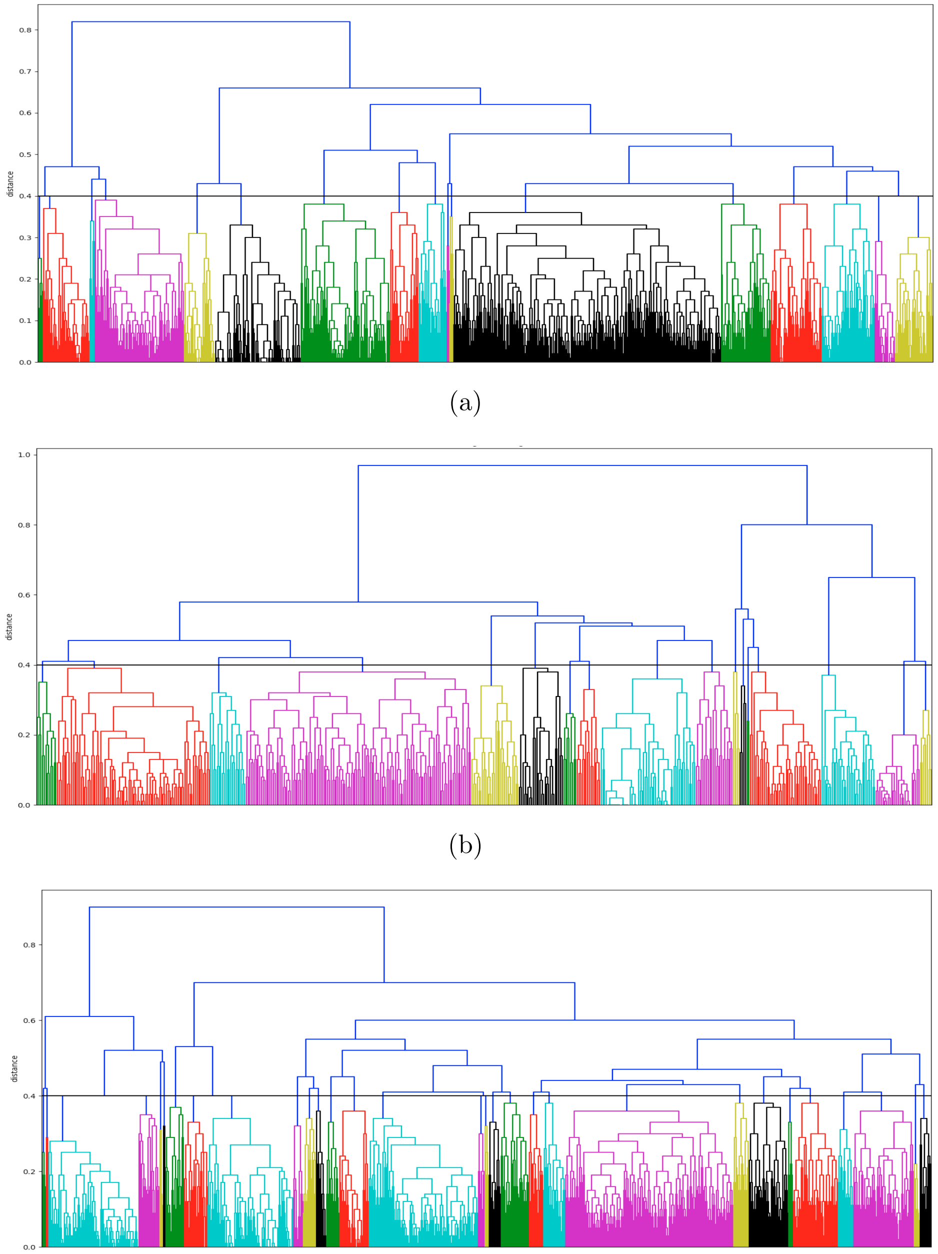
Regarding the depth parameter of the prescription tree, we set it to the number of prescription drugs M to be recommended. For the number of treatment periods n, it is not easy to be confirmed by domain knowledge. Indeed, in our work, it plays a role as to which degree our recommendation method concerns about the chronological order of groups of recommendation drugs. The higher the value of n, the more important to force these groups to preserve the order of delivered time. Figure 4.6 shows an illustration of plotting indication changing score and how we split the varying-length prescription records of a given patient. For this particular work, all prescription records of training patients are split into three periods where the two cutpoints are the timestamps of which the associated accumulated scores change most. For the parameters $w_{main}, w_{symp}, w_{risk}$ in Algorithm 2, we consulted with a physician to assign the values of these parameters to 1, 0.7, 0.5, respectively. The threshold ϵ to stop expanding the prescription tree for each node was set to five patients.

4.3.3 Result and output analysis

In this section, intermediate results from clustering steps and indication labeling framework are presented first, followed by an example of a prescription tree and its interpretation. Lastly, we make a qualitative comparison of the key features of our treatment learning method and the relevant studies.

Table 4.5 shows UMLS terms extracted from the text describing about myocardial infarction. While the term “myocardial infarction” is directly relevant to the illustration cohort, the term “heart attack” is its close concept. Figures 4.7, 4.8, 4.9 present typical sign/symptom or disease terms in three groups characterizing the three cohorts drugs curing theese signs/symptoms or diseases in every group.

Tables 4.6, 4.7, 4.8 provide lists of extracted UMLS sign/disease terms for the three text sections belonging to the three indication groups of septicemia cohort, respiratory cohort, and myocardial infarction cohort, respec-



(a)

(b)

(c)

Figure 4.5: The cluster dendrograms of myocardial infarction cohort (a), respiratory cohort(b), and septicemia cohort (c)

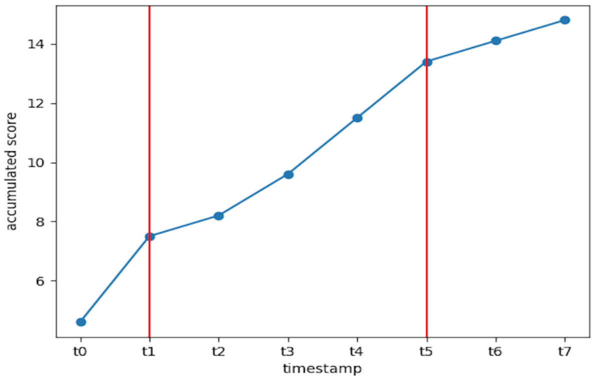


Figure 4.6: A sample plot of accumulated indication changing scores of a given patient. In this example, the patient was prescribed at 8 timestamps $t_0, ..t_7$. The green line is the plot of accumulated score corresponding to each timestamp. We split the longitudinal prescription records at timestamps where significant changes in the accumulated score happen (the red line). For this particular example, we split the prescription records into three periods $(t_0), (t_1, t_2, t_3, t_4), (t_5, t_6, t_7)$

Sample text	Myocardial infarction(MI), commonly known as a heart attack, occurs when blood flow decreases or stops to a part of the heart causing damage to the heart muscle
Extracted terms	Myocardial Infarction, Heart Attack

Table 4.5: Sample extracted UMLS terms by using cTAKES

tively. We note that the extraction of UMLS terms for the definition, the sign/symptom and the risk factor text sections of each cohort requires a little effort to double-check and eliminate a few possible irrelevant terms returned by cTAKES.

Figure 4.10 illustrates a prescription tree of a septicemia subcohort in a specific period. The prefixes “m”,“s”, “r” are used to mark the associated prescription drugs according to which group they belong to. The “m” stands for the primary group, “s” stands for the sign/ symptom group and “r” stands for the risk factor group. For each node, the attached number shows information about the number of patients who were treated by drugs

Primary	bacterial infections; bacteremia; sepsis; communicable diseases; bacterial sepsis
Signs/symptoms	oliguria; hyperglycemia; dehydration; common cold; alkalosis; diarrhea; lightheadedness; actual discomfort; chest pain; syncope; vomiting; dyspnea; nausea; pain; death anxiety; cold intolerance; exanthema; agitation; tremor; dizziness; weakness; chills; fever with chills; single organ dysfunction; myalgia; fever
Risk factors	infections, hospital; chronic disease; acquired immunodeficiency syndrome; diabetes mellitus; kidney diseases; sepsis due to fungus; infections of musculoskeletal system; pneumonia; soft tissue infections; hiv infections; urinary tract infection; candidiasis; senility

Table 4.6: Extracted UMLS terms for the text sections about the definition, the typical symptoms and the risk factors of septicemia cohort

Primary	respiratory failure without hypercapnia; acute respiratory failure; acute-on-chronic respiratory failure; spastic ataxia, charlevoix-saguenay type; respiratory failure; chronic respiratory failure; respiratory depression; respiratory tract structure; respiratory tract infections; lower respiratory tract structure; lower respiratory tract infection
Signs/symptoms	increased sweating; restlessness; shallow breathing; cardiac arrhythmia; irregular heart beat; unconscious state; drowsiness; tachypnea; sweating; confusion; anxiety
Risk factors	chronic obstructive airway disease; lung diseases; communicable diseases; asthma; heart failure; airway obstruction; chronic lung disease; infectious disease of lung; thrombophilia; pneumothorax; respiration disorders

Table 4.7: Extracted UMLS terms for the text sections about the definition, the typical symptoms and the risk factors of respiratory cohort

Primary	myocardial infarction
Signs/symptoms	pain;tachycardia;anxiety;heartburn;influenza;coughing;actual discomfort;nausea;vomiting;observation of attack
Risk factors	hypercholesterolemia; hypertensive disease; stress; diabetes mellitus; spasm; arteriopathic disease; hyperglycemia; preeclampsia; coronary artery disease

Table 4.8: Extracted UMLS terms for the text sections about the definition, the typical symptoms and the risk factors of myocardial infarction cohort

on the path from the root until that node excluding drugs with higher prescription frequency on the same level of that drug and those with greater node frequency on the same level of parent nodes. We interpret the illustrated prescription tree as follows. Starting with the root , the risk factor drug vancomycin is the most prescription drug which was prescribed for 26 patients. Considering the patients who were prescribed with vancomycin, the primary drug metronidazole is the most prescription drug. It was used to treat 16 patients among those who used vancomycin. Considering the patients who were not prescribed with vancomycin, insulin lispro is the most prescription drug. It was used to treat 6 patients among those who did not use vancomycin. The other nodes of the tree could be interpreted in a similar way. One can notice that in some nodes the number in a parent node may not be equal to the sum of the numbers in its child nodes. Such unbalance mass sometimes happens as there are some drugs in the datasets with unknown indication labels. Those nodes are excluded from the prescription tree, and therefore, cause such inequality of node frequency between a parent node and its children nodes.

In the above prescription tree, the drug sequences {VANCOMYCIN, METRONIDAZOLE, FUROSEMIDE, INSULIN LISPRO, ACETAMINOPHEN, ASPIRIN} is the set of prescription drugs with high node frequency. Those drugs are often prescribed together and hence they can be considered as the frequent treatment pattern of the subcohort.

In the indication labeling approach described in the previous section, it

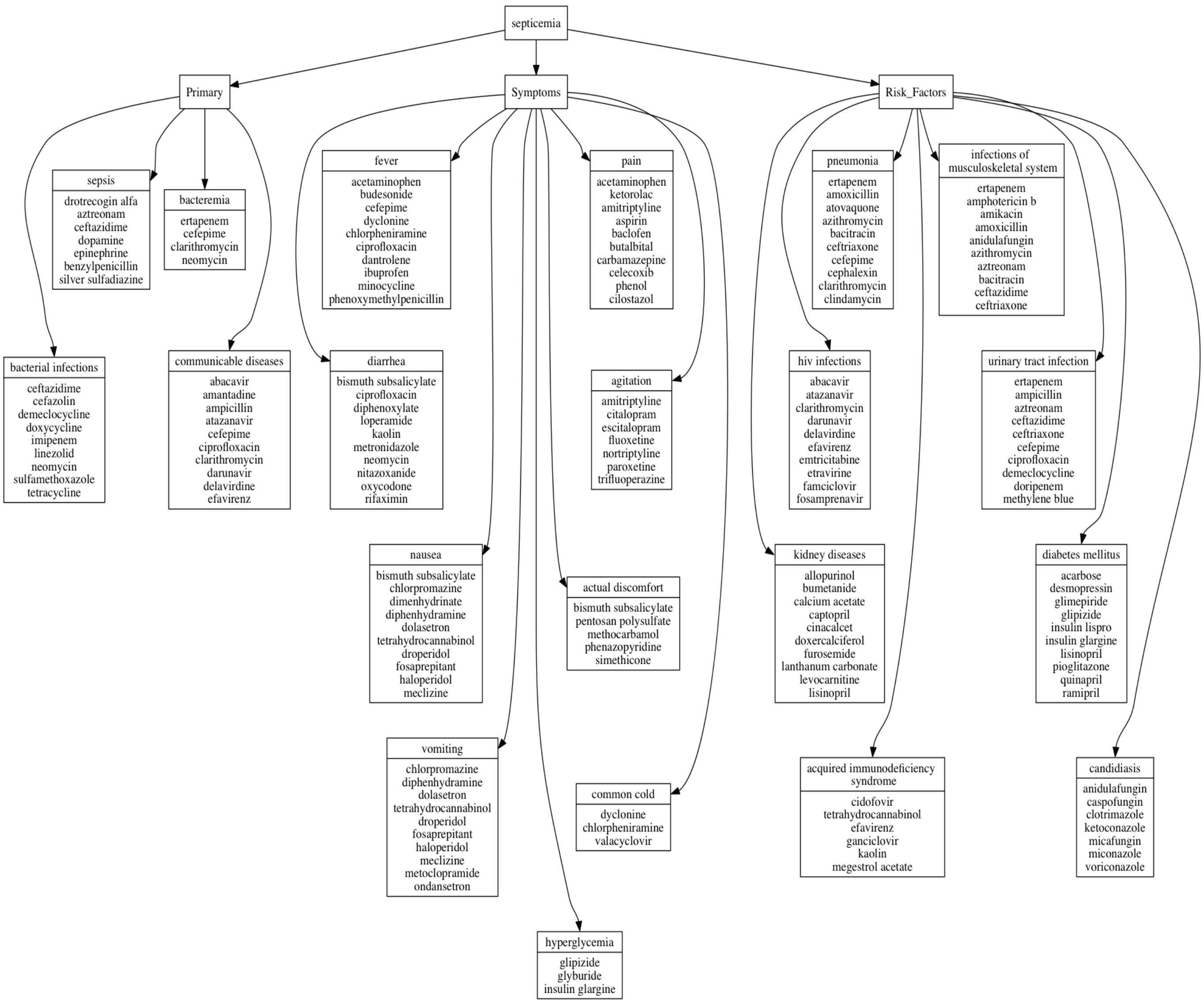


Figure 4.7: Extracted typical symptoms and drugs classified in three groups for septicemia cohort

is noted that we can recognize both the indication group of a prescription drug and the signs/symptoms treated by that drug as well. These characteristics allow one can understand in deep what symptoms underlying the

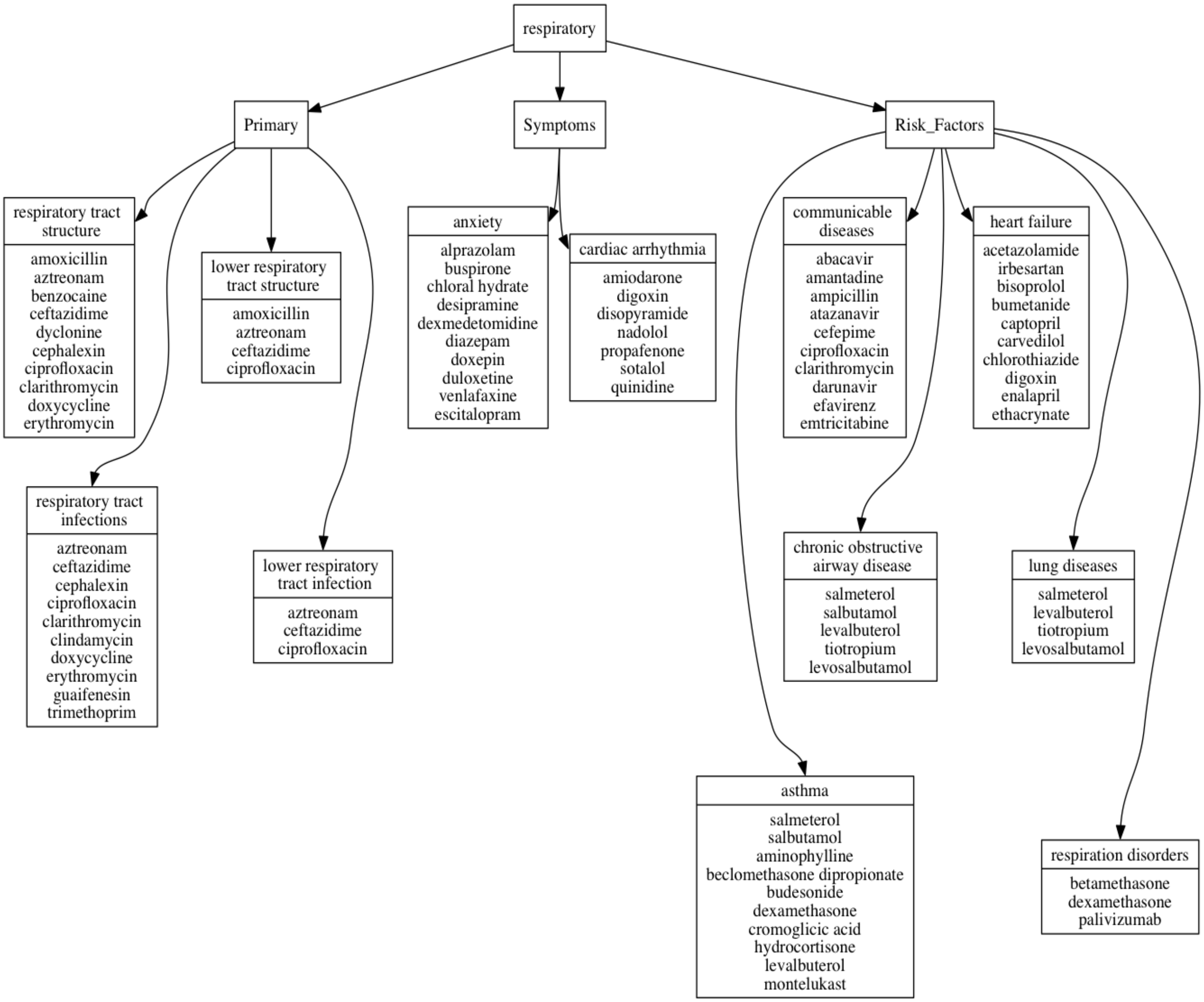


Figure 4.8: Extracted typical symptoms and drugs classified in three groups for respiratory cohort

derived treatment patterns. For instance, based on the indication labeling framework, one can infer that metoprolol, furosemide, insuplin lispro are often used to treat the infection of musculoskeletal system, kidney diseases and

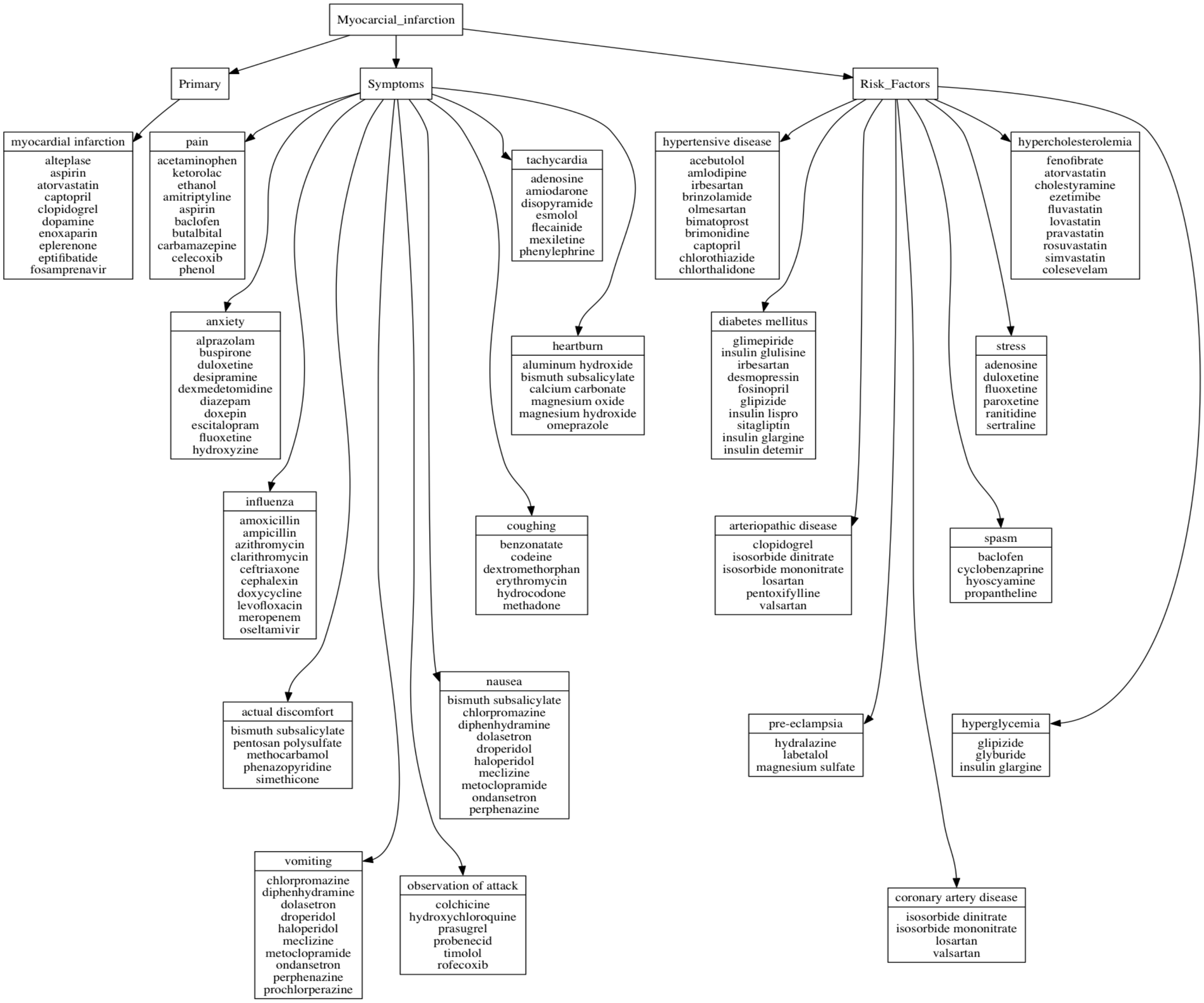


Figure 4.9: Extracted typical symptoms and drugs classified in three groups for myocardial infarction cohort

diabetes, respectively, while acetaminophen, aspirin have pain relief effect. By putting together those supporting facts, it is believed that the patients in the sample prescription tree mostly have some issues related to muscu-

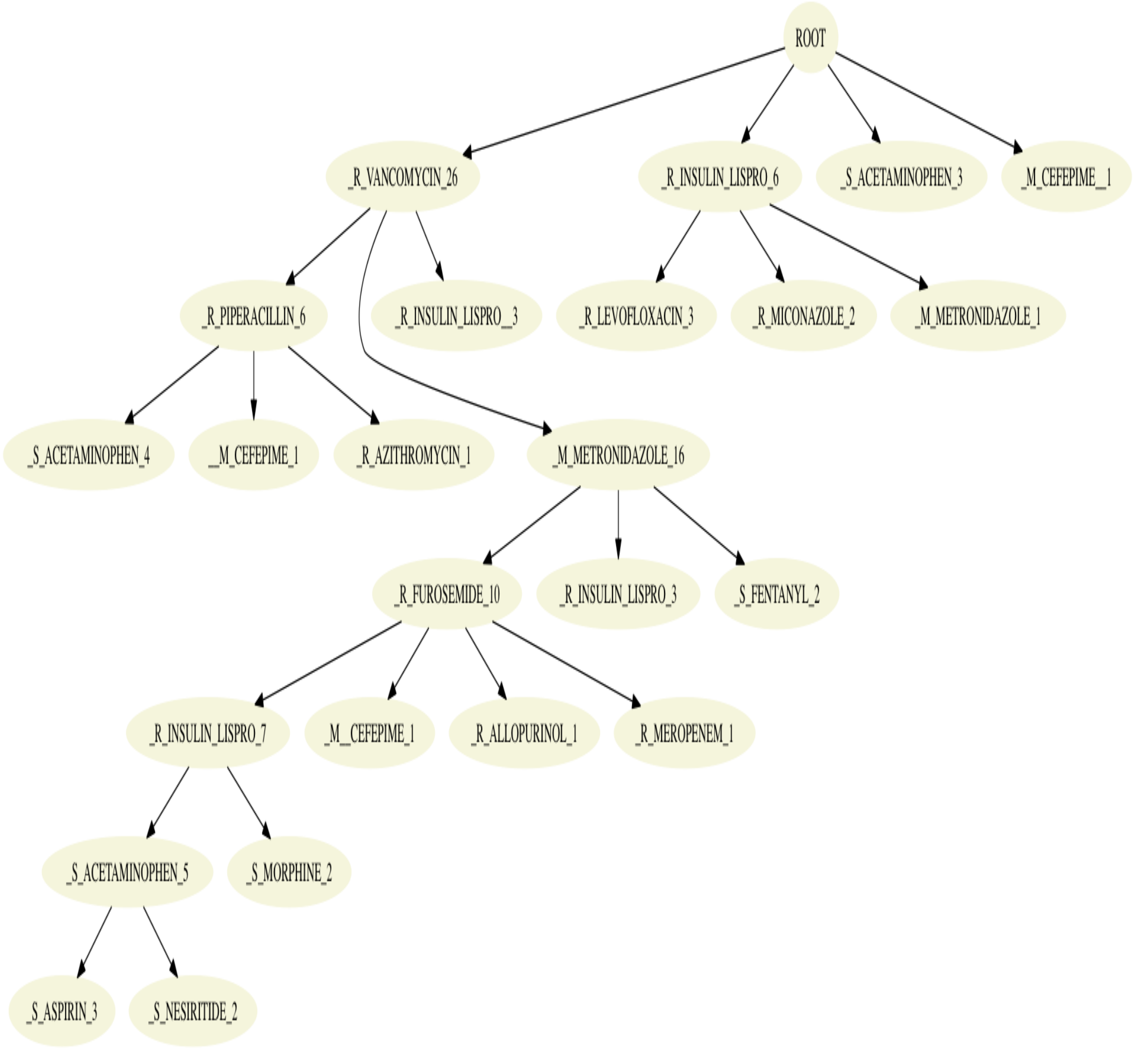


Figure 4.10: An example of resulting prescription tree

loskeletal, diabetes or kidney diseases. These are highly risky elements that possibly cause septicemia.

In addition to frequent pattern drugs, the tree also provides information about sets of infrequently prescription drugs. For example, it can be seen that piperacillin is unlikely to be used together with metronidazole. Another kind of treatment patterns is conditional treatment patterns, i.e which drugs are usually prescribed or not prescribed together given some must use drugs. For instance, among patients who were treated with vancomycin, it can be seen that furosemide is almost not prescribed without metronidazole. The derived patterns demonstrate the helpfulness of constructing prescription trees in our work. During the construction phase, frequency usage of drugs in the subcohort are preserved as much as possible and help the tree can reveal more kinds of patterns compared to other works in the literature. Doctors can look up the trees to quickly identify different patterns in making a treatment decision.

Table 4.9 provides a comparision of different features between the proposed learning method and the related works. [Sun et al., 2016, Huang et al., 2012, Huang et al., 2015]. We note that only a qualitative comparison is made as there is a difference in our problem formulation compared to previous work. This makes our problem setting are uneasy to be transformed into the known ones in the literature. Said differently, it is not straightforward to conduct experimental evaluation between our work and the related works.

Feature	**Our work**	**Related works**
Using different kinds of patient info	Yes	Limited
Domain incorporation	Yes	Limited
Addressing the treatment period identification	Yes	No
Understanding disease and drug relation	Yes	No
Treatment patterns reveal frequent pattern drugs	Yes	Yes
Treatment patterns reveal drugs used in conjunction with other drugs	Yes	Limited
Understanding symptoms underlying the treatments	Yes	Limited

Table 4.9: A qualitative comparison between our proposed treatment learning method and the related works

To evaluate the obtained results from the domain perspective, we asked for feedback from a few doctors. Generally, they understood and accepted the immediate results regarding curing relationships. Some groups should be re-categorized with little effort. For the prescription trees, since patient symptoms are very diverse, it is uneasy to verify the correctness of obtained prescription trees. Despite that, the consulting doctors, in general, recognized the meaningfulness of constructed prescription trees. They also advised that the prescription tree in each subcohort should not be too large to keep the treatment consistent.

4.4 Discussion and Conclusion

Different from previous studies, we address different issues and achieve more interesting outputs by exploiting domain knowledge. Instead of splitting prescription records by fixed periods, we take into account drug indication as an element to measure the change of indication in prescription records over time and cut them into periods flexibly. In our algorithm, different kinds of delivered drugs are taken into account for measuring the strength of indication change. The milestone points are marked as the timestamps with significant changes in prescription indication. The idea tries to capture the intuition of period detection given their treatment data. To the best of our knowledge, our study is the first work addressing the treatment period identification issue using electronic medical records.

Next, by deriving new representation via prescription trees, the learning method not only reflects almost completely the frequency of drugs in a concise form, but also enables physicians to identify sets of frequent and infrequent prescription drugs for every subcohort. Therefore, our treatment learning method seems to be superior to most of the current studies where merely frequent patterns are focused.

More interestingly, in addition to the treatment learning method, we have also developed an indication assignment framework that allows extracting typical signs, symptoms and commonly used drugs for a particular disease or group of diseases. By incorporating a comprehensive medical ontology like UMLS in the proposed method, we have reasonably addressed many challenges in clinical text processing, for example, resolving synonym drug or clinical term variation issues. The output of the indication assignment framework intuitively presents different factors characterizing a disease and medications curing these factors. It is helpful for everybody to understand easily about typical symptoms, drugs and their relations.

In short, this chapter has introduced a treatment learning method which is featured with the ability to identify treatment periods, label prescription indication, extract clinical information, maximize data utilization, and represent knowledge more appropriately. The proposed treatment learning method not only is able to assist physicians but also encourage researchers to exploit medical domain knowledge for a better interpretation of data-driven models.

Chapter 5

Treatment Recommendation Method

5.1 Introduction

Modern life brings a lot of facilities together with potential harmful factors affecting our health. Nowadays, more and more people are getting multiple diseases. Treating those patients is so complicated that a straightforward combination of clinical guidelines for individual diseases may result in conflicting decisions. For example, patients with a gastrointestinal bleeding problem are advised to avoid using aspirin. On the other hand, patients with kidney disease are suggested to be prescribed with aspirin in most of the clinical guidelines. A patient having both the gastrointestinal bleeding problem and the kidney disease may cause trouble for inexperienced physicians to determine whether aspirin should be used in those cases.

The above example illustrates the need for computer-aided systems assisting physicians in making treatment decisions. In recent years, many machine learning approaches have been proposed to address the treatment recommendation problem. The two promising ones are deep learning-based approach and reinforcement-based approach. While the deep learning-based studies have extended deep neuron network models [Snow et al., 1994, Zhang et al., 2017, Nezhad et al., 2019, Katzman et al., 2018] or long-short-term memory models [Liao and Ahn, 2016, Le et al., 2018, Shang et al., 2018, Pham et al., 2017] to feed the disease, treatment sequences, and their relation; the reinforcement-based studies often aim at optimizing a sequence of treatments that maximize the treatment outcome [Murphy, 2003, Weng et al., 2017, Liu et al., 2017, Prasad et al., 2017]. Both approaches, despite their promise, are hard-interpretable under the perspective of medical context or hard to be applied for real-world electronic medical records where treatment outcome could not be identified easily.

Motivated from the above drawback of current works, we seek an interpretable treatment recommendation models that can be applied for generic electronic medical records. Intuitively, one can suggest treatment for a new patient based on prescription drugs of the most similar known patient, i.e. the nearest neighbor patient. However, it is uneasy to identify such patient since his data may not have been recorded in the database, or he is just similar to the new patient partially in several aspects. As a result, the treatment of a new patient and the treatment of his nearest neighbor patient may not be identical in reality. To this end, we propose a neighbor-based treatment recommendation method that mimics human intuition to suggest treatment for new patients based on their K neighbors' treatments.

In the previous chapter, we have introduced a treatment learning method that is able to learn the treatment patterns of patient subcohorts. This section presents a treatment recommendation method that utilizes the results of the learning method to assist physicians in the treatment recommendation task. We first suggest different ways to discover patient subcohorts under the two learning aspects named symptom-based learning treatment-based learning aspects. We then propose different methods to exploit derived treatment paths for the recommendation task that can be used for each learning aspect or both of them.

5.2 Methodology

First, we consider different aspects named symptom-based and treatment-based aspects that could be used to derive treatment patterns. The first aspect groups training patients by symptom-based features, i.e. non-treatment features such as symptoms, laboratory indicators or demographics whereas the second aspect identifies patient sub-cohorts based on treatment-based features, i.e. prescription records of training patients. The underlying assumption for the symptom-based recommendation method is that similar patients in terms of symptom-based features can be treated in similar ways. In the inverse direction, the treatment-based method assumes that patients who treated in similar ways may have commons in symptom-based features.

We hypothesize that each assumption is the complementary one of the other to cover the case of treatment variants. The following example illustrates the motivation of considering two learning aspects. Suppose symptom S could be treated by two treatments T_1 and T_2. If we solely base on the first assumption that symptom S can be treated only by treatment T_1 , it can not be used to explain the case symptom S can be treated by treatment T_2. Similarly, the second assumption is not sufficient to explain the case different symptoms can be treated with the same treatment. Therefore, taking into account the treatment-based and symptom-based aspects that lead to the form of sub-cohorts is hypothesized to cover variants of treatment for similar patients.

5.2.1 Symptom-based learning aspect

Figure 5.1 redraw the overview of the symptom-based learning method that has been presented in the previous chapter. In this section, we shortly summarize its steps. The goal of symptom-based learning method is to divide training patients into sub-cohorts and learn treatment patterns for each sub-cohort. To cluster patients, our method represents training patients by their symptom-based features, e.g. initial laboratory indicators, signs or symptoms extracted from nursing notes, or demographics data. Since those features are usually mixed types of numerical, binary, categorical values or text, we employ MV.RBM, a representation model for mixed data, to transform patients' heterogeneous input vectors to homogeneous binary vectors. The representation vectors are then ready to be input for many clustering methods.

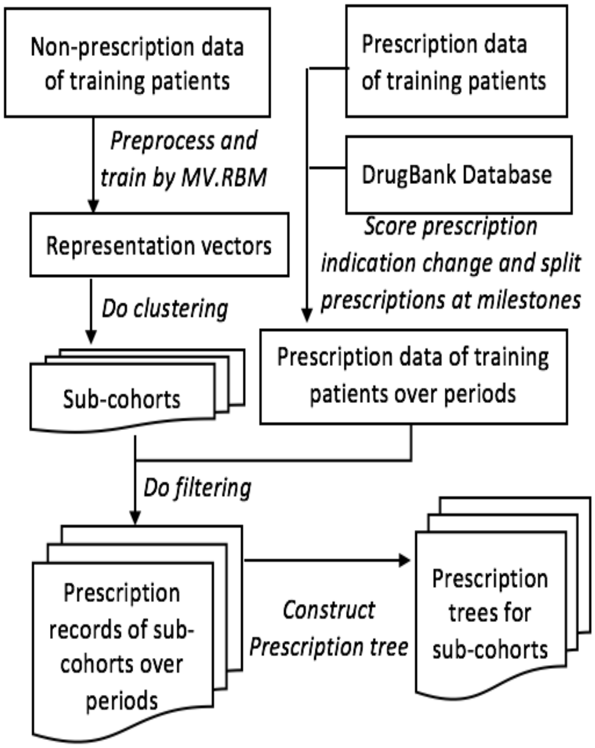

Non-prescription data of training patients
Prescription data of training patients
DrugBank Database
Preprocess and train by MV.RBM
Representation vectors
Score prescription indication change and split prescriptions at milestones
Do clustering
Sub-cohorts
Prescription data of training patients over periods
Do filtering
Prescription records of sub-cohorts over periods
Construct Prescription tree
Prescription trees for sub-cohorts

Figure 5.1: The symptom-based learning method

To deal with the varying lengths of prescription records when identifying treatment periods, the symptom-based learning method tries to capture the intuition that a new treatment period is made when there is a considerable change in the indication of prescription records. To characterize the strength of indication change, we compute an accumulated score at each timestamp when a patient is prescribed. It then splits the whole prescription records into intervals such that significant changes in the prescription indication happen at the beginning stage of each interval. The accumulated score at each timestamp is the sum of the accumulated score at previous timestamp and the cardinalities of sets of newly delivered drugs, recent-stopped delivered drugs, redelivered drugs weighted by the importance of prescription indications.

To construct treatment patterns for each sub-cohort in a period, we organize treatments in a tree structure. Along the nodes of the treatment tree, each node is the next highest frequency among prescription records of patients who were treated by drugs from the root until its parent node and were not treated by drugs from the left-hand side nodes in the same level of the tree. Each node also keeps track of patient ID of those patients prescribed with the drugs from the root until the considering node. It is noted that there is no chronological order among drugs on a treatment path of a sub-cohort. Instead, these drugs are sorted by prescription frequency in the sub-cohort.

5.2.2 Treatment-based learning aspect

We explore another view that can be used to construct sub-cohorts for training patients. In the symptom-based learning method, while patients are grouped according to their symptom-based features, the treatment-based learning method supposes that training patients can also be grouped by their treatments. Figure 5.2 describes an overview of the treatment-based learning method. It is almost identical to the symptom-based learning one except for the features used to represent training patients.

In the previous chapters, we have introduced an indication assignment framework which can extract the sign/symptom/disease terms belonging to the main group, symptom group and risk factor group of the considering cohort. Let $Terms = \{t_1^M, t_2^M, ..., t_1^S, t_2^S, ..., t_1^R, t_2^R, ...\}$ denote these terms, respectively. As prescription records are complicated objects varying in lengths, dosages and drug labels, we represent each patient by new prescription-based features, namely abstract features $F = \{f_{t_1^M}, f_{t_2^M}, ..., f_{t_1^S}, f_{t_2^S}, ...f_{t_1^R}, f_{t_2^R}, ...\}$, where each element in the set is the number of drugs in patient prescription records that cure the corresponding sign/symptom/disease. In concrete, we calculate the proportion of drug curing each medication term in F. The

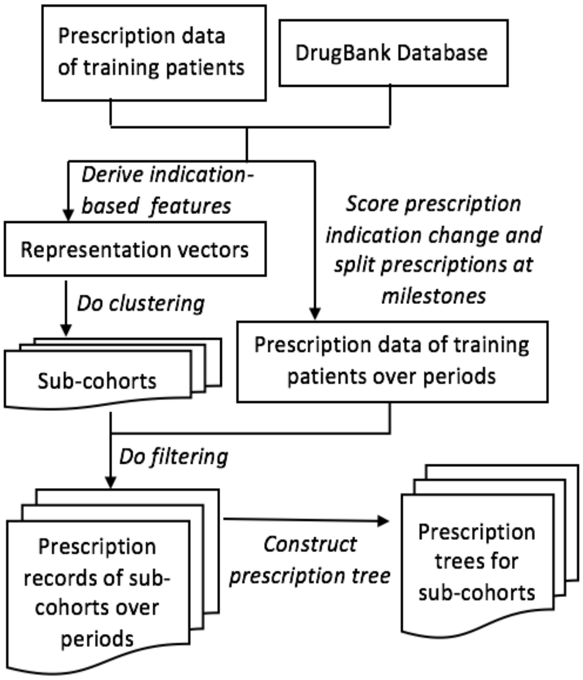

Prescription data of training patients
DrugBank Database
Derive indication-based features
Representation vectors
Score prescription indication change and split prescriptions at milestones
Do clustering
Sub-cohorts
Prescription data of training patients over periods
Do filtering
Prescription records of sub-cohorts over periods
Construct prescription tree
Prescription trees for sub-cohorts

Figure 5.2: The treatment-based learning method

Algorithm 4: Deriving treatment features for a patient

```
Data: $\Theta$, $T$
Result: return $F$ as a vector of new prescription-based features
Initialize all element in $F$ to 0 ;
for each $t \in T$ do
    $D := \{dr \mid \forall dr \in \Theta \wedge dr.startdate == t\}$ ;    ▷ date d's delivered drugs
    for each $dr \in D$ do
        for each $term \in Terms$ do
            if $Cure(dr, term) == True$ then
                $f_{term} = f_{term} + 1$
Return $F$
```

Algorithm 4 provides the detailed derivation of treatment-based features for training patients.

5.2.3 Recommendation over single learning aspect

Figure 5.3 describes our idea for the treatment recommendation method. Given a new patient p, we collect his initial non-treatment-based feature vector and transfer it to the binary form through the parameters $\hat{\theta} = (\hat{\mathbf{a}}, \hat{\mathbf{b}}, \hat{\mathbf{w}})$ learned by the trained MV.RBM used for representing mixed type vectors of training patients. Let h^p be the L dimension binary hidden vector of p, $h^{p'}$ be the binary hidden vector of training patient p', the similarity $d^p_{p'}$ between two patients is defined by the Hamming distance between their latent representations.

$$d^p_{p'} = Hamming(h^p, h^{p'}) = \frac{1}{L}\sum_{i=1}^{L} I(h^p_i \neq h^{p'}_i)$$

In our proposed method, we utilize the resulting prescription trees to find the K associated treatment paths of the neighbors. Each treatment path is considered as a set of typical drugs treating one of the K neighbors and therefore it is taken into consideration in the construction of treatment for new patient p. It is worth noting that to capture the variant of treatments of similar patients, the K neighbor patients and their associated treatment paths can belong to different subcohorts. Let $\delta^{p'} = \{dr^{p'}_1, dr^{p'}_2, ...dr^{p'}_M\}$ denote the set of drugs linked with the treatment path of p', $p_1, p_2, ..., p_K$ denote the K neighbors of p. Their associated treatment paths $\delta^{p_1}, ..., \delta^{p_K}$ and the distances $d^p_{p_1}, d^p_{p_2}, ...d^p_{p_K}$ are then utilized to recommend top M drugs for p. Let $C = \{dr_1, dr_2, ..., dr_j\}$ be the set of distinct drugs named candidate drugs jointed from $\delta^{p_1}, ..., \delta^{p_K}$; H^{train} be the matrix consisting of representation vectors of training patients.

The intuition underlying our approach is that prescription drugs delivered to many neighbors are likely to be used for the new patient. Therefore, for every candidate drug dr in C, we compute its path frequency $freq^p_d$, i.e. the number of treatment paths contains dr as one of the criteria for

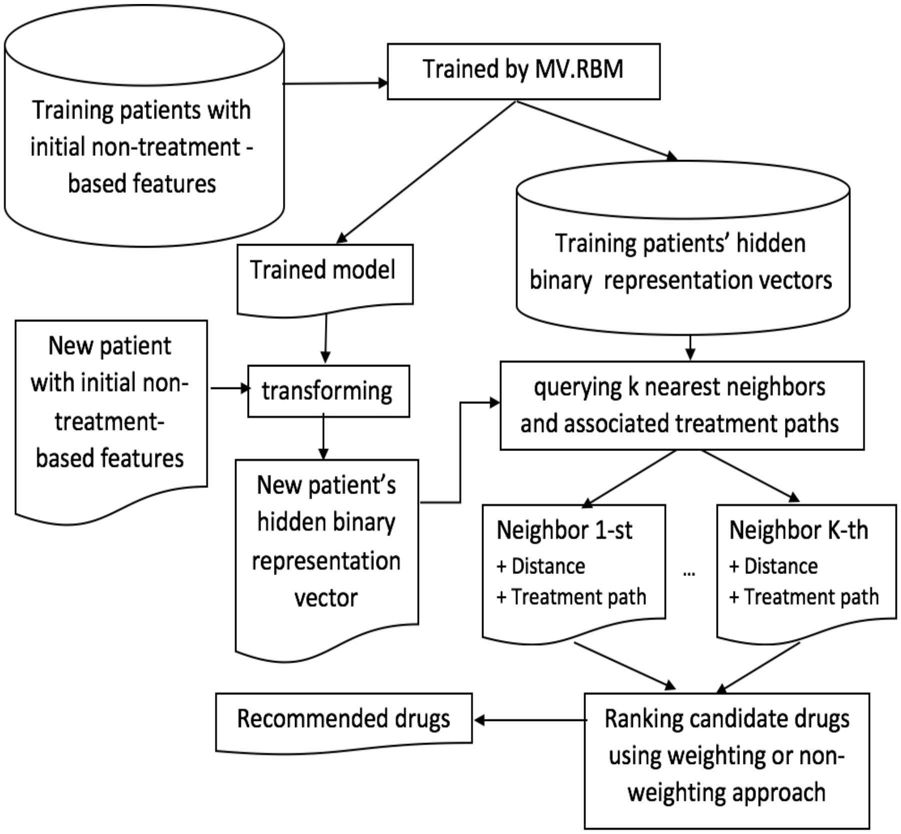

Training patients with initial non-treatment - based features
Trained by MV.RBM
Trained model
Training patients' hidden binary representation vectors
New patient with initial non-treatment-based features
transforming
querying k nearest neighbors and associated treatment paths
New patient's hidden binary representation vector
Neighbor 1-st
+ Distance
+ Treatment path
...
Neighbor K-th
+ Distance
+ Treatment path
Recommended drugs
Ranking candidate drugs using weighting or non-weighting approach

Figure 5.3: An overview of the treatment recommendation method

recommendation. Drugs with higher path frequency indicate that they are prescribed for a greater number of neighbor patients and hence, have a higher chance to be recommended. The formula of $freq^p_{dr}$ is provided below.

$$freq^p_{dr} = \sum_{i=1}^{K} I(dr \in \delta^{p_i}) \tag{5.1}$$

To solve the case dr has the same path frequency with other drugs, we consider a distance priority metric d^p_{dr}, another measure which takes into account the distance from test patient to the neighbors whose treatment paths contain dr. The greater the sum of the inverse distance from those neighbors to p, the higher priory the drug is selected as recommendation drugs. We provide the formula of d^p_{dr} as follows.

Algorithm 5: Recommending prescription drugs for new patient p in a specific treatment period by using neighbors' treatment patterns derived from single learning aspect

```
Data: $\Lambda$, $\theta$, $v^p$, $H^{train}$
Result: return top M recommended drugs
Compute $h^p = P(h|v^p, \theta)$;
Compute similarity between $h^p$ and each training patient's representation
  vector $h^{p'}$ in $H^{train}$;
Select $K$ most similar patients $p_1, p_2, ..., p_K$;
Construct prescription tree in the considering learning aspect using
  algorithm 3;
Query associated treatment paths $\delta^{p_1}, \delta^{p_2}, ..., \delta^{p_K}$ through tracing variables
  $\Lambda$ ;
$C = \bigcup_{i=1}^{K} \delta^{p_i}$ ;
for each $dr \in C$ do
    Compute $freq^p_{dr}$ by equation (5.1);
    Compute $dist^p_{dr}$ by equation (5.2);
Return top $M$ drugs sorted by ($freq^p$, $dist^p$);
```

$$dist^p_{dr} = \sum_{i=1}^{K} I(dr \in \delta^{p_i}) \times \frac{1}{d^p_{p^i}} \tag{5.2}$$

Algorithm 5 provides the pseudocode describing the procedure to recommend prescription drugs for new patient p in a period over a single learning aspect.

5.2.4 Ensemble recommendation over dual learning aspect

While the proposed recommendation method seems to be able to capture treatment variants of neighbor patients in different sub-cohorts, it simply combines treatment paths of neighbor patients from a single learning aspect, e.g treatment-based learning aspect or symptom-based learning aspect. As there are different aspects that could form the patient sub-cohorts and the learned treatment patterns, it may be useful to adopt the ensemble idea on the multi-learning aspect to enhance the ability of recommendation method in capturing treatment variants.

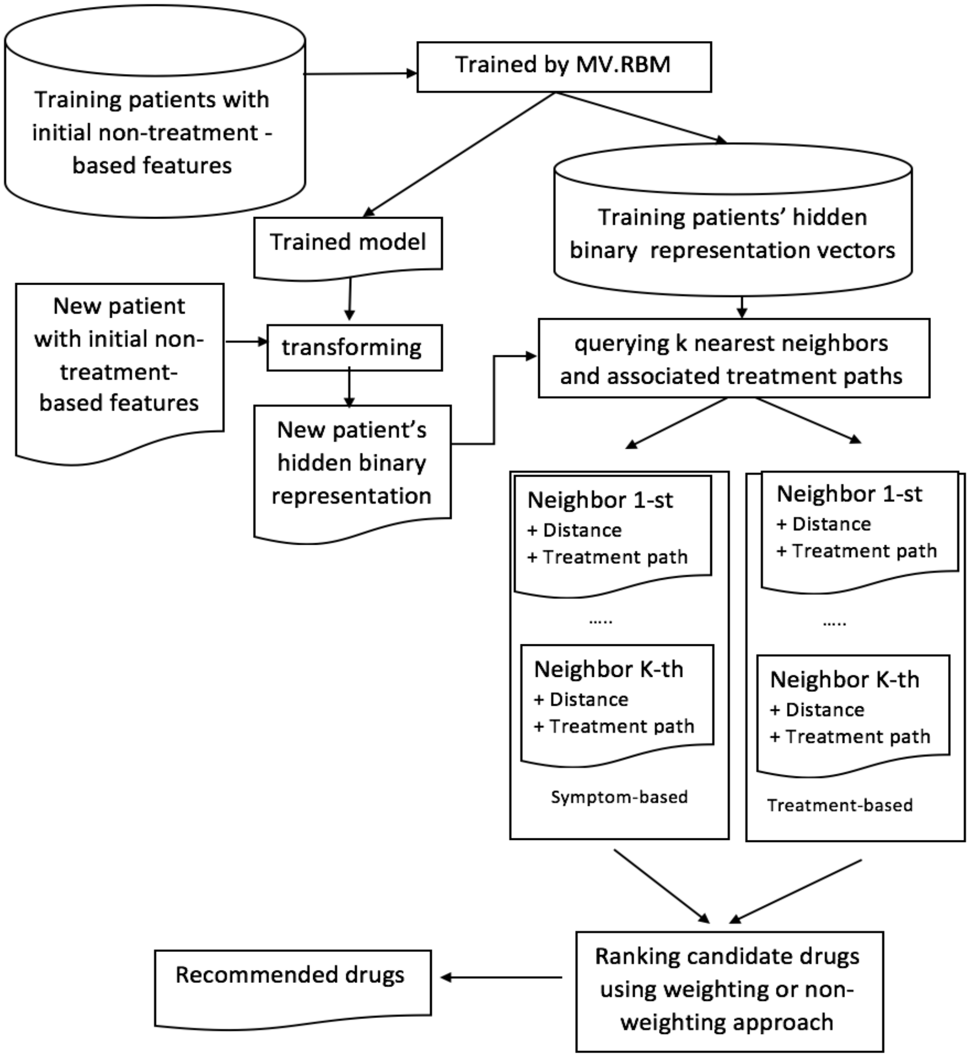

Training patients with initial non-treatment - based features
Trained by MV.RBM
Trained model
Training patients' hidden binary representation vectors
New patient with initial non-treatment-based features
transforming
querying k nearest neighbors and associated treatment paths
New patient's hidden binary representation
Neighbor 1-st
+ Distance
+ Treatment path
.....
Neighbor K-th
+ Distance
+ Treatment path
Symptom-based
Neighbor 1-st
+ Distance
+ Treatment path
.....
Neighbor K-th
+ Distance
+ Treatment path
Treatment-based
Ranking candidate drugs using weighting or non-weighting approach
Recommended drugs

Figure 5.4: The ensemble treatment recommendation method over dual learning aspects

For that reason, we develop an ensemble treatment recommendation method

Algorithm 6: Recommending prescription drugs for new patient p in a specific treatment period by using neighbors' treatment patterns derived from dual learning aspect

```
Data: $\Lambda$, $\theta$, $v^p$, $H^{train}$
Result: return top M recommended drugs
Compute $h^p = P(h|v^p, \theta)$;
Compute similarity between $h^p$ and each training patient's representation vector $h^{p'}$ in $H^{train}$;
Select $K$ most similar patients $p_1, p_2, ..., p_K$;
Construct prescription tree in the symptom-based learning aspect;
Construct prescription tree in the treatment-based learning aspect;
Trace associated treatment paths $\delta^{p_1}, \delta^{p_2}, ..., \delta^{p_K}$ under the symptom-based learning aspect through tracing variables $\Lambda$ ;
Trace associated treatment paths $\delta'^{p_1}, \delta'^{p_2}, ..., \delta'^{p_K}$ under the treatment-based learning aspect through tracing variables $\Lambda'$ ;
$C = \bigcup_{i=1}^{K} \delta^{p_i}, \delta'^{p_i}$ ;
for each $dr \in C$ do
    Compute $freq^p_{dr}$ by equation (5.3);
    Compute $dist^p_{dr}$ by equation (5.4);
Return top $M$ drugs sorted by ($freq^p$, $dist^p$);
```

that span candidate treatment paths not only sub-cohorts but also the learning aspects that form patients sub-cohorts. Figure 5.4 provides an overview of the proposed ensemble recommendation method over dual learning aspects. Our approach is similar to the previous, but it looks for treatment paths of neighbor patients in both treatment-based and symptom-based learning aspects. For each neighbor, we concatenate the associated treatment paths in two learning aspects and follow the same aggregation mechanism as described for the method on a single learning aspect to suggest possibly prescribed drugs. It is noted that the prescription frequency $freq^p_{dr}$ and the distance priority metric are adjusted as follows.

$$freq_{dr}^{p} = \sum_{i=1}^{K} \big(I(dr \in \delta^{p_i}) + I(dr \in \delta'^{p_i})\big) \tag{5.3}$$

$$dist_{dr}^{p} = \sum_{i=1}^{K} \big(I(dr \in \delta^{p_i}) \times \frac{1}{d_{p^i}^{p}} + I(dr \in \delta'^{p_i}) \times \frac{1}{d_{p^i}^{p}}\big) \tag{5.4}$$

Algorithm 6 provides the pseudocode of the procedure to generate recommendation drugs by using neighbors' treatment paths learned from dual learning aspects.

We note that although each patient is treated with a unique set of drugs in a treatment period, the treatment paths in two learning aspects may be different. This is because in two learning aspects, a patient can belong to two slightly different subcohorts. As a result, this property can affect the selection of typical drugs in each learning aspect since it strongly depends on the parameter ϵ and the subcohort that patient belongs to. Table 5.1 shows an illustration of treatment paths of a patient taken from our experiment.

Actual drug	ciprofloxacin, metronidazole, insulin lispro, vancomycin, aspirin, fentanyl, haloperidol
Treatment path in symptom learning aspect	aspirin, vancomycin, metronidazole
Treatment path in treatment learning aspect	insulin lispro, fentanyl, vancomycin, metronidazole, ciprofloxacin

Table 5.1: Example of the treatment paths in symptom and treatment learning aspects for the same patient

5.3 Experimental Evaluation

5.3.1 Evaluation metric

The notations for the evaluation metrics are denoted as follows.

- M: the number of recommended drugs.
- T: the test set, i.e set of new patients.
- n: the number of treatment periods.
- $\hat{D}_p^{M,\pi_j}$: the top M recommended drugs for the testing patient p over period π_j.
- $D_p^{\pi_j}$: the set of actual prescription drugs for p in period π_j.

We use precision, recall, and F1 score, the three well-known evaluation metrics, to evaluate the performance of our treatment recommendation method. The formulas of these metrics are given as follows.

$$recall@M = \frac{1}{|T| \times n} \sum_{p \in T} \sum_{j=1}^{n} \frac{|\hat{D}_p^{M,\pi_j} \cap D_p^{\pi_j}|}{|D_p^{\pi_j}|}$$

$$precision@M = \frac{1}{|T| \times n} \sum_{p \in T} \sum_{j=1}^{n} \frac{|\hat{D}_p^{M,\pi_j} \cap D_p^{\pi_j}|}{M}$$

$$F1@M = \frac{2 \times precision \times recall}{precision + recall}$$

For every experiment, we repeat it three times on different training and testing set. The report values of precision, recall and F1 scores in this dissertation are the mean values.

5.3.2 Dataset and parameter setting

We used the same three datasets which were described in the previous chapter. In each dataset, 80% of the dataset was used as the training set and the rest was used as the testing set.

Regarding the parameter number of neighbors K, we investigated the behavior of the proposed methods on different values of K among following values $\{3, 5, 7, 10, 15, 20, 30, 40, 50, 60, 80, 100, 150, 200\}$. We set the parameter $\epsilon = 5$.

5.3.3 Baselines

We note that the methodologies of our proposed method and most of the related works are not easy to be applied to each other due to the difference in the problem formulation. This is a common situation in most of the treatment recommendation studies where there is a high variance of problem formulations and data collection methods. Therefore we could not conduct experiments using approaches in the related works in our dissertation.

Instead, we consider our treatment recommendation problem as the top-M item recommendation problem where users are patients and items are prescription drugs. Although rich side information about patients such as patient demographics, indicators, nursing notes is available, it is not straightforward to exploit such information to leverage user preferences, i.e how likely a prescription drug is given to a patient.

Thus, we compared the efficacy of our proposed recommendation methods to an API dedicated for implicit collaborative filtering recommender system (ICF) implemented in Graphlab [1]. This API was implemented based on the idea presented in [Koren et al., 2009, Hu et al., 2008, Rendle, 2010]. In the implicit feedback dataset, there is no target value, the API uses the logistic loss to fit a model that attempts to predict all the given (user, item) pairs in the training data as 1 and all others as 0 [2].

We compare the proposed methods to the following three baselines.

- Using Graphlab library for implict recommendation problem and the implicit alternative least square [Hu et al., 2008] solver, namely ICF + IALS.

- Using Graphlab library for implict recommendation problem and the stochastic gradient descent [Bottou, 2012] solver , namely ICF + SGD.

[1] `https://turi.com/products/create/docs/graphlab.toolkits.recommender.html`

[2] `https://turi.com/products/create/docs/generated/graphlab.recommender.ranking_factorization_recommender.RankingFactorizationRecommender.html`

- Using Graphlab library for implict recommendation problem and the adaptive stochastic gradient descent [Kingma and Ba, 2014] , namely ICF + ADA.

We note that the above baselines are not completely black-box. They can be explained in terms of mathematical viewpoint. However, it is hard to understand the treatment mechanism under the healthcare perspective since they try to find parameters that optimize functions taking all patients and drugs into consideration. In other words, they can not point out which neighbors the treatment recommendation process based on.

Our proposed treatment recommendation method are conducted in three following cases:

- Using the treatment recommendation method over the symptom-based learning aspect, namely TRoS.
- Using the treatment recommendation method over the treatment-based learning aspect, namely TRoT.
- Using the treatment recommendation method over both aspects, i.e dual learning aspect, namely TRoD.

We also compare our proposed methods with the approach that recommends treatment for a new patient based on the top M drug on the treatment path of the nearest neighbor's treatment path.

5.3.4 Illustration of recommendation procedure

To illustrate how the proposed recommendation methods work, we take an example of recommending prescription drugs for a new patient using symptom learning aspect with $K = 5$. We note that the following example is a real one extracted from our experiment for septicemia cohort.

Table 5.2 shows the prescription drugs of five neighbors of a test patient, their prescription drugs in a period and the corresponding treatment paths. Each path is selected as a sequence of drugs among the actual drugs

	Prescription drugs	**Treatment path**
Neighbor 1	metronidazole, insulin lispro, vancomycin, fentanyl, cefepime, ibuprofen, nesiritide, prochlorperazine	**metronidazole**, **vancomycin**, **cefepime**
Neighbor 2	aspirin, insulin lispro, meropenem, vancomycin, hydromorphone, levofloxacin, trimethoprim	aspirin, **insulin lispro**, **vancomycin**, levofloxacin
Neighbor 3	metronidazole, desmopressin , insulin lispro, piperacillin, levofloxacin	**piperacillin**
Neighbor 4	metronidazole, ceftriaxone, acetaminophen, insulin lispro, vancomycin, piperacillin	**metronidazole**, acetaminophen, **insulin lispro**, **vancomycin**, **piperacillin**
Neighbor 5	meropenem, metoclopramide	meropenem

Table 5.2: Illustration of treatment recommendation procedure

delivered to the corresponding neighbor such that they are also prescribed together in his sub-cohort with high frequency. The bold drugs are top five selected drugs based on their path frequency and the distance priority. In this example, all five recommendation drugs are matched with the prescription drugs of the new patient. The above example shows the interpretability of our recommendation mechanism. Recommendation drugs are derived from the common drugs of neighbors in the combination with the sub-cohort they belong to. This feature could not be found easily when using black-box or hard-domain interpretable approaches.

5.3.5 A comparison with using the nearest neighbor-based treatment (K=1)

This section presents a comparison of the efficacy of recommending treatment using K-neighbor based recommendation methods, i.e the TRoS, TRoT and TRoD methods with the nearest neighbor-based approach, i.e recommending treatment based on the top M drugs on the treatment path of the nearest neighbor patient.

Figure 5.5 shows evaluation measures of the two recommendation approaches. We report the obtained results of TRoS, TRoT and TRoD with $K = 5$ (randomly selected without parameter tuning), $M = \{3, 5, 10\}$. It can be seen that in all cases, the proposed methods obtain better precision and F1-score compared to using $K = 1$. The result shows the necessity of combining treatments from many neighbor patients' treatments.

5.3.6 A comparison between K-neighbor-based approaches

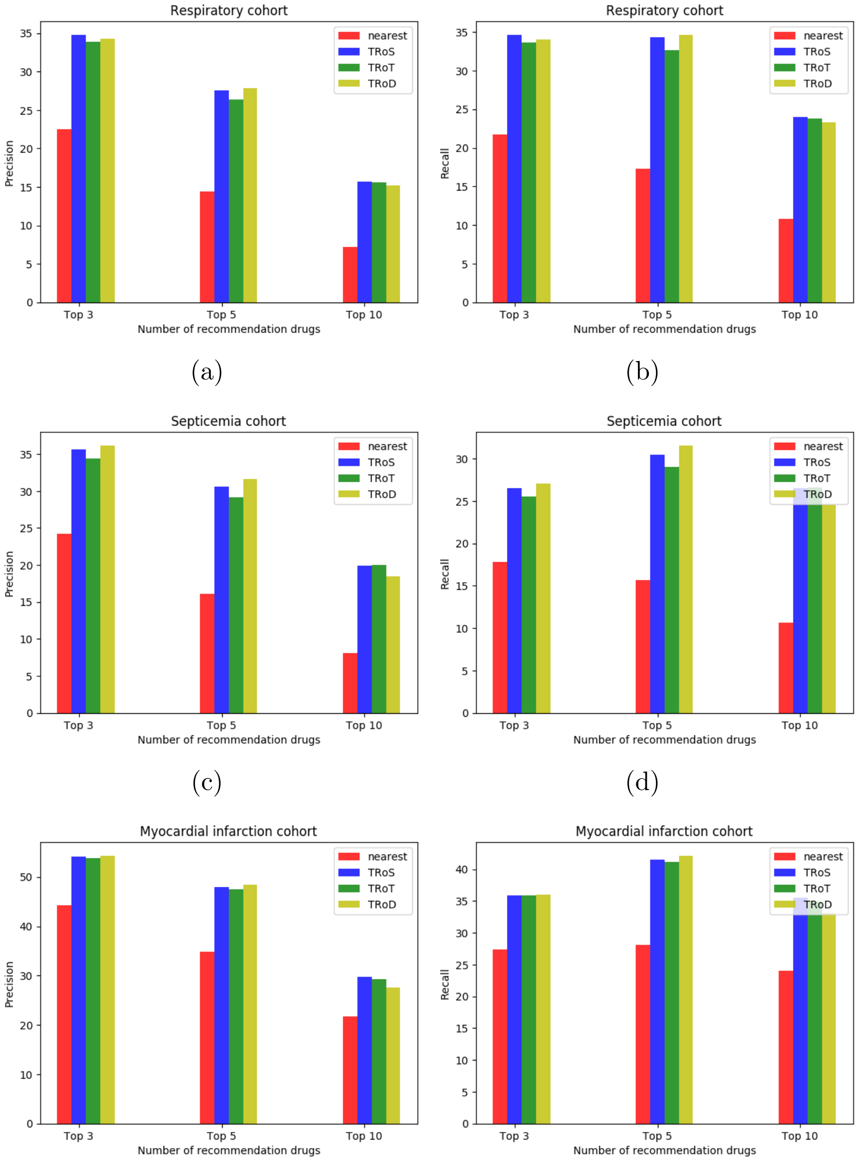
We next show the comparison among the proposed neighbor-based treatment recommendation methods. Through our intensive experiments, we report the obtained F1 score of the TRoS, TRoR, and TRoD when recommending a small number of drugs ($M = 3$) and a relatively large number of drugs ($M = 15$).

Figure 5.6 reports the behaviors of the three methods when the number of neighbors K is varied from 20 to 100. The obtained results indicate that for the small M case, there is no method that obviously outperforms the others. By contrast, when recommending a large number of drugs, the TRoD seems to be superior to the TRoS and the TRoT. All methods seem to perform steadily with large M. The F1 value tends to increase when the number of neighbors increases, but it is not significantly improved for the large K cases.

5.3.7 A comparison to the baselines

This section shows a comparison between the proposed methods and the baselines. We conducted the experiments to recommend top 3, top 5, and top 10 drugs. The parameters are varied with values selected in the parameter setting section. It is noted that the obtained results are reported by using the K with the best F1 score, namely K_{best}.

Tables 5.3, 5.4, 5.5 report the three evaluation measures obtained by the baselines and the proposed methods on three datasets. It can be seen that in the best cases, our proposed methods are comparable to the baselines. There are cases the baselines work slightly better and there are cases our proposed methods yield sharper results. In general, the difference is not significantly large. Since the reported values correspond to the small M, there are no dominated ones among the three proposed methods.



(a)

(b)

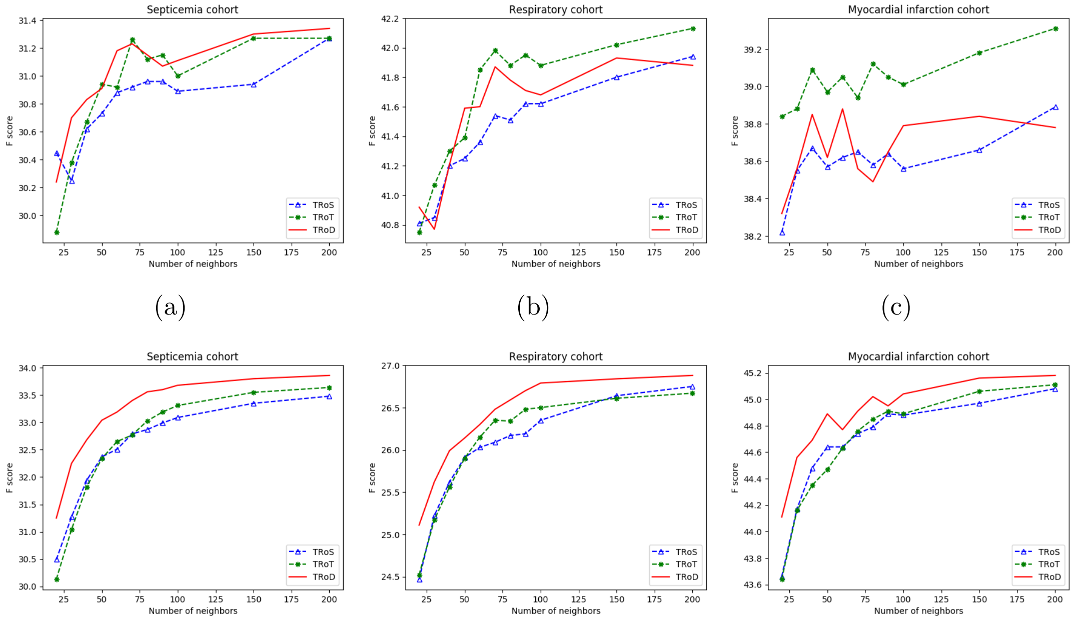
(c)

(d)

(e)

(f)

Figure 5.5: A comparison of the treatment recommendation methods in two cases: using the proposed K-neighbor-based recommendation approach and using the nearest neighbor-based recommendation approach



(a) (b) (c)

(d) (e) (f)

Figure 5.6: A comparison of treatment recommendation efficacy among the TRoS, TRoT and TRoD methods. Here we report the obtained F1 score performed on $M = 3$ (Figures (a), (b), (c)), and on $M = 10$(Figures (d), (e), (f))

Table 5.6 reports the corresponding K that yields the best F1 score. In most cases, the proposed methods have to take into account the treatment paths of a large number of neighbors to achieve the highest F1 score. In small M cases, the TRoD seems to take fewer neighbors' treatment than the TRoS and TRoT methods.

5.3.8 How good when using the small K

Although the proposed neighbor-based methods obtain comparable results with the baselines, they only work with a large number of K. In the healthcare context, this somewhat limits the interpretation of the proposed methods since physicians may only need to look through a few most relevant neighbors

Method	**Precision**			**Recall**			**F score**		
	@3	**@5**	**@10**	**@3**	**@5**	**@10**	**@3**	**@5**	**@10**
TRoS	41.33	33.06	21.62	42.58	56.52	71.07	41.94	41.71	33.15
TRoT	41.48	33.01	21.61	42.82	56.43	70.63	42.13	41.65	33.09
TRoD	41.31	33.04	21.64	42.58	56.16	70.83	41.93	41.6	33.14
ICF + SGD	42.13	33.4	21.75	43.83	56.68	71.16	42.95	42.03	33.31
ICF + IALS	20.1	23.19	17.96	19.33	39.0	59.89	19.7	29.06	27.62
ICF + ADA	42.15	33.44	21.72	43.48	56.78	70.9	42.8	42.08	33.25

Table 5.3: A comparison between proposed recommendation methods and the baselines on respiratory cohort

Method	**Precision**			**Recall**			**F score**		
	@3	**@5**	**@10**	**@3**	**@5**	**@10**	**@3**	**@5**	**@10**
TRoS	41.4	35.91	27.81	25.12	36.28	55.3	31.27	36.09	37.01
TRoT	41.4	36.09	27.95	25.13	36.49	55.71	31.27	36.29	37.22
TRoD	41.49	36.01	28.13	25.19	36.29	55.65	31.34	36.14	37.37
ICF + SGD	41.49	35.98	28.36	25.18	36.71	56.01	31.34	36.34	37.66
ICF + IALS	17.74	21.57	18.82	10.76	22.26	37.61	13.39	21.91	25.08
ICF + ADA	41.97	36.16	28.51	25.82	36.44	56.39	31.97	36.3	37.87

Table 5.4: A comparison between proposed recommendation methods and the baselines on septicemia cohort

to make treatment for new patients. Therefore, it is necessary to investigate the efficacy of them when being used with a small K. We consider the difference between the values of evaluation measures obtained by using a small K and by using K_{best}. Let $F1_{K_{best}}$ be the F1 score, obtained with K_{best}, $F1_K$ be the F1 score obtained with K. We compute Δ_K^F as the difference between these measures.

Method	**Precision**			**Recall**			**F score**		
	@3	**@5**	**@10**	**@3**	**@5**	**@10**	**@3**	**@5**	**@10**
TRoS	57.51	51.08	41.19	29.38	40.54	57.42	38.89	45.2	47.96
TRoT	57.75	51.36	41.33	29.8	40.93	57.89	39.31	45.55	48.22
TRoD	57.41	50.94	41.31	29.39	40.31	57.63	38.88	45.0	48.12
ICF + SGD	58.36	51.07	41.16	29.65	39.93	56.64	39.32	44.82	47.67
ICF + IALS	34.59	34.1	31.41	15.95	27.69	47.28	21.81	30.54	37.73
ICF + ADA	58.24	51.38	41.38	29.44	40.43	57.27	39.11	45.24	48.04

Table 5.5: A comparison between proposed recommendation methods and the baselines on myocardial infarction cohort

Method	**Respiratory**			**Septicemia**			**Myocardial Infarction**		
	@3	**@5**	**@10**	**@3**	**@5**	**@10**	**@3**	**@5**	**@10**
TRoS	200	200	200	200	200	200	200	200	200
TRoT	200	200	200	200	200	200	200	200	200
TRoD	150	100	200	200	200	200	60	100	200

Table 5.6: K values that yield the best F1 score on three datasets of recommendation methods over symptom-based, treatment-based and dual learning aspects

$$\Delta_K^F = F_{K_{best}} - F_K$$

Tables 5.7, 5.8, 5.9 report the Δ_K^F for $K = 7, K = 15, K = 50$ on three datasets. From the tables, we can see that the values of F1 for small Ks are considerably lower than those of the best cases. The difference is reduced but still remain substantially high when K is increased. With $K = 50$, the Δ_K^F is around ±1%

Method	**K = 7**			**K = 15**			**K = 50**		
	@3	**@5**	**@10**	**@3**	**@5**	**@10**	**@3**	**@5**	**@10**
TRoS	4.08	4.54	7.49	2.18	1.96	4.36	1.24	0.4	1.61
TRoD	4.1	3.96	8.23	2.36	1.78	3.43	1.06	0.48	1.24
TRoT	4.69	5.73	7.96	3.01	2.74	4.79	1.03	0.83	1.81

Table 5.7: Reported Δ_K^F of TRoS, TRoT, TRoD methods on respiratory cohort

Method	**K = 7**			**K = 15**			**K = 50**		
	@3	**@5**	**@10**	**@3**	**@5**	**@10**	**@3**	**@5**	**@10**
TRoS	7.0	6.88	6.73	3.45	3.86	3.34	1.7	1.52	1.14
TRoT	7.99	8.62	6.79	3.6	4.34	3.63	1.56	1.46	1.06
TRoD	7.44	6.91	6.58	3.43	3.63	2.63	1.36	1.33	0.82

Table 5.8: Reported Δ_K^F of TRoS, TRoT, TRoD methods on septicemia cohort

Method	**K = 7**			**K = 15**			**K = 50**		
	@3	**@5**	**@10**	**@3**	**@5**	**@10**	**@3**	**@5**	**@10**
TRoS	2.29	3.53	7.16	1.35	1.91	1.85	0.75	1.02	0.69
TRoT	2.65	3.61	7.67	1.08	1.69	1.97	0.35	0.28	0.56
TRoD	2.18	3.23	8.45	1.36	1.68	1.5	0.7	0.69	0.43

Table 5.9: Reported Δ_K^F of TRoS, TRoT, TRoD methods on myocardial infarction cohort

5.4 Discussion and Conclusion

In this chapter, we have introduced a simple K neighbor-based method to recommend treatment for a new patient. Our method takes treatment patterns of neighbor patients into account and selects the frequent prescription drugs as recommendation drugs. The proposed method mimics human intuition in capturing the treatment creation procedure where treatment for new patients is often derived from physicians' medical knowledge and their gained experiences through treating similar patients in the past. The idea of selecting frequent drugs is naturally based on the assumption that frequent prescription drugs among similar patients are essential drugs that will probably be given to new patients. Therefore, it can be said that our method seem to be more explainable in terms of healthcare perspective.

The comparison result between using the K-neighbor based methods and the nearest neighbor-based approach ($K = 1$) has pointed out the necessity of taking treatment of K similar patients into account. The inefficient of the nearest neighbor based treatment recommendation approach can be explained by complicated relationships between symptoms and drugs. Such complicated relationships can be the combination of multiple symptoms, the combination of multiple drugs, the many-to-many mapping of drug-symptom relations, i.e. one treatment could be treated by different drugs and vice verse. In addition, there are cases that treating a patient is performed without any knowledge on past cases since information about similar patients were not found in the database. As a result, it is more likely that a new patient is only similar to each other in several parts of his treatment. Consequently, combining treatment from multiple patients seems to be inevitable.

To capture the variety of treatment due to the complicated relationship between symptoms and drugs, we have explored the construction of treatment patterns of neighbor patients in either the treatment-based aspect or symptom-based aspect. We have also proposed a treatment recommendation method for each learning. The experimental results in Figure 5.6 have indicated that there is no learning aspect that obviously outperforms the other. This result can be the assumption that each aspect is a complementary view where both of them try to capture full mapping relation among drug-symptom relations.

We have also proposed an ensemble approach that recommends prescription drugs for new patients by combining treatment patterns of neighbor patients from both learning aspects. The dual recommendation method slightly outperforms the symptom-based and treatment-based recommendation method for large M. This result shows the promise to consider different aspects of solving the treatment recommendation problem. For small M, our dual method works unsteadily. We hypothesize that in such cases, the number of recommendation drugs is too small to make the prescription frequency of candidate drugs among neighbor treatment paths stable for applying our approach.

We have compared the baselines and the proposed methods with the best parameter K. The obtained results have shown that our domain-interpretable recommendation method is able to achieve comparable results to the state-of-the-art recommender systems designed for implicit feedback data. Since solving the recommendation problems remains very challenging, we stress on the need of developing domain interpretable neighbor-based methods that can work as good as black-boxed or complicated approaches. It can be seen that interpretable methods are important to convince physicians' faith in computer-aided systems applied to the healthcare domain.

Although our proposed methods can yield comparable results to the baselines, the number of required neighbors in the best cases are relatively large. This drawback again could be explained due to the complicated drug symptom relationship which can result in thousands of treatment patterns that difficult to learn on a limited subset of patients. In addition, the combination idea is applied to recommend multiple drugs that probably require more neighbor patients than predicting a single value by the classification or prediction task.

Chapter 6

Weighting Treatment Recommendation Method

6.1 Motivation

The previous chapter has demonstrated a simple idea to recommend treatment for new patients by selecting frequent prescription drugs among neighborhood treatments from either single or dual learning aspect. However, the efficacy of the proposed methods is only comparable with state-of-art-treatment recommendation models for implicit feedback dataset when being used with a large number of neighbor patients. This major drawback weakens the interpretation of treatment recommendation model considerably.

The above drawback could be attributed to a strong assumption that treatment of the neighbor patients and treatment of the considering new patient are similar. This ideal assumption may not be fit the reality. For example, hypertension patients can be treated by Angiotensin-converting enzyme (ACE) inhibitors including enalapril, lisinopril, perindopril and ramipril; or by Angiotensin-2 receptor blockers (ARBs) including candesartan, irbesartan, losartan, valsartan and olmesartan. Therefore, patients who are similar in terms of symptom-based features may not be similar in terms of treatment-based features. As a result, taking frequency of candidate drugs among K treatment paths seems to be over-optimistic for cohorts that can be treated in very different ways. In this chapter, we aim to develop a more sophisticated recommendation method that overcomes the above drawback.

6.2 Proposed Method

We derive a more deliberate mechanism to estimate the possibility of candidate drugs to be selected as recommendation drugs based on the following observation. Prescription drugs on the treatment paths of a neighbor patient p are given a high confidence score if they also appear in the treatments of training patients who have p as one of their neighbor patients. In other words, we re-utilize the known treatment paths of training patients to estimate the confidence of nodes of their neighbors. The following section describes in detail our weighting approach.

We split the training patients into several subsets where each subset is considered as a sub-testing set and the rests are sub-training set. For each patient in the sub-testing sets, we query his K_1 neighbors $p_1, p_2, ...p_{K_1}$ and their associated treatment paths $\delta^{p_1}, ..., \delta^{p_{K_1}}$. For each patient p_j in the sub-training set, let S^{p_j} be the set of patients who have p_j as one of their K_1 neighbors. We calculate a hitting-score $hit_{dr}^{\delta^{p_j}}$ for each drug dr on the treatment path δ^{p_j} of training patient p_j as follows.

$$hit_{dr}^{\delta^{p_j}} = \sum_{p_k \in S^{p_j}} d_{p^k}^{p_j} \times I(dr \in \delta^{p_k})$$

In the above formula, every time drug dr was used to treat a patient p_k in S^{p_j}, we add to the hitting score $hit_{dr}^{\delta^{p_j}}$ a reward equal to the distance $d_{p^k}^{p_j}$. The meaning is that when p_j and p_k are far neighbors and dr has been found in the treatment of p_k, it is added more weight than the closer neighbors as a compensation for the possibility of "incorrectly" identifying close neighbors. The term "incorrectly" means those far neighbors who are considerably different in terms of non-treatment-based features are similar in terms of treatment-based features. In case p_j and p_k are close neighbors and dr has been found in the treatment of p_k, we add a relatively small award

Algorithm 7: Recommending prescription drugs for new patient p in a specific period using weighting approach

```
Data: $\Lambda$, $\theta$, $v^p$, $H^{train}$
Result: return top M recommended drugs
Randomly split training set into sub-training and sub-testing sets ;
Initialize all nodes in the prescription trees with 0 hitting score;
for each pair (sub-training , sub-testing) do
    for each $p$ in the sub-testing do
        Select $K'$ most similar patients $p_1, p_2, ..., p_{K'}$ among sub-training patients ;
        Trace associated treatment paths $\gamma^{p_1}, \gamma^{p_2}, ..., \delta^{p_{K'}}$ through tracing variables $\Lambda$ ;
        for each $\gamma^{p_i}$ do
            for each $dr$ in $\gamma^{p_i}$ do
                If $p$ was treated with $dr$  $hit_{dr}^{\delta^{p_j}} = hit_{dr}^{\delta^{p_j}} + d_{p^i}^{p}$
Compute $h^p = P(h|v^p, \theta)$;
Compute similarity between $h^p$ and each training patient in $H^{train}$;
Select $K$ most similar patients $p_1, p_2, ..., p_K$;
Trace associated treatment paths $\gamma^{p_1}, \gamma^{p_2}, ..., \gamma^{p_K}$ through tracing variables $\Lambda$ ;
$C = \bigcup_{i=1}^{K} \gamma^{p_i}$ ;
for each $dr \in C$ do
    Compute $\overline{hit_{dr}^{p}}$ by equation (6.1);
Return top $M$ drugs sorted by $\overline{hit_{dr}^{p}}$;
```

equal to their distance to the hitting score since there is a high possibility dr can be found in the treatment of p_k.

After calculating the hitting score for all nodes in the prescription trees, we perform the procedure for ranking recommendation drugs for testing patient p. For each candidate drug dr in the set C, we compute an average hitting score $\overline{hit_{dr}^{p}}$ weighted by the distances from the test patient to neighbors whose treatment paths include dr. The formula of $\overline{hit_{dr}^{p}}$ is given below.

$$\overline{hit^p_{dr}} = \frac{1}{\sum_{i=1}^{K} I(dr \in \delta^{p_i})} \sum_{i=1}^{K} I(dr \in \delta^{p_i}) \times hit^{\delta^{p_i}}_{dr} \times d^{p}_{p^i} \tag{6.1}$$

Algorithm 7 summarizes the main steps of the treatment recommendation method using the weighting approach. It is worth noting that the proposed weighting treatment recommendation method follows a similar workflow of data processing, patient clustering, prescription tree construction steps as described in the previous chapters. Therefore, in this chapter, we only present a new mechanism to combine treatment paths of neighbor patients.

6.3 Experimental Evaluation

This section presents our experimental evaluation for the proposed weighting treatment recommendation method. We use the symptom-based learning aspect to illustrate the efficacy of the proposed model. We compare the weighting treatment recommendation method on symptom learning aspect named WTRoS with the TRoS and the baselines ICF + SGD, ICF + IALS, ICF + ADA that have been mentioned in the previous chapters.

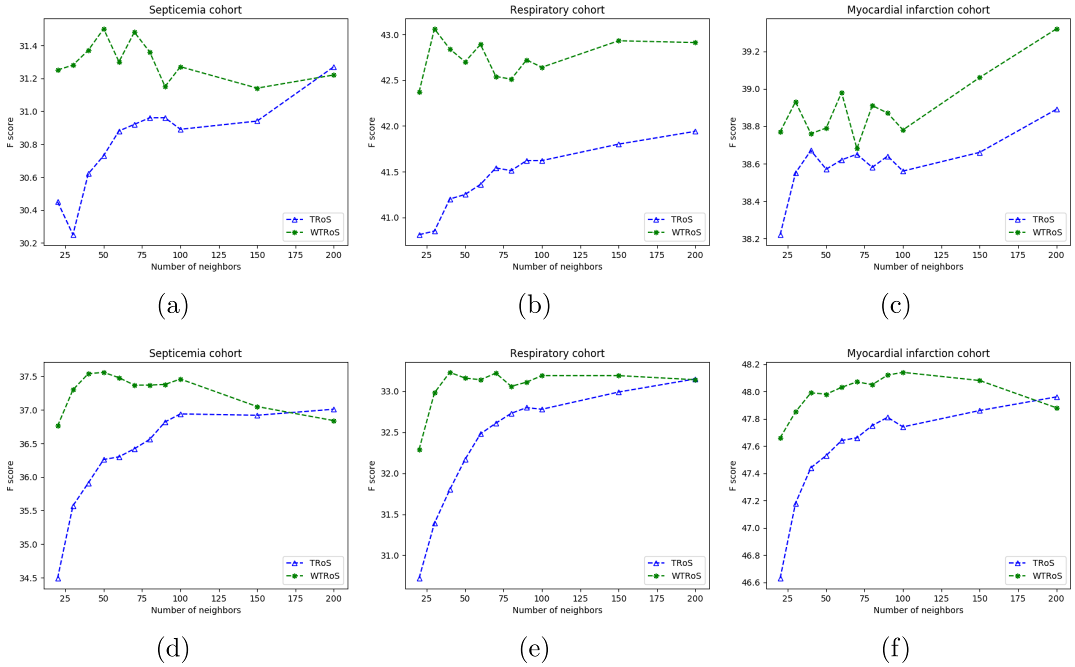
6.3.1 Dataset and parameter setting

We used the same three datasets described in the treatment learning chapter. For the weighting approach, we split the training set into five subsets and then learn the hitting score of nodes in the treatment paths of training patients. The number of sub training neighbors K_1 is varied with 50, 100, 150, 200 neighbors. We found that it is better to select a relatively large number of K_1. We report the result with $K_1 = 100, \epsilon = 5$. Most of the other relevant parameters were set similarly to the previous chapter.

6.3.2 A comparison to the non-weighting approaches

First, we compare the efficacy of WTRoS and TRoS. Figure 6.1 shows the F1 score obtained by two methods when recommending top 3 drugs and top

10 drugs on three datasets.



(a) (b) (c)

(d) (e) (f)

Figure 6.1: A comparison on the efficacy of treatment recommendation task in two cases: using the non-weighting recommendation method (TRoS) and using the weighting recommendation method (WTRoS) over the symptom learning aspect

We varied the number of neighbors from 20 to 200 neighbors. It can be seen that the WTRoS outperforms the TRoS in most of the cases over three datasets. This shows the effectiveness of the proposed weighting approach.

6.3.3 A comparison to the baselines

We next compare the efficacy between WTRoS, TRoS and the baselines in best cases. Similarly to the previous chapter, we report the result with the choice of K_{best} that returns the best F1 score corresponding to each M and report the three evaluation measures on the test set with parameters K_{best}. Tables 6.1, 6.2, 6.3 show the obtained results on the septicemia, respiratory

Method	**Precision**			**Recall**			**F score**		
	@3	**@5**	**@10**	**@3**	**@5**	**@10**	**@3**	**@5**	**@10**
TRoS	41.33	33.06	21.62	42.58	56.52	71.07	41.94	41.71	33.15
WTRoS	42.15	33.59	21.68	44.02	57.44	71.19	43.06	42.38	33.23
ICF + SGD	42.13	33.4	21.75	43.83	56.68	71.16	42.95	42.03	33.31
ICF + IALS	20.1	23.19	17.96	19.33	39.0	59.89	19.7	29.06	27.62
ICF + ADA	42.15	33.44	21.72	43.48	56.78	70.9	42.8	42.08	33.25

Table 6.1: A comparison between the proposed weighting recommendation method with the non-weighting one and the baselines over respiratory cohort

Method	**Precision**			**Recall**			**F score**		
	@3	**@5**	**@10**	**@3**	**@5**	**@10**	**@3**	**@5**	**@10**
TRoS	41.4	35.91	27.81	25.12	36.28	55.3	31.27	36.09	37.01
WTRoS	41.48	36.22	28.26	25.39	36.49	55.99	31.5	36.35	37.56
ICF + SGD	41.49	35.98	28.36	25.18	36.71	56.01	31.34	36.34	37.66
ICF + IALS	17.74	21.57	18.82	10.76	22.26	37.61	13.39	21.91	25.08
ICF + ADA	41.97	36.16	28.51	25.82	36.44	56.39	31.97	36.3	37.87

Table 6.2: A comparison between the proposed weighting recommendation method with the non-weighting one and the baselines over septicemia cohort

and myocardial infarction cohorts, respectively. From the three tables, we can see that the WTRoS achieves competitive results to the baselines and substantially better than the TRoS in most cases. This observation partially shows the effectiveness of our weighting approach for recommendation task in comparison to the non-weighting approach.

Table 6.4 reports the values of K that yields the best F1 score. Compared to the TRoS, the WTRoS requires fewer neighbors to achieve the best F1 score in many cases. For example, we only need around 30 neighbors to obtain the highest F1 score in recommending top 3 prescription drugs on the

Method	**Precision**			**Recall**			**F score**		
	@3	**@5**	**@10**	**@3**	**@5**	**@10**	**@3**	**@5**	**@10**
TRoS	57.51	51.08	41.19	29.38	40.54	57.42	38.89	45.2	47.96
WTRoS	57.92	51.12	41.36	29.76	40.33	57.61	39.32	45.09	48.14
ICF + SGD	58.36	51.07	41.16	29.65	39.93	56.64	39.32	44.82	47.67
ICF + IALS	34.59	34.1	31.41	15.95	27.69	47.28	21.81	30.54	37.73
ICF + ADA	58.24	51.38	41.38	29.44	40.43	57.27	39.11	45.24	48.04

Table 6.3: A comparison between the proposed weighting recommendation method with the non-weighting one and the baselines over myocardial infarction cohort

Method	**Respiratory**			**Septicemia**			**Myocardial Infarction**		
	@3	**@5**	**@10**	**@3**	**@5**	**@10**	**@3**	**@5**	**@10**
TRoS	200	200	200	200	200	200	200	200	200
WTRoS	30	200	40	50	50	50	200	50	100

Table 6.4: K values that yield the best F score on three datasets of recommendation method using weighting (WTRoS) and non-weighting approach (TRoS)

respiratory cohort.

6.3.4 How good when using the small K

This section provides an investigation of the behavior of the proposed method in small K. We use the same evaluation measure to characterize the difference between F1 scores in the best K and the small K. Tables 6.5, 6.6, 6.7 show the Δ_K^F of the WTRoS and TRoS in comparison to the best cases. It can be seen that the WTRoS can reduce the delta on each evaluation measure considerably in comparison to the TRoS. The gap of the F1 score in the small K and the best case has been reduced considerably.

Method	**K = 7**			**K = 15**			**K = 50**		
	@3	**@5**	**@10**	**@3**	**@5**	**@10**	**@3**	**@5**	**@10**
TRoS	7.11	7.18	6.73	3.56	4.16	3.34	1.81	1.82	1.14
WTRoS	1.84	3.85	6.6	0.79	1.1	2.15	0.36	0.24	0.15

Table 6.5: Reported Δ_K^F of WTRoS and TRoS methods on respiratory cohort

Method	**K = 7**			**K = 15**			**K = 50**		
	@3	**@5**	**@10**	**@3**	**@5**	**@10**	**@3**	**@5**	**@10**
TRoS	4.08	4.55	7.49	2.18	1.97	4.36	1.24	0.41	1.61
WTRoS	1.41	1.29	6.07	0.87	0.28	1.64	0.47	0.0	0.31

Table 6.6: Reported Δ_K^F of WTRoS and TRoS methods on septicemia cohort

Method	**K = 7**			**K = 15**			**K = 50**		
	@3	**@5**	**@10**	**@3**	**@5**	**@10**	**@3**	**@5**	**@10**
TRoS	2.29	3.22	7.08	1.35	1.6	1.77	0.75	0.71	0.61
WTRoS	0.7	0.8	3.27	0.78	0.2	0.51	0.53	0.15	0.16

Table 6.7: Reported Δ_K^F of WTRoS and TRoS methods on myocardial infarction cohort

6.4 Discussion and Conclusion

This chapter has presented a weighting method that aims to overcome the drawback of the non-weighting one presented in the previous chapter. Our method addresses the issue where unreliable neighbors whose treatment may be very different from the treatment of new patients even they are similar in terms of symptom-based features.

To tackle this, we consider neighbor patients among training patients as test data to estimate the confidence of the drugs on the treatment paths of the training patients. The improvement of obtained F1 scores in Figure 6.1, the competitive results of the WTRoS in comparison to the baselines and the fewer neighbor needed for the best case have shown the superiority of our weighting strategy to the non-weighting one. It can be considered as a "self-correction" strategy where a larger amount of weight is given to nodes that appear in far training neighbor patients' treatment paths to address the problem of possibly wrong identification of neighbor patients. We note that the low obtained precision in all three datasets indicates that there are many neighbor patients who have been identified "incorrectly". As a result, this characteristic of the datasets seems to fit the weighting method, and therefore lead to better results in terms of both efficacy and interpretability.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

In this dissertation, we have introduced domain-based treatment learning methods and treatment recommendation methods that try to incorporate medical domain knowledge and provide interpretable data-driven methods for healthcare problems. The main findings of our work are summarized as follows.

Chapter 4 has proposed a treatment learning method that aims to derive treatment patterns of patient groups. First, we address the challenge in dealing with heterogeneous and longitudinal EMR objects. In concrete, we proposed adopting a mixed variate restricted Boltzmann machine for representing different types of patient records. Our method is more generic in terms of data utilization than most of the current studies in the literature that merely used a limited subset of patient features. To address the challenge of handling longitudinal prescription records, we proposed a scoring algorithm which adopted medical domain information to split patient records into periods automatically. Our scoring algorithm reflects significant changes in prescription indication and seems to more flexible than fixed interval treatment periods often used in the literature.

Second, we have proposed an indication labeling framework which is able to reveal signs or symptoms of a set of diseases, and drugs curing these signs, symptoms. The framework illustrates how we use medical domain sources for information extraction task. It is useful to grasp about diseases and treatments quickly. In addition, the indication labeling is helpful for identifying drug indication, an important component to measure the significant change that probably indicates a new treatment period stage in prescription records. More interestingly, drugs with labeled indication can help understand to some extent what symptoms or diseases are underlying treatment patterns of each patient group and therefore, help to understand the characteristics of each patient group.

Third, in this chapter, we have also suggested an alternative way to organize drug frequency of each patient group in a tree form. This kind of knowledge representation can not only reveal the sequence of frequent prescription drugs but also allow identifying drugs that are frequently or infrequently prescribed given a set of other prescription drugs. In other words, we propose a more flexible way that derives different types of treatment patterns in comparison to the conventional approaches using association analysis.

Chapter 5 has presented neighbor-based methods which recommend top M prescription drugs over treatment periods for new patients. The key idea of the methods is to take into account the typical prescription drugs of neighbor patients' treatments to suggest drugs for new patients. To capture as many as treatment variation caused by the complicated drug-disease mapping, we have proposed exploring different ways to find out the typical drugs of neighbor patients under treatment-based learning aspect or symptom-based learning aspect. The recommendation mechanism could be done via each of the above learning aspect or both of them. Experimental results have shown the superiority of the proposed K-neighbor-based recommendation methods to the nearest neighbor-based approach. In best cases, our neighbor-based methods are able to yield similar results but more promising in terms of interpretability compared to the baselines. The dual recommendation method has shown to be effective in recommending a large number of drugs. This result shows that the consideration of synthesizing different learning aspects is promising to address the treatment recommendation problem.

Chapter 6 has provided a weighting recommendation mechanism which partially addresses the issue of inconsistent similarity of symptom-based and treatment-based features. Different from the recommendation methods proposed in the previous chapter where recommendation drugs are ranked according to their appearance frequency among neighbor patients' treatment paths, the weighting method estimates the confidence of each drug in training patients through neighbor patients' treatment paths among training set itself. The experimental results have pointed out the effectiveness of the weighting method in terms of evaluation measures and interpretability. It is able to yield competitive results to the baselines with fewer neighbors in comparison to the non-weighting method. This result has shown there are plenty of rooms to develop different strategies for solving the treatment recommendation problem using neighbor-based approach.

7.2 Future Work

We have proposed various methods for solving the treatment recommendation problem. However, the precision of the proposed methods is still quite low. Based on the results and analysis of the pros and cons of the proposed recommendation methods, we suggest the following research directions that could be considered further to improve the treatment recommendation efficacy.

First, although we extracted sign/symptom features to represent patients, we hypothesize that many among them could be relevant or close to other features. As a result, similarity measures designed for binary features seem to unfit with highly relevant symptom features. To resolve this issue, it would be necessary to identify sets of relevant features and propose a new similarity for highly correlated features. The first challenge can be addressed by word embedding techniques that represent each extracted symptom term as a vector in the clinical context and measure the similarity among the terms, i.e the symptom attributes. We then need to develop a new similarity between two patient vectors given the similarity of attributes. In this case, the weighting approach for similarity measures [Mihalcea et al., 2006, Candillier et al., 2008, Luo et al., 2011, Matsuo and Ho, 2018] could be useful to reweigh the similarity more appropriately for the clinical context.

Second, there is a complicated relationship between treatment view and symptom view. Patients who are similar to others for only a few symptoms could be treated in a very similar way. Therefore, it may be useful to utilize the similarity in treatment view of training patients to adjust the similarity in symptom view. A high-quality patient subcohort at that time is defined as a group of patients who are similar in both symptom and treatment-based features.

Third, the use of initial symptom could limit capturing the change in prescription according to patient health status. A dynamic modle that captures sequences of treatment and patient phenotype would be more practical to improve the recommendation efficacy. The challenge of this approach is one has to recognize patient phenotypes after each period of drug use. Though this is a very challenging task, it could be addressed by sentiment analysis methods for medical text which have been developed recently [Thelwall et al., 2010, Deng et al., 2014, Bui and Zeng-Treitler, 2014, del Arco et al., 2016].

List of Publication

International Conference Papers

- Hoang K.H., Ho T.B. (2018) Learning Treatment Regimens from Electronic Medical Records. In: Phung D., Tseng V., Webb G., Ho B., Ganji M., Rashidi L. (eds) *Advances in Knowledge Discovery and Data Mining. PAKDD 2018.* Lecture Notes in Computer Science, vol 10937. Springer, Cham.

Journal Papers

- Hoang K.H., Ho T.B. (2019) Learning and Recommending Treatments Using Electronic Medical Reords. *Journal of Knowledge-based Systems* (accepted).

Bibliography

[Association et al., 2008] Association, B. M. et al. (2008). *Withholding and withdrawing life-prolonging medical treatment: guidance for decision making*. John Wiley & Sons.

[Blumstein, 1980] Blumstein, J. F. (1980). Rationing medical resources: A constitutional, legal, and policy analysis. *Tex. L. Rev.*, 59:1345.

[Bottou, 2012] Bottou, L. (2012). Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, pages 421–436. Springer.

[Bouarfa and Dankelman, 2012] Bouarfa, L. and Dankelman, J. (2012). Workflow mining and outlier detection from clinical activity logs. *Journal of biomedical informatics*, 45(6):1185–1190.

[Bui and Zeng-Treitler, 2014] Bui, D. D. A. and Zeng-Treitler, Q. (2014). Learning regular expressions for clinical text classification. *Journal of the American Medical Informatics Association*, 21(5):850–857.

[Candillier et al., 2008] Candillier, L., Meyer, F., and Fessant, F. (2008). Designing specific weighted similarity measures to improve collaborative filtering systems. In *Industrial Conference on Data Mining*, pages 242–255. Springer.

[Char et al., 2018] Char, D. S., Shah, N. H., and Magnus, D. (2018). Implementing machine learning in health careaddressing ethical challenges. *The New England journal of medicine*, 378(11):981.

[Chen et al., 2018] Chen, J., Sun, L., Guo, C., Wei, W., and Xie, Y. (2018). A data-driven framework of typical treatment process extraction and evaluation. *Journal of biomedical informatics*, 83:178–195.

[Chen et al., 2015] Chen, Y., Xie, W., Gunter, C. A., Liebovitz, D., Mehrotra, S., Zhang, H., and Malin, B. (2015). Inferring clinical workflow efficiency via electronic medical record utilization. In *AMIA annual symposium proceedings*, volume 2015, page 416. American Medical Informatics Association.

[del Arco et al., 2016] del Arco, F. M. P., Valdivia, M. T. M., Zafra, S. M. J., González, M. D. M., and Cámara, E. M. (2016). Copos: corpus of patient opinions in spanish. application of sentiment analysis techniques. *Procesamiento del Lenguaje Natural*, 57:83–90.

[Deng et al., 2014] Deng, Y., Stoehr, M., and Denecke, K. (2014). Retrieving attitudes: Sentiment analysis from clinical narratives. In *MedIR@ SIGIR*, pages 12–15.

[Elghazel et al., 2007] Elghazel, H., Deslandres, V., Kallel, K., and Dussauchoy, A. (2007). Clinical pathway analysis using graph-based approach and markov models. In *Digital Information Management, 2007. ICDIM'07. 2nd International Conference on*, volume 1, pages 279–284. IEEE.

[Fischer and Igel, 2012] Fischer, A. and Igel, C. (2012). An introduction to restricted boltzmann machines. In *Iberoamerican Congress on Pattern Recognition*, pages 14–36. Springer.

[Gräßer et al., 2017] Gräßer, F., Beckert, S., Küster, D., Schmitt, J., Abraham, S., Malberg, H., and Zaunseder, S. (2017). Therapy decision support based on recommender system methods. *Journal of healthcare engineering*, 2017.

[Group et al., 2009] Group, K. D. I. G. O. K. C.-M. W. et al. (2009). Kdigo clinical practice guideline for the diagnosis, evaluation, prevention, and treatment of chronic kidney disease-mineral and bone disorder (ckd-mbd). *Kidney international. Supplement*, (113):S1.

[Healy et al., 1998] Healy, W. L., Ayers, M. E., Iorio, R., Patch, D. A., Appleby, D., and Pfeifer, B. A. (1998). Impact of a clinical pathway and implant standardization on total hip arthroplasty: a clinical and economic study of short-term patient outcome. *The Journal of arthroplasty*, 13(3):266–276.

[Hinton, 2012] Hinton, G. E. (2012). A practical guide to training restricted boltzmann machines. In *Neural networks: Tricks of the trade*, pages 599–619. Springer.

[Hirano and Tsumoto, 2014] Hirano, S. and Tsumoto, S. (2014). Mining typical order sequences from ehr for building clinical pathways. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 39–49. Springer.

[Hu et al., 2008] Hu, Y., Koren, Y., and Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 263–272. Ieee.

[Huang et al., 2015] Huang, Z., Dong, W., Bath, P., Ji, L., and Duan, H. (2015). On mining latent treatment patterns from electronic medical records. *Data Mining and Knowledge Discovery*, 29(4):914–949.

[Huang et al., 2012] Huang, Z., Lu, X., and Duan, H. (2012). On mining clinical pathway patterns from medical behaviors. *Artificial intelligence in medicine*, 56(1):35–50.

[Ireson, 1997] Ireson, C. L. (1997). Critical pathways: effectiveness in achieving patient outcomes. *Journal of Nursing Administration*, 27(6):16–23.

[Jensen et al., 2012] Jensen, P. B., Jensen, L. J., and Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405.

[Jin et al., 2018] Jin, B., Yang, H., Sun, L., Liu, C., Qu, Y., and Tong, J. (2018). A treatment engine by predicting next-period prescriptions. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1608–1616. ACM.

[Johnson et al., 2016] Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3.

[Jutel, 2009] Jutel, A. (2009). Sociology of diagnosis: a preliminary review. *Sociology of health & illness*, 31(2):278–299.

[Katzman et al., 2018] Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. (2018). Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):24.

[Kingma and Ba, 2014] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[Koren et al., 2009] Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, (8):30–37.

[Lazarou et al., 1998] Lazarou, J., Pomeranz, B. H., and Corey, P. N. (1998). Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *Jama*, 279(15):1200–1205.

[Le et al., 2018] Le, H., Tran, T., and Venkatesh, S. (2018). Dual control memory augmented neural networks for treatment recommendations. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 273–284. Springer.

[Le Roux and Bengio, 2008] Le Roux, N. and Bengio, Y. (2008). Representational power of restricted boltzmann machines and deep belief networks. *Neural computation*, 20(6):1631–1649.

[Liao and Ahn, 2016] Liao, L. and Ahn, H.-i. (2016). Combining deep learning and survival analysis for asset health management. *International Journal of Prognostics and Health Management*.

[Lin et al., 2001] Lin, F.-r., Chou, S.-c., Pan, S.-m., and Chen, Y.-m. (2001). Mining time dependency patterns in clinical pathways. *International Journal of Medical Informatics*, 62(1):11–25.

[Liu et al., 2017] Liu, Y., Logan, B., Liu, N., Xu, Z., Tang, J., and Wang, Y. (2017). Deep reinforcement learning for dynamic treatment regimes on medical registry data. In *Healthcare Informatics (ICHI), 2017 IEEE International Conference on*, pages 380–385. IEEE.

[Lu et al., 2016] Lu, H.-M., Wei, C.-P., and Hsiao, F.-Y. (2016). Modeling healthcare data using multiple-channel latent dirichlet allocation. *Journal of biomedical informatics*, 60:210–223.

[Luo et al., 2011] Luo, Q., Chen, E., and Xiong, H. (2011). A semantic term weighting scheme for text categorization. *Expert Systems with Applications*, 38(10):12708–12716.

[Ma et al., 2017] Ma, F., Chitta, R., Zhou, J., You, Q., Sun, T., and Gao, J. (2017). Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1903–1911. ACM.

[Matsuo and Ho, 2018] Matsuo, R. and Ho, T. B. (2018). Semantic term weighting for clinical texts. *Expert Systems with Applications*, 114:543–551.

[Mei et al., 2015] Mei, J., Liu, H., Li, X., Xie, G., and Yu, Y. (2015). A decision fusion framework for treatment recommendation systems. *Studies in health technology and informatics*, 216:300.

[Mihalcea et al., 2006] Mihalcea, R., Corley, C., Strapparava, C., et al. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, volume 6, pages 775–780.

[Murphy, 2003] Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355.

[Nemati et al., 2016] Nemati, S., Ghassemi, M. M., and Clifford, G. D. (2016). Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach. In *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*, pages 2978–2981. IEEE.

[Newdick et al., 2005] Newdick, C. et al. (2005). Who should we treat?: rights, rationing, and resources in the nhs. *OUP Catalogue*.

[Nezhad et al., 2019] Nezhad, M. Z., Sadati, N., Yang, K., and Zhu, D. (2019). A deep active survival analysis approach for precision treatment recommendations: Application of prostate cancer. *Expert Systems with Applications*, 115:16–26.

[Ordonez, 2006] Ordonez, C. (2006). Association rule discovery with the train and test approach for heart disease prediction. *IEEE Transactions on Information Technology in Biomedicine*, 10(2):334–343.

[Palaniappan and Awang, 2008] Palaniappan, S. and Awang, R. (2008). Intelligent heart disease prediction system using data mining techniques. In *Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on*, pages 108–115. IEEE.

[Park et al., 2017] Park, S., Choi, D., Kim, M., Cha, W., Kim, C., and Moon, I.-C. (2017). Identifying prescription patterns with a topic model of diseases and medications. *Journal of biomedical informatics*, 75:35–47.

[Pattekari and Parveen, 2012] Pattekari, S. A. and Parveen, A. (2012). Prediction system for heart disease using naïve bayes. *International Journal of Advanced Computer and Mathematical Sciences*, 3(3):290–294.

[Pham et al., 2016] Pham, T., Tran, T., Phung, D., and Venkatesh, S. (2016). Deepcare: A deep dynamic memory model for predictive medicine. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 30–41. Springer.

[Pham et al., 2017] Pham, T., Tran, T., Phung, D., and Venkatesh, S. (2017). Predicting healthcare trajectories from medical records: A deep learning approach. *Journal of biomedical informatics*, 69:218–229.

[Prasad et al., 2017] Prasad, N., Cheng, L.-F., Chivers, C., Draugelis, M., and Engelhardt, B. E. (2017). A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. *arXiv preprint arXiv:1704.06300*.

[Rendle, 2010] Rendle, S. (2010). Factorization machines. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 995–1000. IEEE.

[Rushing, 1971] Rushing, W. A. (1971). Public policy, community constraints, and the distribution of medical resources. *Soc. Probs.*, 19:21.

[Savova et al., 2010] Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., and Chute, C. G. (2010). Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

[Shang et al., 2018] Shang, J., Hong, S., Zhou, Y., Wu, M., and Li, H. (2018). Knowledge guided multi-instance multi-label learning via neural networks in medicines prediction. In *Asian Conference on Machine Learning*, pages 831–846.

[Snow et al., 1994] Snow, P. B., Smith, D. S., and Catalona, W. J. (1994). Artificial neural networks in the diagnosis and prognosis of prostate cancer: a pilot study. *The Journal of urology*, 152(5):1923–1926.

[Soni et al., 2011] Soni, J., Ansari, U., Sharma, D., and Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, 17(8):43–48.

[Sun et al., 2016] Sun, L., Liu, C., Guo, C., Xiong, H., and Xie, Y. (2016). Data-driven automatic treatment regimen development and recommendation. In *KDD*, pages 1865–1874.

[Tamasauskas et al., 2012] Tamasauskas, D., Sakalauskas, V., and Kriksciuniene, D. (2012). Evaluation framework of hierarchical clustering methods for binary data. In *Hybrid Intelligent Systems (HIS), 2012 12th International Conference on*, pages 421–426. IEEE.

[Thelwall et al., 2010] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558.

[Tran et al., 2014] Tran, T., Phung, D., and Venkatesh, S. (2014). Mixed-variate restricted boltzmann machines. *arXiv preprint arXiv:1408.1160*.

[Ubel et al., 1996] Ubel, P. A., DeKay, M. L., Baron, J., and Asch, D. A. (1996). Cost-effectiveness analysis in a setting of budget constraintsis it equitable? *New England Journal of Medicine*, 334(18):1174–1177.

[Wada et al., 2013] Wada, H., Thachil, J., Di Nisio, M., Mathew, P., Kurosawa, S., Gando, S., Kim, H., Nielsen, J., Dempfle, C.-E., Levi, M., et al. (2013). Guidance for diagnosis and treatment of disseminated intravascular coagulation from harmonization of the recommendations from three guidelines. *Journal of thrombosis and haemostasis*, 11(4):761–767.

[Wang et al., 2018] Wang, L., Zhang, W., He, X., and Zha, H. (2018). Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2447–2456. ACM.

[Weng et al., 2017] Weng, W.-H., Gao, M., He, Z., Yan, S., and Szolovits, P. (2017). Representation and reinforcement learning for personalized glycemic control in septic patients. *arXiv preprint arXiv:1712.00654*.

[Wilson and Evans, 1993] Wilson, P. W. and Evans, J. C. (1993). Coronary artery disease prediction. *American journal of hypertension*, 6(11_Pt_2):309S–313S.

[Xu et al., 2016] Xu, X., Jin, T., Wei, Z., Lv, C., and Wang, J. (2016). Tcpm: topic-based clinical pathway mining. In *Connected Health: Applications, Systems and Engineering Technologies (CHASE), 2016 IEEE First International Conference on*, pages 292–301. IEEE.

[Yao et al., 2018] Yao, L., Zhang, Y., Wei, B., Zhang, W., and Jin, Z. (2018). A topic modeling approach for traditional chinese medicine prescriptions. *IEEE Transactions on Knowledge and Data Engineering*, 30(6):1007–1021.

[Zhang et al., 2017] Zhang, Y., Chen, R., Tang, J., Stewart, W. F., and Sun, J. (2017). Leap: learning to prescribe effective and safe treatment combinations for multimorbidity. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1315–1324. ACM.

[Zhao et al., 2011] Zhao, Y., Zeng, D., Socinski, M. A., and Kosorok, M. R. (2011). Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics*, 67(4):1422–1433.