INTEGRATING HETEROGENEOUS DATA INTO ELECTRONIC MEDICAL RECORD ANALYSIS

BY

EDWARD W HUANG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Doctoral Committee:

Professor ChengXiang Zhai, Chair
Professor Roy Campbell
Assistant Professor Farzad Farnoud, University of Virginia
Assistant Professor Jian Peng
Professor Saurabh Sinha

ABSTRACT

Electronic medical records (EMRs) are the digital equivalent of paper records at a clinician's office. They contain patient information such as treatment and medical history, and have been shown to have a wide variety of benefits.

However, EMRs typically contain a multitude of diverse data, including images, doctor notes, medical test results, and genomic data. This heterogeneity generates high dimensionality and data sparsity, which are two of the most prevalent culprits that exacerbate already difficult computational problems. Additionally, domain-specific characteristics, such as the existence of synonyms in the medical vocabulary, introduce ambiguity. This can further reduce the data mining potential of EMRs.

This thesis is a systematic study that addresses these issues associated with EMRs. In particular, I utilized heterogeneous data sources that are typically incompatible, and then developed frameworks in which these data sources complement one another. As a result, these methods have the potential for direct clinical translation, paving the way for improving healthcare from a data-driven perspective.

To improve a variety of downstream healthcare applications, such as patient subcategorization, survival analysis, and visualization, I used external networks of domain knowledge consisting of drug-symptom relationships, protein-protein interactions, and genetic information to enhance patient records. I found that this enhancement process increased the data mining capabilities as well as the interpretability of the EMRs.

To improve EMR retrieval systems, I developed a query expansion method that frames symptoms and treatments as two different languages. I found that a topic modeling method that follows this dual-language framework yielded the highest performance. Lastly, I showed that due to pathological similarities, jointly studying Alzheimer's disease and Parkinson's disease resulted in higher computational power by effectively increasing the size of the training datasets. This allowed for the accurate prediction of the onset of dementia in both diseases.

Each of these results can lay the groundwork for applications that have the potential to be implemented directly in clinical practice, improving the safety and quality of patient care.

To my family.

ACKNOWLEDGMENTS

I would first like to thank my advisor, Professor ChengXiang Zhai, for his guidance in both research projects and in life and career advice. Our research was always a conversation between us rather than a dictation. He valued and considered my voice early in my graduate studies, which was largely responsible for my ability to perform independent research. Furthermore, Professor Zhai's capacity to always see the bigger picture allowed me to shape my career path into something of which I am greatly proud. For all of his kindness and patience, I express my deepest gratitude.

I would also like to thank my committee members. In alphabetical order, they are Professor Roy Campbell, Professor Farzad Farnoud, Professor Jian Peng, and Professor Saurabh Sinha. Professor Roy Campbell gave me the opportunity to work on one of my favorite projects through his course. I've learned much from his attention to detail and his enthusiasm for rigorous and high-impact healthcare research. Professor Farzad Farnoud was my mentor at Caltech, and our close collaboration was what allowed me to discover my interest in information retrieval and data mining. His sense of humor is matched only by his incredible work ethic. Professor Peng introduced me to the technique of network embeddings, the basis for much of my thesis. His peerless domain expertise, research suggestions, and kind words of encouragement have been invaluable. Professor Sinha was the first person to introduce me to bioinformatics and computational biology, and his knowledge and confidence have been a source of inspiration ever since. I always knew that I could go to him for an enlightening answer to any question.

There are many others who I need to thank for all the help I've received throughout the years for their contributions leading up to my thesis. Professor Fu-Ming Tao gave me my first research opportunity at California State University, Fullerton. At Caltech, I met Professor Mani Chandy and Professor Shuki Bruck, who were amazing mentors. They helped me narrow my focus in computer science research and always went above and beyond when helping me with my career. I would also like to thank Dr. Sheng Wang for his friendship and mentorship while we were students together. Lastly, I would like to thank the National Science Foundation Graduate Research Fellowship Program for providing funding.

Finally, I want to thank my family for their unwavering support. To my brother, Efrem, who I have looked up to in every aspect of life since the day I was born. To my parents, there is nothing I can say in words, no story I could tell, that could begin to express my gratitude for everything that they have done for me. This thesis is for them.

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION 1

CHAPTER 2 PATIENT RECORD SUBCATEGORIZATION FOR PRECISION TRADITIONAL CHINESE MEDICINE 3
2.1 Problem Definition 5
2.2 Methods 7
2.3 Related Work 10
2.4 Dataset Description 11
2.5 Experimental Design 12
2.6 Applications of PaReCat 14
2.7 Conclusions and Future Work 19

CHAPTER 3 INTEGRATION OF ELECTRONIC MEDICAL RECORDS WITH MOLECULAR INTERACTION NETWORKS AND DOMAIN KNOWLEDGE FOR SURVIVAL ANALYSIS 21
3.1 HEMnet Overview 22
3.2 Related Work 25
3.3 Dataset Description 26
3.4 Experimental Design 27
3.5 Results and Discussion 29
3.6 Conclusions and Future Work 34

CHAPTER 4 INTEGRATING EXTERNAL KNOWLEDGE INTO ELECTRONIC MEDICAL RECORD VISUALIZATION 37
4.1 Dataset Description 38
4.2 Patient Profile Matrix 38
4.3 VisAGE Overview 38
4.4 Related Work 41
4.5 Results and Discussion 41
4.6 Conclusions and Future Work 47

CHAPTER 5 FRAMING ELECTRONIC MEDICAL RECORDS AS POLYLINGUAL DOCUMENTS IN QUERY EXPANSION 49
5.1 Retrieval Problem Formulation 50
5.2 Methods 51
5.3 Evaluation 55
5.4 Results and Discussion 58
5.5 Related Work 59
5.6 Conclusions and Future Work 60

CHAPTER 6 COMBINING ALZHEIMER'S AND PARKINSON'S DISEASES DATA TO IDENTIFY PATIENTS AT RISK FOR DEMENTIA 61
6.1 Methods 62
6.2 Results and Discussion 67
6.3 Related Work 71
6.4 Conclusions and Future Work 72

CHAPTER 7 CONCLUSIONS AND FUTURE DIRECTIONS 73
7.1 Text Mining in Clinical Notes 73
7.2 Expanding Knowledge Networks with Information from Multiple Diseases . . 74
7.3 Utilizing Gene Expression Data 75

REFERENCES 76

CHAPTER 1: INTRODUCTION

Electronic medical records (EMRs) are the digital equivalent of paper records at a clinician's office, containing patient information such as treatment and medical history. Since their adoption into healthcare practices across the United States, EMRs have been estimated to save more than $81 billion annually [1]. This could be attributed to, for example, the fact that they can improve patient safety by alerting doctors to potentially harmful interactions that may result from prescribing multiple drugs [2]. Furthermore, by using EMR databases, doctors can apply stronger statistical methods to accomplish previously unfeasible tasks, such as large-scale relationship mining and clinical prediction of survival.

One of the most important aspects of EMRs is that they are typically heterogeneous, consisting of images, doctor notes, medical test results, genomic data, and more. Although each data type is uniquely informative (e.g., medical images can provide information that genomic data cannot), it is oftentimes difficult to analyze these different data sources within the same framework, which reduces the efficacy of many computational models. Inconsistencies among recordkeeping practices in different hospitals also introduce data sparsity, which further exacerbates this problem.

While the heterogeneity of EMRs presents drawbacks, the distinctive data types pave the way for tremendous power in knowledge discovery. For my thesis, I examined data from not only standard western medicine, but also traditional Chinese medicine in order to tackle domain-specific problems. I found that EMRs in both fields suffer from similar complications, including heterogeneous data types, inherent sparsity and incompleteness of medical records, and synonyms in the medical vocabulary, allowing for the development of general methods.

The goal of this thesis is to introduce models that explore the different data types in EMRs to improve their accessibility, universality, and analytical power. These models can be used in a wide variety of applications and have the potential to be directly implemented in clinical settings.

First, I designed PaReCat, a framework that was the first to use an external herb-symptom dictionary to identify subcategories of patients within specific diseases [3]. I used network embeddings to mine latent connections among similar concepts in the dictionary to enhance the information in the medical records. PaReCat can potentially help refine our understanding of diseases, as well as facilitate precision medicine and medical research.

Next, I discuss HEMnet, which enhances survival analysis by building upon PaReCat's concept of mining external knowledge [4]. This was the first method to integrate molecular interactions and medical records into a single network. By improving survival analysis, HEMnet can help doctors understand the features that differentiate patients who live for a long time from those who die more quickly.

I further built upon these works with VisAGE, which can create higher quality patient visualizations [5]. VisAGE was the first method to explore the effects of adding chemical-protein interactions and genetic information embeddings to enrich EMRs. The resulting visualizations can lead to more interpretable patient clusters and a better overall understanding of EMR data.

These works were based on many previous EMR analyses, such as disease subcategorization [6] and patient visualization [7]. Additionally, I drew inspiration from works that experimented with heterogeneous data sources [8, 9, 10, 11]. However, this work was the first to combine these two aspects by directly using heterogeneous data sources to enhance the analysis of EMR data.

Embedding vectors, which are a key component of my works, have also been explored in other papers. Choi *et al.* computed embeddings directly from EMRs [12], but their method relies on sequential information and does not mine external networks. Deep Patient also learns embedding vectors from EMRs [13]. However, it learns representations of patients rather than individual medical entities and does not use external networks. Word2vec provided the basis for much of the current research on learning efficient embeddings [14]. However, word2vec can only be applied to the text in patient records and cannot incorporate medical domain knowledge. This thesis was the first work to enrich EMR data by learning embedding vectors from external medical networks.

I also discuss a query expansion method that improves patient record retrieval [15]. This method utilizes a topic model, and was the first to frame symptoms and treatments as distinct languages that can be translated between one another. With this method, we can retrieve more relevant patients when given a new patient, which can be used in search engines for doctors.

Lastly, I discuss using heterogeneous data sources to study Alzheimer's disease and Parkinson's disease in the same feature space. Our work was the first to combine patients from the two diseases into a single dataset. We found that a classifier trained on the combined dataset better identified patients at risk for dementia. This study can save time for future data collections by allowing for a joint analysis of both diseases.

I conclude the thesis with a discussion of how each of these works fits into the future of computational healthcare and EMR analysis, as well as possible directions for additional research.

CHAPTER 2: PATIENT RECORD SUBCATEGORIZATION FOR PRECISION TRADITIONAL CHINESE MEDICINE

Any single disease will manifest itself differently in different patients. This phenomenon is the result of complex interactions between a patient's environment and numerous physiological and pathological factors, including genetic variation [16]. Consequently, patient diagnosis and treatment in real-world settings are extremely difficult due to population variation, creating demand for precision medicine. Western medicine tends to use many one size fits all over-the-counter drugs for common diseases, though increasingly available genomic data has conceived new opportunities for precision medicine.

In contrast, traditional Chinese medicine (TCM), a style of medicine that incorporates herbs and other natural products, has been leveraging personalized treatment as the core principle of clinical practice for thousands of years. Although the exact number of people who seek TCM treatment in the United States is unknown, the World Health Organization (WHO) estimated that 1 in 5 adults regularly consumes herbal products [17]. TCM is often used as a complementary and alternative medicine (CAM) and is especially effective for certain diseases such as stomach ailments. Indeed, several healthcare institutions have begun to integrate CAM therapies into their treatment programs, including the University of California, Los Angeles[1] and the University of California, San Francisco[2] [18].

TCM doctors prescribe mixtures of herbs tailored to each patient based on a thorough assessment of symptoms and physical condition; this is true even for common diseases. As a result, two patients with similar symptoms may receive completely different treatments due to the personalized nature of TCM. This manner by which TCM observes an individual's symptom patterns is reminiscent of precision medicine techniques. Thus, it could complement modern precision medicine, which currently relies mostly on molecular profiles [19].

Like in western medical settings, TCM doctors document symptoms that they observe from their patients. However, diseases consist of two components in TCM. The first roughly translates to *disease entity*, which is simply a set of symptoms. The second, unique to TCM, is the *syndrome*, which has no direct parallel in western medicine as the translation would suggest. Syndromes are best described as disharmonies at the core of the body. In western medicine, doctors often must resort to prescribing treatments based on symptoms alone in the event of failed differential diagnoses. On the other hand, TCM practitioners always assess and treat patients based on combinations of the two aforementioned components.

[1] ccim.med.ucla.edu

[2] osher.ucsf.edu

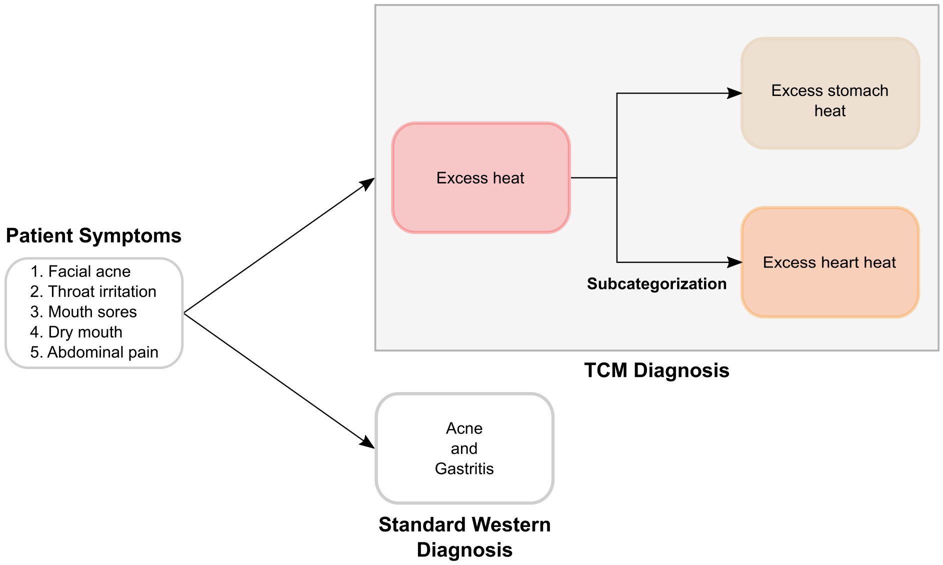


Figure 2.1: A comparison between TCM and modern western medicine. TCM prescribes treatments for the underlying syndromes, while western medicine prescribes treatments based on the symptoms.

The personalized, holistic treatment of a patient in TCM is especially useful for patients with multiple, seemingly unrelated symptoms. For example, a particular patient might have facial acne, throat irritation, mouth sores, dry mouth, and abdominal pain (Figure 2.1). In western medicine, this patient may be separately referred to a dermatologist and a gastroenterologist, each of whom might prescribe superficial medications to alleviate the symptoms. On the other hand, TCM doctors prescribe treatments for the patient's symptoms (disease entities) by determining their underlying causes (syndromes).

Unfortunately, this added dimension introduces an extra layer of complexity to TCM. Because two patients with the same disease might have different underlying patterns, TCM doctors will prescribe different herbs. This complication generates a necessity for subcategorizing patient records, which separates patients into groups based on their blanket diseases, and then further categorizes them into smaller clusters. For example, doctors will subcategorize the patient in Figure 2.1 into the *excess stomach heat* or *excess heart heat* subsyndromes.

In addition, subcategorizations can help inexperienced doctors learn pattern analysis by explaining specific medical cases. Most importantly, patient record subcategorization can help doctors view all possible diagnoses for their patients, reducing the chance of misdiagnosis. Misdiagnoses across all fields not only present ongoing risks to the health and safety of patients, but also cost the United States roughly $750 billion annually [20]. Furthermore, doctors can cross-reference their records with existing databases to obtain comparisons helpful in determining trends in prescriptions and treatment.

We conducted a study of this problem and aimed to accurately compute the similarity between patient records in TCM, a key step toward accurate subcategorization. Computationally, the problem of subcategorization can be solved by applying a clustering algorithm to group similar records together and then induce subcategories. Indeed, such a clustering approach was also the basis of our study. However, a straightforward application of a clustering algorithm for subcategorization in which we directly match two patient records was unlikely to be effective because it does not address the issue of variations in both symptoms and herbs (i.e., comorbid symptoms and functionally similar herbs). For example, yellow tongue coating and greasy tongue coating are comorbid symptoms, and often appear together for patients with the *excess heat* syndrome. The crow-dipper and the Chinese goldthread are functionally similar herbs, and are commonly prescribed to treat vomiting and abdominal pain.

Moreover, TCM patients typically have ten to fifteen symptoms and are prescribed a similar number of herbs, further complicating subcategorization. We concluded that we must be able to match patient records inexactly by tolerating variations in both symptoms and herbs, yet allowing related symptoms or herbs to somehow match with each other. This was the main technical challenge that we addressed in this work.

Our approach, which we called PaReCat (**Pa**tient **Re**cord Sub**cat**egorization), uses a dictionary-based embedding. Our hypothesis was that the prior knowledge of herb-symptom associations in a TCM dictionary could be used to discover latent relationships induced by comorbid symptoms and functionally similar herbs, thereby improving the quality of subcategorization. We performed extensive experiments on large-scale real-world datasets. As expected, this led to more accurate matchings of patient records than baseline approaches, and thus better subcategorization results. We also showed that PaReCat can be used immediately in multiple TCM clinical applications, such as retrieving similar patients as well as discovering meaningful cases of similar symptoms treated by different herbs and different symptoms treated by similar herbs.

2.1 PROBLEM DEFINITION

PaReCat takes two entities as input. The first is a set of n TCM patient records, $R = \{r_1, \ldots, r_n\}$, where $r_i \in R$ is a patient record and consists of $H_i = \{h_1, \ldots, h_p\}$ and $S_i = \{s_1, \ldots, s_l\}$. Here, H_i is the set of herbs and S_i is the set of symptoms for patient i. The second input is a set of known herb-symptom associations available in the TCM dictionary, which are functions of herbs outputting sets of symptoms, $f : H \to S$. An herb might have zero, one, or multiple symptom associations. PaReCat outputs a set of patient record

Table 2.1: An illustration of the conundrum of identifying discriminative attributes in a patient record.

Herbs prescribed only to asthma patient	**Herbs prescribed only to diarrhea patient**	**Herbs prescribed to both patients**
Chinese liquorice	*P. flos* extract	*A. propinquus*
Cinnamon	Japanese raisin tree	Poor man's ginseng
Ginseng	Sarsaparilla	White atractylodes rhizome
Tuckahoe	Lesser galangal	Barrenwort
Xanthium		Chinese parsnip root
		Yam extract
		Bamboo extract

categories $C = \{c_1, \ldots, c_m\}$, where each category $c_j \in C$ contains a subset of R.

We distinguished two kinds of categorizations. The first aims to group patients by considering both symptom and herbs. This allows us to create subcategories useful for cross-referencing patient treatments and discovering prescription trends. The second involves using only symptoms. This type of categorization emulates situations in which doctors are presented with a new patient and would like to view similar ones. We did not consider categorizations using only herbs because doctors do not prescribe treatments without first identifying symptoms.

Desirable categories will contain patients that are closely related by their symptoms, prescribed herbs, or both. These categories can be obtained by grouping similar patient records together. However, simply clustering TCM patients by their symptoms and prescriptions is ineffective. One cause for this difficulty is that TCM patients may display comorbid symptoms or be prescribed functionally similar herbs, but have very different ailments. These situations occur because many symptoms, such as coughing or bleeding, are not specific to any one disease.

In a similar vein, when prescribing treatments to patients, TCM doctors will assign batches of herbs, some of which are intended to treat the underlying syndromes instigating the observed symptoms. As a result, two patients who suffer from different diseases may be prescribed very similar herbs if they have similar syndromes. For example, in our dataset, a patient suffering from diarrhea was prescribed eleven herbs, seven of which were also prescribed to an asthma patient (Table 2.1).

Conversely, TCM patients might have identical diagnoses, yet very different herb prescriptions. This happens because the patients belong to different subcategorizations. For instance, between two herbs h_1 and h_2 that treat the same disease, a doctor might opt to prescribe h_1 over h_2 to a patient with chronic gastritis because the patient displays the *ex-*

Table 2.2: Two patients, both diagnosed with asthma, but prescribed completely different herbs. We show fifteen of them here.

Asthma patient *A*	**Asthma patient *B***
Chinese ephedra	White mulberry leaves
Tibetan apricot	White mulberry bark
Gyspum fibrosum	Mulberries
Chinese liquorice	Chinese taxillus twig
Powdered water buffalo horn	*A. japonica*
R. glutinosa	Horned holly leaves
Chinese tree peony	Leopard lily
Chinese figwort root	*A. propinquus*
Weeping forsythia	Chinese parsnip root
Japanese honeysuckle	White atractylodes rhizome
Rhubarb root	Roasted coltsfoot
F. thunbergii	Barrenwort
A. asphodeloides	Coixseed
Baikal skullcap	Green mandarine peel
Chinese gourd	Mandarine peel

cess heat syndrome, which h_1 specifically treats, but h_2 does not. In a particular case in our dataset, a patient suffering from asthma was prescribed 66 different herbs, none of which overlapped with another asthma patient's prescription of 57 herbs (Table 2.2).

These circumstances occur only in TCM and do not arise in western medical cases, in which doctors prescribe only a handful of drugs that each treats specific symptoms or diseases. Systematic discovery of the same symptoms treated by variations of herbs and variations of symptoms treated by the same herbs is essential in transforming empirical TCM data into useful medical knowledge. This was the main goal of our study. PaReCat aims to capture these complications by leveraging a dictionary containing prior TCM knowledge.

2.2 METHODS

In this section, we first give an overview of the model, then discuss each component in detail.

2.2.1 PaReCat Overview

We started by obtaining a set of known herb-symptom associations in the dictionary [21], which contained rules that map herbs to the symptoms they treat. For example, the crow-dipper had multiple entries, treating symptoms from vertigo to breathing difficulties. There were 1,995 herbs, 1,635 symptoms, and 27,824 treatment rules in the dictionary.

From these associations, we constructed a bipartite network, G, in which one part of the network consisted of symptoms and the other part consisted of herbs. Symptoms that were associated with similar herbs were close to each other in the network, and *vice versa*.

We then applied a network embedding approach to learn a low-dimensional vector representation for each herb and symptom. These low-dimensional vectors optimally preserved the original associations between symptoms and herbs. We then computed the similarity between each pair of features by using the cosine similarity between their corresponding low-dimensional vectors. By using these similarity scores, which enabled inexact matchings of symptoms and herbs, we determined the similarity between any two patients, even if they shared no herbs or symptoms.

Finally, we applied agglomerative clustering on the patient record and learned the cluster for each patient. The main novelty of our method lied in utilizing external knowledge, the herb-symptom dictionary, to cluster TCM patient records.

2.2.2 Building the Bipartite Network

First, from the set of known herb-symptom associations, $A : H \rightarrow S$, we constructed a bipartite network, $G = (H, S, E)$, in which the two disjoint sets, H and S, were the sets of herbs and symptoms, respectively. In our dictionary, $|H| = 1,995$ and $|S| = 1,635$.

For each mapping $h \rightarrow s$ $(h \in H, s \in S)$ in A, we added a node for h and a node for s if they were not already in G, and created an edge between them. Thus, $|E| = |A|$. In our dictionary, $|E| = 27,824$.

2.2.3 Network Embedding

Next, we performed network embedding, which took as input the bipartite network, G. We used a recently developed network embedding approach, diffusion component analysis (DCA), to learn low-dimensional vector representations of the herbs and symptoms in the network [22]. DCA has been shown to achieve state-of-the-art results in learning network structure for gene function prediction [23].

DCA takes a network as input and outputs a low-dimensional representation for each node in the network. It ensures that two nodes will have very similar low-dimensional representations if they are topologically close in the network. Thus, related symptoms and related herbs tend to have similar low-dimensional vector representations, enabling inexact matchings. Between each pair of vectors, we computed the cosine similarity score to find the association between the corresponding symptoms or herbs. Each symptom and herb had a similarity score with every other symptom and herb. For example, *Streptococcus dysgalactiae*, a bacteria strain that causes indigestion, and *yinianjin*, a powder consisting of several herbs, had a similarity score of 0.85272. Indeed, a capsule with *yinianjin* as the primary ingredient has been developed to treat children afflicted with *S. dysgalactiae* [24].

We computed a total of $(|H| + |S|)^2$ similarity scores. The similarity score between a feature and itself is always equal to 1. With these scores, we output a similarity matrix, M, of shape $(|H| + |S|) \times (|H| + |S|)$. The diagonal consisted of only 1's, ensuring that an exact matching was always treated as the most reliable match. Non-zero values off the diagonal were in the range of $[-1, 1]$. These values captured the extent to which different symptoms and herbs match with each other. Furthermore, we set a similarity score threshold s, and only considered entity pairs with a similarity score larger than s. Scores in M below s were changed to 0.

For any particular run of PaReCat, the set of features were symptoms or both symptoms and herbs. As a standard step in text mining, we found that filtering out high-frequency words and low-frequency words led to better results. Thus, we also introduced parameters to remove herbs that appeared in fewer than β_{herbs} patient visits ($\beta_{\text{herbs}} \in \mathbb{Z}$) and more than α_{herbs} of all patient visits ($0.0 \leq \alpha_{\text{herbs}} \leq 1.0$). Our model similarly removed symptoms for α_{symptoms} and β_{symptoms}. If the features were only symptoms, then the values of α_{herbs} and β_{herbs} were inconsequential.

After filtering, M became an $m \times m$ similarity matrix, where $m \leq |H| + |S|$ was the number of symptoms and herbs post-processing. Note that the inexact matchings allowed by matrix M surpass symptom-symptom and herb-herb matchings to also allow symptom-herb matchings.

2.2.4 Enhancing the Patient Profile Matrix

We defined the patient profile matrix, P, to be an $n \times m$ matrix, where n was the number of patients in the medical record and m was the number of symptoms and herbs, as in the similarity matrix, M. Each row of P was a binary vector corresponding to a patient, where a 1 denoted that the patient was prescribed the herb or possessed the symptom corresponding to the column, 0 otherwise.

After learning the similarity matrix, M, from the dictionary, we used it to enhance P. The purpose of this enhancement was to make some zero elements (absent symptoms or herbs) in the patient profile matrix non-zero if they had enough support from known associated symptoms and herbs. M provided support for any given element. Formally, we performed the following matrix multiplication:

$$P' = P \times M \tag{2.1}$$

where P' was the enhanced $n \times m$ patient profile matrix. After multiplication, the dimensions were preserved such that each patient was still regarded as a sample, and each herb or symptom a feature.

Intuitively, after the matrix multiplication, symptoms and herbs that had high similarity to many other symptoms and herbs in the original patient vector tended to have higher values. In effect, this augmented the original patient vector to potentially include additional related symptoms and herbs.

2.2.5 Agglomerative Clustering

Lastly, we clustered on the enhanced patient vectors, which accommodated inexact matchings and enabled more accurate patient record matching. We could have used any similarity function to compute the similarity between two patient records. In our experiments, we chose cosine similarity as the affinity measure in the clustering, which was utilized in the similar task of western medical record linkage [25]. Similarly, our enhanced vector representations could also support any clustering algorithm. In our experiments, we chose agglomerative clustering with average linkage, which has been shown to be useful for clustering herbs in TCM data [26]. The advantage of such a hierarchical clustering algorithm was that we could obtain a detailed subcategorization at different levels.

2.3 RELATED WORK

Our method was the first to combine patient record analysis and herb-symptom associations for subcategorizing traditional Chinese medicine records. However, data-driven approaches on TCM data have attracted more attention in recent years. In one particular study, the authors employed Chi-Squared Automatic Interaction Detection (CHAID) decision trees to identify and differentiate syndromes associated with coronary heart disease, performing classifications on the syndromes with k-core network analysis [27]. One study combined network construction and cluster analysis to study relationships and associations among TCM patient records [6]. However, it performed analysis with only symptoms and did not leverage herb-symptom associations to solve the issues associated with comorbid symptoms and functionally similar herbs in TCM. He *et al.* also used agglomerative clustering on TCM data, but categorized herbs based on their efficacies in order to analyze their chemical components [26]. Roque *et al.* used cosine similarity between patient records to analyze disease co-occurrence [25]. However, they did not consider treatment information, which is crucial to patient record subcategorization. Furthermore, methods designed for western medical records are difficult to apply to TCM records because firstly, some symptoms (e.g., vacuity and depletion) are unique to TCM [28]. Secondly, TCM generally has more symptoms and herbs per patient. Our experimental results confirmed that the methods proposed to cluster western medical records indeed perform poorly on TCM datasets.

2.4 DATASET DESCRIPTION

We conducted experiments on two datasets to quantitatively and qualitatively evaluate the effectiveness of PaReCat. After obtaining clusters of patient records, we evaluated with disease labels in the patient records as the ground truth: records that had the same label were expected to be in the same cluster.

We used a TCM textbook containing 2,276 patient medical records. These patient medical records were organized into a three-level hierarchy based on their identifying categories. There were three categories at the topmost level, which we refer to as level 1: *women and children*, *internal medicine*, and *surgical acupuncture*. The middle level, which we call level 2, was more specific, and included categories such as *exogenous seasonal diseases*. The bottommost level, or level 3, included specific ailments such as the common cold.

We used these disease categories as cluster labels for their corresponding patient records. There were three level 1 labels, 51 level 2 labels, and 274 level 3 labels. Because we could interpret these labels as the ground truth, this dataset was ideal for quantitative evaluation.

In addition to the medical textbook, we further evaluated our model on a much larger medical record containing 9,529 anonymous patients, obtained from a major hospital in China. These patients all had some variety of stomach disease. This dataset did not have detailed labels for quantitative evaluation. We used it to understand whether our approach could offer interesting insights into stomach disease subcategories. The doctor who treated these patients manually assessed the subcategorization results.

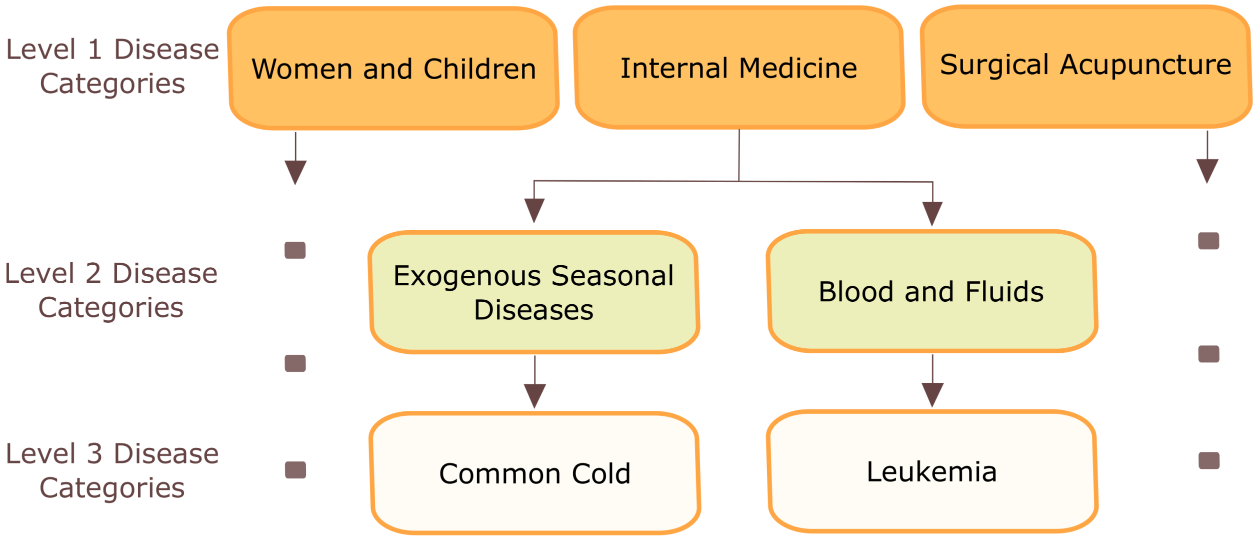

Level 1 Disease Categories
Women and Children
Internal Medicine
Surgical Acupuncture
Level 2 Disease Categories
Exogenous Seasonal Diseases
Blood and Fluids
Level 3 Disease Categories
Common Cold
Leukemia

Figure 2.2: The disease label hierarchy in our TCM dataset.

2.5 EXPERIMENTAL DESIGN

2.5.1 Setting the Parameters

Tuning values of α_{symptoms}, β_{symptoms}, α_{herbs}, and β_{herbs}, which were the thresholds controlling the filtering of frequent and rare symptoms and herbs, we achieved the best results for clustering without embedding on symptom features with $\alpha_{\text{symptoms}} = 0.05$ and $\beta_{\text{symptoms}} = 1$. For clustering without embedding on symptom and herb features, we obtained the best results with $\alpha_{\text{symptoms}} = 0.2$, $\beta_{\text{symptoms}} = 5$, $\alpha_{\text{herbs}} = 0.1$, and $\beta_{\text{herbs}} = 2$. To strengthen our baseline, we fixed these optimal parameters for their corresponding counterparts with embedding (i.e., clustering with embedding on symptom features also used $\alpha_{\text{symptoms}} = 0.05$ and $\beta_{\text{symptoms}} = 1$). Tuning s for clustering with embedding, we achieved the best results for symptom features with $s = 0.96$ and for combined symptom and herb features with $s = 0.98$.

2.5.2 Evaluation Metrics

We scored the quality of the clusters with the adjusted Rand index [29]. It has been widely used in evaluating clustering results [30], defined as

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}} \tag{2.2}$$

where n_{ij}, a_i, and b_j are values from the contingency table generated from the overlap between two groupings $X = \{X_1, \ldots, X_r\}$ and $Y = \{Y_1, \ldots, Y_s\}$. Each entry n_{ij} denotes

Table 2.3: Best performances for each feature type, measured by the adjusted Rand index.

	Symptoms	**Symptoms and Herbs**
k-means	0.0174	0.0770
Spectral	0.0653	0.0843
Agglomerative	0.1613	0.2717
PaReCat	**0.1672**	**0.2754**

the number of objects in common between X_i and $Y_j : n_{ij} = |X_i \cap Y_j|$. a_i denotes the sum of the entries for the group X_i, and b_j denotes the sum of the entries for the group Y_j. A higher score indicates a better clustering with respect to the ground truth. The adjusted Rand index is in the range of $(-\infty, 1]$.

2.5.3 Quantitative Evaluation

In addition to agglomerative clustering, we employed k-means, a common clustering technique, and spectral clustering, which has been used to cluster western medical records for the task of predicting healthcare costs for individuals [31].

The adjusted Rand indices for k-means, spectral clustering, and agglomerative clustering are shown in Table 2.3. For clustering without embedding (first three rows), we note that agglomerative clustering performed the best, as expected.

The bottom two rows of the table show the results comparing agglomerative clustering without and with embedding (baseline and PaReCat, respectively). We observed improvement by embedding known herb-symptom associations for both types of features. Embedded feature vectors of symptoms and herbs achieved an adjusted Rand index of 0.2754, which was higher than that of feature vectors without embedding, 0.2717. Similarly, embedded feature vectors of symptoms alone achieved an adjusted Rand index of 0.1672, an improvement over that of feature vectors without embedding, 0.1613. The clustering results were very sensitive to s. For features of both symptoms and herbs, decreasing s to 0.9 lowered the adjusted Rand index of the resulting clusters to 0.22142.

Though the improvement was small, we showed that PaReCat can indeed improve clustering performance by adding external information. Furthermore, we can gain new knowledge from clusters that mismatch the ground truth.

2.5.4 Qualitative Evaluation

To investigate whether PaReCat could effectively cluster patients into informative subcategories, we applied PaReCat to the larger set of stomach disease medical records. Here, we randomly selected three clusters and had the TCM doctor who treated the patients verify PaReCat's effectiveness. The expert found two of them to be especially coherent and informative (Tables 2.4 and 2.5, where each row denotes a patient record).

In the first cluster (Table 2.4), all of the patients were diagnosed with chronic gastritis. In addition to this similarity, the records stated that these patients showed syndromes of either *qi* deficiency or *yin* deficiency. The related symptoms here were weakness and pale tongue for *qi* deficiency, dry mouth and dry stool for *yin* deficiency, and deep, thready pulse for both deficiencies. The corresponding herbs for these patients specifically target *qi* deficiency or *yin* deficiency. For example, poor man's ginseng, *A. propinquus* honey, tuckahoe, and fried white atractylodes strengthen the spleen. *R. glutinosa*, female ginseng, and Chinese peony are blood enrichers. These benefits supplement and fortify the *qi* and *yin*. The symptoms and herbs of interest are bolded in the table. This example illustrates how TCM treats its patients by identifying the underlying syndromes.

We show another meaningful cluster in Table 2.5. The records stated that all of the patients had symptoms associated with the *excess heat* syndrome, such as coughing, chest tightness, dry mouth, and dry stool. Consequently, the patients were prescribed herbs specifically for these symptom-syndrome combinations. We can see that there were not many overlaps in symptoms and herbs among the patients, which may be the reason that traditional clustering methods failed to group the patients together. In spite of this difficulty, PaReCat successfully clustered the patients by using embedding-based similarity measurements.

2.6 APPLICATIONS OF PARECAT

As mentioned in the previous section, PaReCat can be used to achieve high quality, informative subcategorizations of patients. Here, we show how we can use them in three different applications to directly support doctors.

2.6.1 Similar Patient Retrieval

PaReCat can be used to retrieve similar patients. We show the results of a sample retrieval here. We extracted the five most similar pairs of patients as computed by PaReCat, of which we highlight two.

Table 2.4: In this subcategorization example, each row is a patient with *qi* or *yin* deficiency syndromes.

Diseases	Symptoms	Herbs
chronic gastritis	**weakness**; **pale tongue**; crenated tongue; yellow tongue coating; **deep, thready pulse**	*A. propinquus*; bugbane; Chinese goldthread; coixseed; crow-dipper; **female ginseng**; **fried white atractylodes**; ginger; gingered magnolia bark; mandarine peel; Mongolian dandelion; *P. arecae*; **poor man's ginseng**; **tuckahoe**
chronic gastritis	**weakness**; weight loss; hiccups; **dry mouth**; recurrent oral ulcers; joint pain; lack of sleep; **pale tongue**; tongue lacerations; yellow tongue coating; **deep, thready pulse**	*A. asphodeloides*; ***A. propinquus* honey**; Baikal skullcap; calamus; Chinese figwort; Chinese liquorice; crow-dipper; **female ginseng**; *A. lancea*; fried barley; **fried Chinese peony**; fried jujube; **fried white atractylodes**; gypsum fibrosum; Mongolian dandelion; pine silk tree; **poor man's ginseng**; sweet wormwood; **tuckahoe**
chronic gastritis; gastroptosis	abdominal pain; acid reflux; halitosis; **dry mouth**; poor sleep; hair loss; dysmenorrhea; **pale tongue**; crenated tongue	*A. propinquus*; Baikal skullcap; bitter orange; black sesame; Chinese knotweed; Chinese liquorice honey; Chinese thuja leaves; Chinese tree peony; curcuma; **female ginseng**; fried jujube; **fried white atractylodes**; ***R. glutinosa***; red sage; **tuckahoe**
hiatal hernia; **chronic gastritis**; reflux esophagitis	abdominal distension; acid reflux; belching; **dry mouth**; **weakness**; **dry stool**; poor sleep	*A. propinquus*; Baikal skullcap; bitter orange; Chinese goldthread; Chinese liquorice honey; crow-dipper; cuttlebone; *F. thunbergii*; **female ginseng**; **fried Chinese peony**; **gingered magnolia bark**; **immature bitter orange**; **Japanese *Inula***; **perilla stem**; **poor man's ginseng**; **red thorowax**; **redstem wormwood**; ***T. ruticarpum***; **turmeric rhizome**

Table 2.5: In this subcategorization example, all patients had symptoms associated with the *excess heat* syndrome.

Diseases	**Symptoms**	**Herbs**
chronic gastritis; URTI	acid reflux; heartburn; **dry mouth**; halitosis; belching; excess phlegm; poor appetite; **chest tightness**; **coughing**; **dry stool**; cracked tongue	red sage; crow-dipper; monkeygrass; umbrella polypore; magnolia bark; *A. propinquus* honey; Oriental water-plantain; perilla leaf; Chinese liquorice honey; yam extract; Chinese cornel dogwood; American silvertop; *R. glutinosa*; gypsum fibrosum; *Lophatherum*; Chinese gourd; bishop's weed; cinnamon; tuckahoe
chronic gastritis	acid reflux; **coughing**; white sputum; **dry stool**; vomiting; hypouresis	*A. chinensis*; nacre; cuttlebone; tuckahoe; Chinese goldthread; peach seeds; Chinese liquorice; Baikal skullcap; crow-dipper; Chinese peony; gingered magnolia bark; Chinese rhubarb; citron fruit; bitter orange; perilla stem
chronic gastritis	abdominal pain; belching; heartburn; acid reflux; constipation; **coughing**; dark, crenated tongue	crow-dipper; Chinese liquorice honey; bitter orange; Chinese goldthread; gingered magnolia bark; red thorowax; bamboo extract; Mongolian dandelion; fried Chinese peony; *O. diffusa*; turmeric rhizome; cuttlebone; Baikal skullcap
chronic gastritis	abdominal pain; fullness; **chest tightness**; lumbago; **coughing**; **dry mouth**; **dry stool**; dark, swollen tongue; greasy tongue coating; thready pulse	perilla leaf; fried white atractylodes; bugbane; Chinese bellflower; red thorowax; Baikal skullcap; tuckahoe; immature bitter orange; Chinese goldthread; ginger; Japanese *Inula*; Java grass; mandarine peel; crow-dipper; curcuma; gingered magnolia bark; cinnamon; chicken gizzard; American silvertop; bitter orange; bishop's weed

Table 2.6: Example of a pair of similar patients with *cold* and *heat* syndromes.

Diseases	Symptoms	Herbs
chronic gastritis	**body chills**; **dry mouth**; *heat* excess; **cold hands and feet**; yellow complexion; watery stool; **bitter taste**; glossodynia; **halitosis**; dark red, swollen tongue; **yellow tongue coating**	poor man's ginseng; fried white atractylodes; Chinese cinnamon; Chinese liquorice; Chinese goldthread; *T. ruticarpum*; Amur cork tree; female ginseng; cinnamon; Chinese plantain; Chinese tree peony; *R. glutinosa*; shrubby sophora; gypsum fibrosum; common rush; fried *A. lancea*
chronic gastritis	stomach pain; **cold hands and feet**; **body chills**; **dry mouth**; **halitosis**; dizziness; watery stool; **bitter taste**; sore gums; dark red, swollen tongue; **yellow tongue base coating**	*A. propinquus*; fried white atractylodes; kudzu; Chinese goldthread; Chinese liquorice honey; fried gardenia; red thorowax; Chinese tree peony; bamboo extract; *R. glutinosa*; Mongolian dandelion; fried Chinese peony; gypsum fibrosum; Baikal skullcap; *A. asphodeloides*; *T. ruticarpum*; female ginseng; Chinese rhubarb

According to the doctor, the first pair of patients (Table 2.6) shared the underlying syndrome of *cold deficiency*, represented by symptoms such as body chills and cold hands and feet. In addition, they also showed symptoms for the *heat* syndrome, such as dry mouth, bitter taste, halitosis, and yellow tongue coating. These two patients' conditions are typical examples of *cold* and *heat* syndromes. Consequently, doctors prescribed both *cold* and *heat* treatment herbs.

The second pair of patients were both diagnosed with upper respiratory tract infection (URTI) in addition to chronic gastritis (Table 2.7). Recall that our method did not utilize the disease diagnosis information in clustering. However, we still managed to find these patients, despite the fact that they shared no symptoms. Among the herbs shared by the two patients, four of them (powdered water buffalo horn [32], Japanese apricot [33], mint [34], and woad root [35]) are specific treatments for URTI. Our methods successfully identified these four herbs, clustering the two patients together.

2.6.2 Similar Symptoms with Different Herbs

Another application is identifying patients that have similar symptoms but are treated with different herbs. Conversely, identifying patients with different symptoms but treated with similar herbs is also an interesting task. These cases are unique to TCM, since western medicine tends to treat the same symptoms with the same drugs. Our method was able to

Table 2.7: Example of a pair of similar patients with no common symptoms.

Diseases	Symptoms	Herbs
chronic gastritis; **URTI**	cracked tongue; pale tongue; bloating; stomach pain; weakness; epigastric chills; white tongue coating; belching; reddish tongue; epigastric pain; bowel discomfort; abdomen chills	ginger; Chinese liquorice honey; **mint**; crow-dipper; Chinese tree peony; Chinese cinnamon; **powdered water buffalo horn**; fried white atractylodes; shrubby sophora; fried coixseed; weeping forsythia; bitter orange; tuckahoe; **Japanese apricot**; Chinese goldthread; American silvertop; hyacinth orchid; **woad root**; Chinese parsnip root; lesser reedmace; Baikal skullcap; fried Chinese peony; Japanese honeysuckle; poor man's ginseng; costus; female ginseng
insomnia; **chronic gastritis**; **URTI**	dry mouth; thready pulse; thirst; dry throat; dry lips; greasy tongue coating; swollen, crenated tongue	Chinese liquorice honey; crow-dipper; **mint**; weeping forsythia; bitter orange; tuckahoe; **Japanese apricot**; Chinese goldthread; Baikal skullcap; fried Chinese peony; female ginseng; costus; ginger; Chinese tree peony; **powdered water buffalo horn**; shrubby sophora; American silvertop; **woad root**; Chinese parsnip root

Table 2.8: Two patients clustered together. They had similar symptoms, but were prescribed herbs that treat different diseases.

Diseases	Symptoms	Herbs
superficial gastritis	**dry mouth**; dry stool; heartburn; halitosis; **insomnia**; abdomen chills; belching; **bloating**; **yellow tongue coating**; **acid reflux**; bitter taste	Chinese goldthread; Chinese liquorice honey; sandalwood; ginger; crow-dipper; cuttlebone; immature bitter orange; female ginseng; ginger; Chinese gourd; red sage; Java grass; bitter orange; Baikal skullcap; *M. toosendan*
gallbladder polyps	**dry mouth**; thready pulse; **insomnia**; **bloating**; **acid reflux**; **yellow tongue coating**; frequent urination; cracked tongue	***O. diffusa***; immature bitter orange; Chinese gourd; cuttlebone; bitter orange; Baikal skullcap; monkeygrass; Chinese figwort; ginger; fried Chinese peony; female ginseng; mandarine peel; crow-dipper; gingered magnolia bark; Chinese cinnamon; Chinese rhubarb; **vinegared chicken gizzard**

identify these two categories of patients.

The first case (similar symptoms, different herbs) is especially important, and plays a large role in TCM misdiagnoses. An inexperienced doctor may see a patient with symptoms similar to one he or she had previously treated. Traditional methods might confound patients with similar symptoms. However, PaReCat can view previous medical records, taking into account both symptoms and herbs, and help new doctors improve decision-making when facing similar circumstances.

For example, a pair of patients in our medical record both displayed dry mouth, bloating, insomnia, acid reflux, and a yellow tongue coating (Table 2.8). However, one patient was treated with *Oldenlandia diffusa* and vinegared chicken gizzard. This patient suffered from gallbladder polyps, which the herbs specifically treat (*O. diffusa* [36], vinegared chicken gizzard [37]). On the other hand, the other patient was not treated with these herbs, and was instead diagnosed with superficial gastritis.

2.6.3 Different Symptoms with Similar Herbs

In addition to patients with similar symptoms but different herbs, it is interesting to study patients that have different symptoms but are treated with similar herbs. Table 2.9 shows two patients that had very different symptoms. However, they were prescribed similar herbs. *M. toosendan* is a particular herb of note, which specifically treats liver issues [38]. Indeed, both patients were diagnosed with liver ailments (hepatic steatosis and cirrhosis). This relationship is not uncommon to TCM; doctors often prescribe a multitude of herbs as a supplement to a main herb (in this case, *M. toosendan*). PaReCat successfully filtered out ambiguous herbs to discover these relationships.

2.7 CONCLUSIONS AND FUTURE WORK

Mining subcategorizations from TCM medical records is an important task for precision medicine. In this work, we proposed a novel patient record subcategorization model called PaReCat. PaReCat was able to obtain patient subcategorizations that characterize the underlying syndromes behind the observed symptoms and herbs. It uses a novel dictionary-based embedding approach to solve challenges associated with comorbid symptoms and functionally similar herbs. We performed experiments on two real-world datasets and observed substantial improvement in patient subcategorizations created by PaReCat. We also verified the subcategorizations to be meaningful for understanding variations of stomach diseases.

Table 2.9: Two patients clustered together with different symptoms, but treated with similar herbs. Shared herbs are bolded.

Diseases	**Symptoms**	**Herbs**
hepatic steatosis; gallbladder inflammation	weakness; laryngitis; white tongue coating; dark, purple tongue; dark, red tongue; deep, thready pulse	**ginger**; curcuma; **vinegared *C. yanhusuo***; **cuttlebone**; red thorowax; **bitter orange**; *E. ulmoides*; monkeygrass; *R. glutinosa*; **Baikal skullcap**; **Chinese peony root**; ***M. toosendan***; lotus leaf; **vinegared Java grass**; **female ginseng**; gardenia
cirrhosis	weakness; pale tongue; dizziness; back pain; abdomen chills; chest tightness; weight loss; belching; blurred vision; chest pain	Chinese liquorice honey; magnolia bark; perilla leaf; **cuttlebone**; **bitter orange**; **Baikal skullcap**; Chinese goldthread; **fried Chinese peony**; **ginger**; **vinegared Java grass**; fried *A. lancea*; **female ginseng**; mandarine peel; crow-dipper; pachouli; **vinegared *C. yanhusuo***; *T. ruticarpum*; false starwort; perilla stem; ***M. toosendan***

PaReCat's generality allows it to be applied to any TCM dataset to discover interesting subcategories that are immediately useful to not only research, but also clinical applications. PaReCat is completely unsupervised, which has the advantage of requiring no manual work. However, to improve the accuracy of subcategorization, we can explore semi-supervised subcategorizations in which we allow a doctor to provide feedback on the clustering results, which can then be used as labels for additional clustering.

To further improve our method, we can employ natural language analysis. One issue in the data that could be solved by natural language analysis was that textual differences separated essentially identical herbs and symptoms. For example, a doctor might prescribe a crushed variant of an herb to a patient, creating a new herb in the record, though the main ingredient remains the same.

CHAPTER 3: INTEGRATION OF ELECTRONIC MEDICAL RECORDS WITH MOLECULAR INTERACTION NETWORKS AND DOMAIN KNOWLEDGE FOR SURVIVAL ANALYSIS

As discussed in the introduction, with more accessible EMR databases, doctors can apply stronger statistical methods to accomplish previously unfeasible tasks, such as relationship mining and clinical prediction of survival (CPS). Relationship mining allows doctors to discover useful associations among entities in medical records, including novel drug usages and adverse drug events [39]. On the other hand, CPS allows doctors to predict a new patient's probability of survival, which can help hospitals optimize resource allocation and treatment planning. In particular, accurate survival estimates for terminally ill patients can prevent inappropriate therapies and avoid unnecessary toxicity [40].

Analyses such as the two aforementioned applications utilize patient features extracted from EMRs. These features include test results, clinical notes, symptoms, diagnoses, and medical history. However, there are two main challenges that frequently appear with these features:

1. **Missing data.** Many methods assume the availability of all features. However, this assumption typically does not hold for many EMR databases. One reason is that doctors do not perform all existing medical tests on each patient. Other reasons include incomplete medical records or inconsistent text data in the form of clinical notes. Mean imputation, the most common method of filling missing values, has been shown to introduce noise rather than reduce it [41].

2. **Semantic mismatching.** Patients with similar but distinct features may be judged to be dissimilar. In the traditional vector space model, in which each unique word in the vocabulary occupies a dimension, patient records require exact matches of features in order to be considered similar (Table 3.1). This problem of semantic mismatching has been addressed with methods such as word2vec, which trains similar representations for similar words [14]. However, word2vec has had limited success in the context of medical records [12]. Comorbid symptoms and functionally similar herbs, described in the previous chapter, contribute to semantic mismatching.

Because of these challenges, existing models have been unable to effectively group similar EMRs together. Fortunately, genetic and protein interactome databases have been rapidly growing due to advances in high-throughput experiments [42] and improved biocuration via text mining [43]. These interactomes typically take the form of molecular interaction networks. We can analyze the topologies of these networks and extract meaningful patterns.

Table 3.1: An example of the semantic mismatching problem in the traditional vector space model. The two patients in this table have similar cancer diagnoses (adenocarcinoma and adenosquamous carcinoma) and synonymous symptoms (halitosis and bad breath). Despite this, standard models do not judge the two patients to be similar.

	Adenocarcinoma	**Adenosquamous carcinoma**	**Halitosis**	**Bad breath**
patient_1	✓	×	×	✓
patient_2	×	✓	✓	×

In this work, I expanded on PaReCat's enhancement of patient records by further exploring the idea of guilt by association: associated or interacting entities in a network are more likely to be functionally related [44]. However, instead of utilizing just an herb-symptom dictionary, I included a molecular interaction network. Using this network, my study revealed implicit relationships among nodes with common neighbors. I also integrated medical records as well as existing domain knowledge into this network to help solve the primary EMR challenges of missing data and semantic mismatching.

I named this framework the **HE**terogeneous **M**edical record **net**work (HEMnet), which consists of information derived from EMRs, molecular interaction networks, and domain knowledge. HEMnet was the first method to integrate medical records and molecular interactions into a single network.

I showed that efficient node representations trained from HEMnet can be used to enhance EMRs. Furthermore, I showed that the enhanced EMRs can better group similar patients whose records may otherwise be misinterpreted due to missing data and semantic mismatching.

Though I performed experiments in the context of EMR enhancement, HEMnet is a general model that can be applied to many different tasks.

3.1 HEMNET OVERVIEW

3.1.1 Definition of HEMnet

HEMnet is a network that consists of nodes and edges associated with different types of information. Formally, we defined HEMnet as a graph $G = (V, E, R)$, where V is the set of typed nodes (i.e., each node belongs to a specified type), E is the set of typed edges, and R is the set of edge types. An edge $e \in E$ in HEMnet is an ordered triplet $e = \{u, v, r\}$, where $u, v \in V$ are typed nodes and $r \in R$ is the corresponding edge type.

In our study, we utilized four distinct categories of edges to create HEMnet. The first three categories were drawn from external databases, while the last category drew directly from the EMR database.

1. **Protein-protein interaction network.** This network was based on HumanNet, an external network of protein-encoding genes [45]. For a functional linkage between two proteins p_1 and p_2, we created a node for p_1, a node for p_2, and an undirected edge $\{p_1, p_2\}$ in HEMnet. This was the molecular interaction network.

2. **Herb targets.** This database contained herbs and the proteins they target, curated from expert knowledge and medical literature. We created a node h, a node p, and an undirected edge $\{h, p\}$ if an herb h targeted a protein p. This was domain knowledge.

3. **Herb-symptom dictionary.** This database is a TCM textbook consisting of herbs and the symptoms they treat, also curated from expert knowledge and medical literature. We created a node h, a node s, and an undirected edge $\{h, s\}$ if an herb h treated a symptom s in the dictionary. This was domain knowledge.

4. **Electronic medical records.** We directly added co-occurrence edges from each medical record. For example, if a patient was diagnosed with cancer type c and prescribed a drug d, then we created a node c, a node d, and an undirected edge $\{c, d\}$. We repeated this for all elements in each patient's medical record.

Because of the sparsity of patient records, the external information was critical in discovering relationships among EMR entities that may have otherwise been hidden. Overall, $|V| = 11,911$, $|E| = 379,715$, and $|R| = 23$ for HEMnet in our experiments. It is important to note that only 449 of the nodes were EMR features, and that the remaining nodes were proteins.

Another feature of note is that the HEMnet is unweighted, which stems from the assumption that all its relationships are equally reliable. We made this assumption because our original databases were derived from high-quality experiments. However, this would weight future edges of lower confidence, such as those obtained from text mining, the same as our current edges. We discuss this limitation in the conclusions.

3.1.2 Network Embedding

Like with PaReCat, we utilized network embedding to reduce the dimensionality of the nodes down to vectors. However, instead of diffusion component analysis (DCA), we used

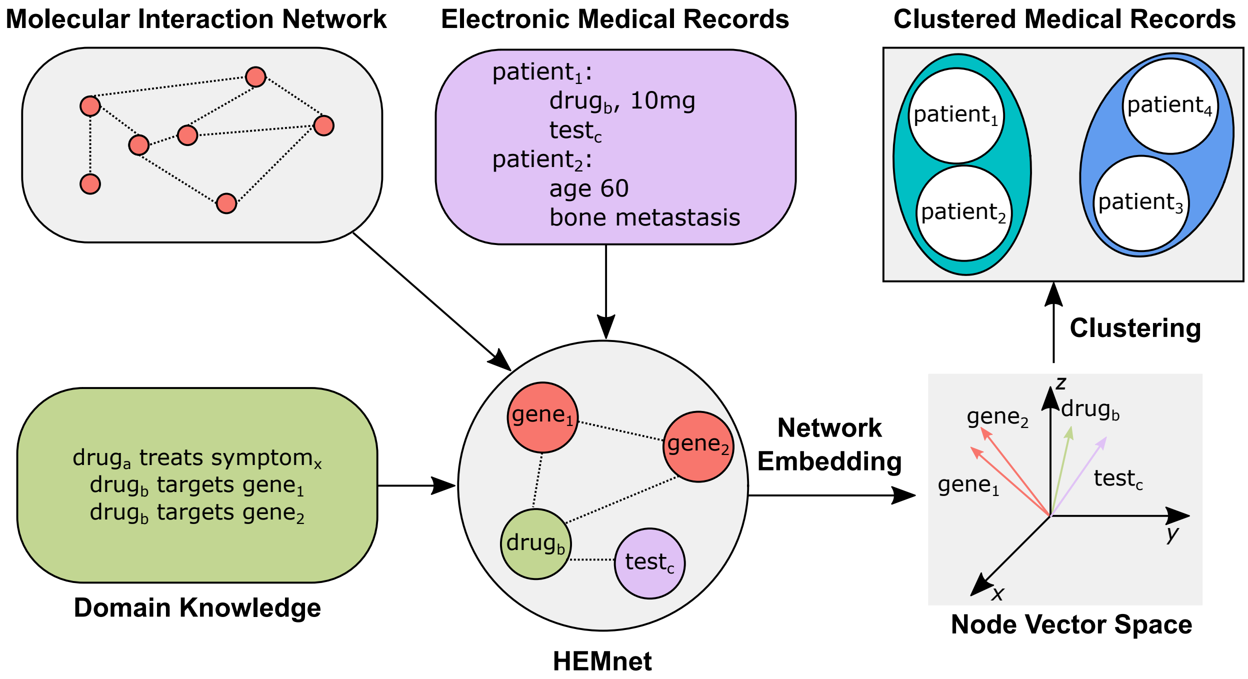


Figure 3.1: The pipeline for HEMnet, our proposed heterogeneous information network consisting of electronic medical records, molecular interaction networks, and domain knowledge. We trained network embedding vectors on HEMnet, obtaining low-dimensional vector representations for each node. With these vectors, we enhanced the original medical records and then clustered the patients into two groups.

a recently proposed embedding method, ProSNet [46], to infer relationships among its constituent nodes.

ProSNet was originally developed to discover functionally similar proteins from large biological networks across multiple species. It takes a heterogeneous network as input, on which it performs a novel dimensionality reduction algorithm to optimize a low-dimensional vector representation for each node. The vectors of two nodes are co-localized in the low-dimensional space if the nodes are close to each other in the heterogeneous network.

A key contribution is that ProSNet obtains low-dimensional vectors through a fast online learning algorithm instead of the batch learning algorithm used in previous works [23]. In each iteration, ProSNet samples a path from the heterogeneous network and optimizes the vectors based on this path instead of all pairs of nodes. Therefore, it can easily scale to large networks containing millions of edges and nodes, making it suitable for training on HEMnet. In our experiments, we chose our vectors to be of the recommended number of 500 dimensions.

3.1.3 Enhancing the Patient Profile Matrix

After generating low-dimensional vector representations of nodes in HEMnet, we tackled the problems of missing data and semantic mismatching in the patient records.

As in the previous chapter, we defined the raw patient profile matrix, P, to be an $n \times m$ matrix, where n was the number of patients and m was the number of features. In our experiments, $n = 133$ and $m = 449$. Unlike in PaReCat, where we used only binary values, each row here consisted of a patient profile with binary, categorical, discrete, or continuous values, depending on the corresponding feature. For example, drugs were continuous features because they were recorded as dosages, while metastasis sites were categorical features that included multiple locations in the body. For our particular dataset, we binarized the categorical features.

We generated the enhanced patient profile matrix, P', by multiplying P with another matrix, M, computed from pairwise embedding vector cosine similarity scores (Equation 2.1), and then normalized.

With this operation, we simultaneously solved the problems of missing data and semantic mismatching for each patient by filling in missing features that were highly similar to existing features. For example, a patient that had halitosis in his or her records would receive a non-zero value for bad breath, as these two symptoms were likely to have similar neighbors, and thus similar embedding vectors. The overview of this pipeline can be seen in Figure 3.1.

This operation is robust to extreme values. This is because min-max normalization does not allow features with naturally high values to dominate the enhancement process. For example, even if we were to multiply all values of a medical test by some arbitrary constant, then we would retain the normalized values of the feature, and P' would not change.

Additionally, this enhancement leads to interpretable results. For example, previously non-zero values of binary features such as symptoms might have a fractional value in the enhanced matrix, since min-max normalization guarantees the values to be between 0 and 1. These resulting values can be interpreted as the probability that the patient suffers from that symptom. However, non-binary values, such as medical tests, are not as interpretable in the enhanced matrix, which is a limitation of our approach. In the future, it would important to obtain more interpretable representations of patients for all features after enhancement.

3.2 RELATED WORK

Though our work was the first to utilize molecular interaction networks in conjunction with electronic medical records, it drew inspiration from many tasks and networks.

A previous work integrated genetics by linking EMR data to biobanked blood samples [47]. However, they used patient-matched DNA samples and patient genetic data rather than prior domain knowledge. Deep Patient is a previous work that also learns embedding vectors from EMRs [13]. However, it learns representations of patients rather than individual medical entities and does not use molecular information networks. Med2Vec is an algorithm that also learns efficient representations of medical concepts, but its joint optimization process relies on sequences of multiple visits [12], which is less applicable to inpatients. Word2vec, the inspiration for Med2Vec, provided the basis for much of the current research on learning efficient word embeddings [14]. However, word2vec can only be applied to patient records directly and cannot incorporate molecular interaction networks or domain knowledge.

Many related works have implemented data mining techniques on heterogeneous networks, such as bibliographic networks [8, 9], gene-phenotype networks [10], and social media networks [11]. However, none of these studies focused on EMR tasks. A previous study performed similarity searches within heterogeneous information networks [48]. It also exploited the idea of paths within a heterogeneous network, but instead used them for similarity computations rather than for vector optimization. Caballero and Akella also performed data imputation on EMR data [49]. However, they used an expectation-maximization method to fill out missing values rather than domain knowledge.

3.3 DATASET DESCRIPTION

The data we used was curated from a hospital that also provides traditional Chinese medicine (TCM) services. We chose this dataset for our experiments because semantic mismatches, caused by comorbid symptoms and functionally similar herbs, are even more prevalent in the TCM field, as discussed in the previous chapter. Thus, if our method could yield strong results on this dataset, it would also work well for EMR datasets in most other domains. All patients in the data were diagnosed with some type of non-small-cell lung carcinoma (NSCLC), including adenocarcinoma, squamous-cell carcinoma, adenosquamous carcinoma, and papillary adenocarcinoma. In addition to TCM herb prescriptions, patients received standard western medical treatments.

In our experiments, we only considered the first visit that each patient made to the hospital. Overall, we selected 133 patients with high degrees of missing data. Patients in the dataset had six feature types: medical history, medical test results, prescribed herbs, prescribed drugs, symptoms, and syndromes. Here, the term *syndrome* again denotes an underlying pattern that is specific to TCM [50]. In total, there were 449 unique features in the dataset. On average, each record was missing 407 of the 449 possible features, which provided a unique challenge of both missing data and semantic mismatching.

In addition, the database contained each patient's survival information as the number of months before the final event. A patient's final event was either hospital discharge or death. The data was right-censored (i.e., if there was no recorded death event, then death occurred at some unknown time after hospital discharge).

3.4 EXPERIMENTAL DESIGN

Typical cancer dataset studies might cluster patients based on a particular characteristic or set of characteristics (e.g., gene expression profiles) [51]. Thus, we identified features that discriminate patients who develop metastases or die from those who remain metastasis-free [52]. To this end, we clustered patients into two groups. Framing this as an unsupervised binary classification task in which a patient's medical record determined if he or she was likely to survive, a good clustering was thus one that had a cluster with a significantly better survival rate than the other.

In summary, our method generated HEMnet, trained network embedding vectors, and then created the enhanced patient profile matrix, P'. We compared the performance of our enhanced matrix with two baselines. The first baseline, which we refer to as the raw baseline, directly used the raw profile matrix, P. The second baseline, which we refer to as the mean imputation baseline, replaced each missing feature in the raw profile matrix with the average of the available values to generate a dense profile matrix, denoted by P_{mean}. We compared the performances of these three profile matrices in identical clustering tasks. In the following sections, we discuss cancer subtypes, the clustering process, and our evaluation methods.

3.4.1 Cancer Subtypes

Different cancer types are affected by different factors. For example, age has been shown to be negatively correlated with survival rates for patients with glioblastoma multiforme [53]. On the other hand, younger breast cancer patients statistically have worse prognoses than older patients [54]. Therefore, we separated the patients in our database into two groups: one of squamous-cell lung carcinoma (SCC) patients and the other of non-squamous-cell non-small-cell lung carcinoma (non-SQ NSCLC) patients. There were 43 SCC patients and 90 non-SQ NSCLC patients in our dataset. We performed experiments and analysis on these two cancer subtypes independently, but built HEMnet with all available records.

3.4.2 Clustering

We generated two clusters from each method's patient profile matrix. Note that the survival information was excluded from the clustering process. We did not necessarily use all features during clustering, but rather combinations of feature categories. For example, we clustered on profile matrices containing only drug features, profile matrices containing drug and medical history features, profile matrices containing all available features, etc. Using symptom and medical history features (a total of 57 features) yielded the best performance.

In addition, we performed dimensionality reduction with principal component analysis (PCA) to dampen the impact of highly correlated features. From the reduced profile matrix, we computed a dissimilarity matrix using pairwise cosine distance, which is commonly used to calculate the dissimilarity between two documents [55]. We ran k-means for two clusters on the resulting dissimilarity matrix.

3.4.3 Survival Analysis (Quantitative Evaluation)

In order to analyze the difference between the survival rates of two given clusters, we first computed the survival curves using the Kaplan-Meier estimator, one of the most frequently used methods in survival analysis [56]. After estimating the survival functions of both groups, we determined whether they were significantly different by comparing the two curves with the log-rank test [57]. The log-rank test computes a χ^2 statistic and a corresponding p-value to indicate if two survival functions are significantly different. We performed survival analysis with the R package `survival`[1]. Recall that a clustering is of high quality if one cluster possesses a significantly higher survival rate than the other.

When reporting the means of survival functions, we used a restricted mean. Since final events were not always deaths in the dataset, the corresponding survival curve estimates did not necessarily go to zero, which resulted in undefined means. Thus, we set the upper limit to be some constant u, so that the restricted mean signified the number of months out of the first u months that each group was expected to experience [58].

3.4.4 Feature Analysis (Qualitative Evaluation)

Throughout our discussion, we denote the cluster with longer survival c_{long} and the cluster with shorter survival c_{short}. After obtaining c_{long}, we wished to identify the features that were responsible for its higher survival rate. Although we only used a subset of features

[1] https://CRAN.R-project.org/package=survival

Table 3.2: SCC patient survival time statistics with HEMnet.

	Cluster Size	**Restricted Mean**	**Restricted Mean Standard Error**	**Median**
c_{long}	28	20.7	2.01	22.7
c_{short}	15	11.0	1.60	11.0

during the clustering phase, we analyzed all features that appeared in the EMRs. This is because all features were used in HEMnet to train embedding vectors, so they may have indirectly impacted the clustering.

We identified interesting features by computing an unpaired t-test for each feature, where one set of samples came from c_{long} and the other set came from c_{short}. The features with significant p-values (< 0.01) were candidates for further exploration.

3.5 RESULTS AND DISCUSSION

We discuss the results of the survival and feature analyses below. First, we discuss the SCC patients, then the non-SQ NSCLC patients.

3.5.1 HEMnet substantially improved stratification on SCC patients

After clustering squamous-cell carcinoma patients into two groups using the HEMnet-enhanced profile matrix, P', c_{long} had 28 patients and c_{short} had 15 patients (Figure 3.2a). By the log-rank test, c_{long}'s survival function was significantly better than c_{short}'s with a χ^2 statistic of 10.79 (p-value = 0.001020). We show the cluster summary statistics in Table 3.2, with $u = 33.3$.

In contrast, when clustering on the raw profile matrix, P, c_{long} had 25 patients and c_{short} had 18 patients (Figure 3.2c). Their survival functions were not significantly different at the 1% significance level with a χ^2 statistic of 6.214 (p-value = 0.01267).

When clustering on the mean-imputed profile matrix, P_{mean}, c_{long} had 25 patients and c_{short} had 18 patients (Figure 3.2e). Here, the survival functions were not significantly different with a χ^2 statistic of 2.489 (p-value = 0.1147).

After computing an unpaired t-test for each feature, we received a p-value denoting how different the feature's values were in c_{long} from its values in c_{short} (Table 3.3). In the table, we place an asterisk after features that were identified by our method but not by the raw baseline. The mean imputation baseline only identified five features, all of which were identified by the raw baseline. In both the raw baseline and the mean imputation baseline,

Table 3.3: Significant SCC patient features with HEMnet (p-value < 0.01). Features not found to be significant by the raw baseline have asterisks. Feature means (μ) and standard deviations (σ) are shown.

Feature Name	p**-value**	c_{long} μ	c_{long} σ	c_{short} μ	c_{short} σ	**Feature Type**
Karnofsky performance status	4.427×10^{-7}	77.86	6.739	62.67	9.978	Medical Test
Loss of appetite	2.691×10^{-6}	0.2857	0.6999	1.600	0.8794	Symptom
Shortness of breath	2.440×10^{-4}	0.6786	0.8886	1.733	0.7717	Symptom
Fatigue	2.687×10^{-4}	0.4286	0.8207	1.467	0.8844	Symptom
S. tuberosa	1.407×10^{-3}	0.03571	0.1856	1.800	2.857	Herb
P. suffruticosa	2.730×10^{-3}	0	0	1.067	1.879	Herb
Umbrella polypore*	3.847×10^{-3}	0.1429	0.7423	1.267	1.806	Herb
Sulphurweed*	4.562×10^{-3}	0.1071	0.4087	1.667	2.891	Herb
Sputum*	5.015×10^{-3}	0.8214	0.7585	1.533	0.8844	Symptom
A. tataricus*	5.678×10^{-3}	0.1071	0.3093	3.667	6.925	Herb
Cough*	9.388×10^{-3}	1.321	0.8474	1.933	0.5735	Symptom

umbrella polypore, sulphurweed, sputum, and *A. tataricus* had p-values of 0.01513, 0.01507, 0.02150, and 0.01352, respectively. Cough had a p-value of 0.1056 in the raw baseline and a p-value of 0.3069 in the mean imputation baseline.

Our method missed one feature deemed significant by the raw baseline: TNM staging system, which describes the stage of the cancer's progression [59]. Since there were many missing values in our dataset, the raw baseline failed to leverage a variety of features, relying on this single feature. However, our method did not have this limitation, and also achieved a more significant separation between the two clusters. Additionally, it missed no features identified by the mean imputation baseline.

Of all significant features, only Karnofsky performance status had a higher mean in c_{long} than in c_{short}. This was expected, as a higher KPS score indicates a relatively healthier cancer patient [60]. All other significant features had lower means in c_{long} than in c_{short}. This was reasonable for symptom features, as patients with longer survival rates generally have less severe symptoms. On the other hand, it might initially appear as if the herbs' higher values in c_{short} indicate their ineffectiveness. However, this can be explained by the fact that patients with more life-threatening conditions require higher dosages of treatments.

The raw baseline did not find sputum and cough to be symptoms that were statistically significantly different between c_{long} and c_{short}, while our method did. However, they are very common symptoms in lung cancer patients, and tend to be more severe in patients with lower survival rates. Furthermore, the other three features that our method discovered were all herbs. We can interpret these features' higher values in c_{short} as potential herb-symptom relationships. In the herb-symptom dictionary, sulphurweed was listed as a treatment for both cough and sputum, while umbrella polypore treated only urinary tract-related symp-

Table 3.4: Non-SQ NSCLC patient survival time statistics with HEMnet.

	Cluster Size	**Restricted Mean**	**Restricted Mean Standard Error**	**Median**
c_{long}	52	23.9	2.67	18.6
c_{short}	38	14.1	2.18	12.7

toms and *A. tataricus* simply did not appear. Despite these relationships not appearing in the domain knowledge, HEMnet was able to capture them in our method. In fact, a study showed that a naturally occurring compound derived from umbrella polypore mycelia induces apoptosis in human lung cancer cells [61]. Furthermore, a recent study showed that a polysaccharide isolated from *A. tataricus* inhibits the growth of cancer cells [62].

Our method was able to capture meaningful herb-symptom relationships while the baselines were not. Thus, we have shown that HEMnet can integrate external herb information in the form of herb-protein targets and protein-protein interaction information.

3.5.2 HEMnet also substantially improved stratification on non-SQ NSCLC patients

After clustering non-SQ NSCLC patients into two sets using P', c_{long} had 52 patients and c_{short} had 38 patients (Figure 3.2b). By the log-rank test, c_{long}'s survival function was statistically significantly better than c_{short}'s with a χ^2 statistic of 8.449 (p-value = 0.003652). We show the cluster summary statistics in Table 3.4, with $u = 50.8$.

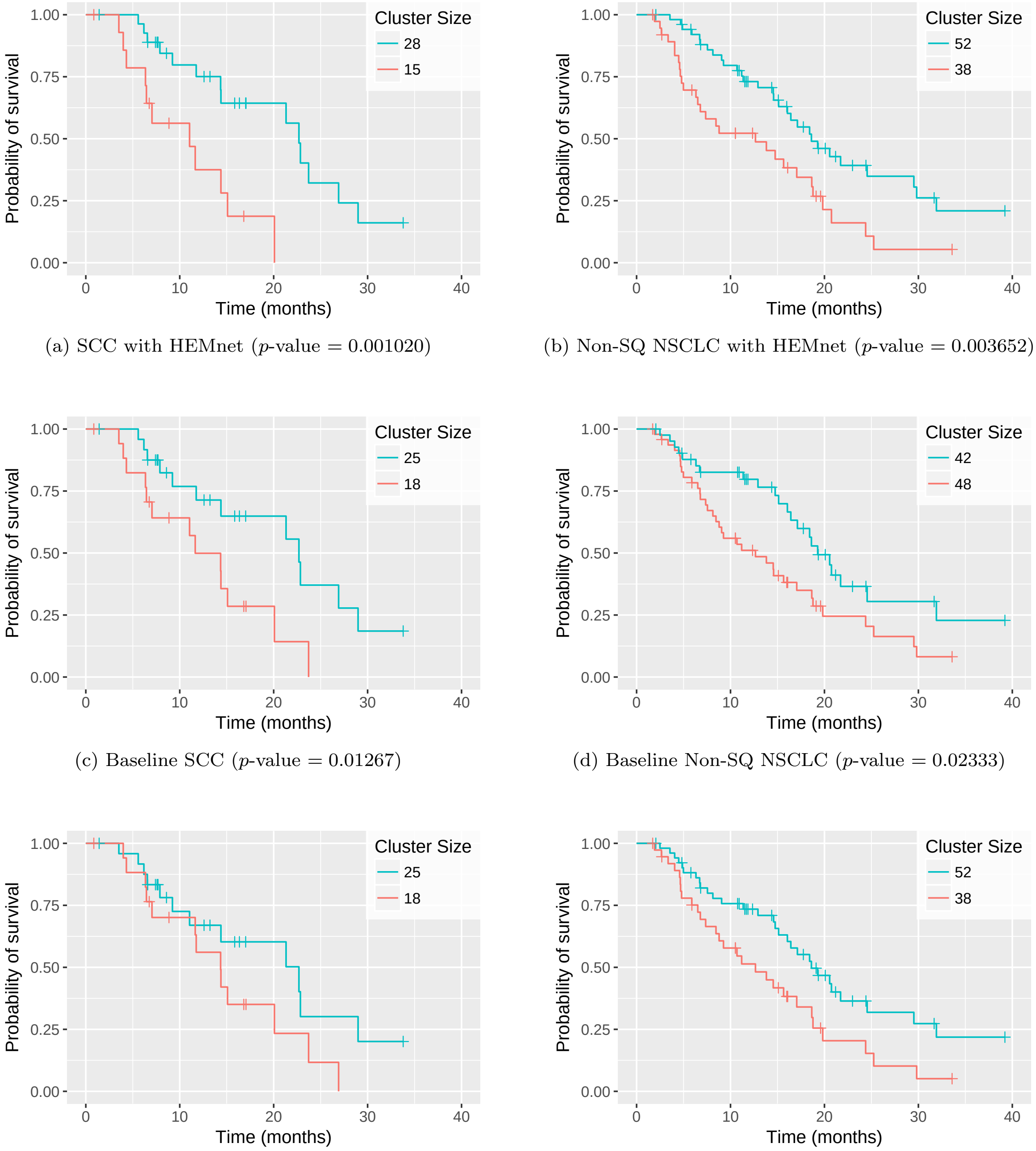
In contrast, when we clustered on P, c_{long} had 42 patients and c_{short} had 48 patients (Figure 3.2d). Again, the survival functions were not significantly different at the 1% significance level with a χ^2 statistic of 5.144 (p-value = 0.02333).

When we clustered on P_{mean}, c_{long} had 52 patients and c_{short} had 38 patients (Figure 3.2f). The survival functions were not significantly different with a χ^2 statistic of 6.115 (p-value = 0.01341).

Again, we computed an unpaired t-test for each feature, receiving a corresponding p-value to indicate significance (Table 3.5). Consistent with earlier results, Karnofsky performance status was the only feature to have a higher mean in c_{long} than in c_{short}.

In the raw baseline, pamidronate disodium injection, atelectasis, white peony, shortness of breath, Chinese cinnamon, and *A. asphodeloides* had p-values of 0.04887, 0.01208, 0.05991, 0.021107, 0.06895, and 0.03629, respectively. Our method obtained p-values below 0.01 for each of these features. The mean imputation baseline did identify pamidronate disodium injection as a significant feature (p-value = 0.008043), but did not identify any herb features.

Because our method found several treatments to be significant, it allowed us to ex-



(a) SCC with HEMnet (p-value = 0.001020)

(b) Non-SQ NSCLC with HEMnet (p-value = 0.003652)

(c) Baseline SCC (p-value = 0.01267)

(d) Baseline Non-SQ NSCLC (p-value = 0.02333)

(e) SCC with Mean Imputation (p-value = 0.1147)

(f) Non-SQ NSCLC with Mean Imputation (p-value = 0.01341)

Figure 3.2: A comparison of survival functions for two patient clusters in SCC patients (subfigures (a), (c), and (e)) and non-SQ NSCLC patients (subfigures (b), (d), and (f)). Subfigures (a) and (b) used the HEMnet-enhanced profile matrices, subfigures (c) and (d) used the baseline profile matrices, and subfigures (e) and (f) used the baseline profile matrices with mean imputation.

Table 3.5: Significant non-SQ NSCLC patient features with HEMnet (p-value < 0.01). Features not found to be significant by the raw baseline have asterisks. Feature means (μ) and standard deviations (σ) are shown.

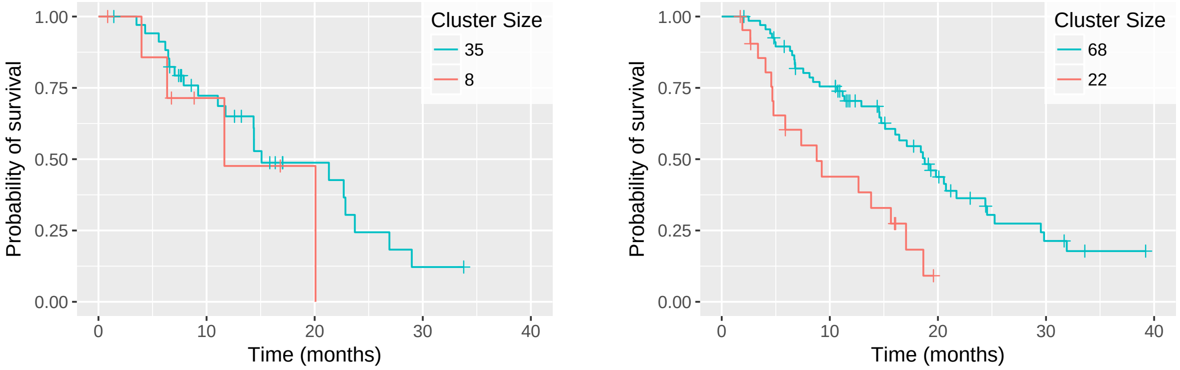
Feature Name	**p-value**	c_{long} μ	c_{long} σ	c_{short} μ	c_{short} σ	**Feature Type**
Karnofsky performance status	5.570×10^{-8}	76.35	8.993	63.42	11.98	Medical Test
Pamidronate disodium injection*	2.196×10^{-7}	0.2115	0.8166	15.34	19.74	Drug
Bone metastasis	1.294×10^{-6}	0.03846	0.1923	0.4211	0.4937	Medical Test
Atelectasis*	1.250×10^{-3}	0.01923	0.1373	0.2105	0.4077	Symptom
TNM staging system	1.307×10^{-3}	7.346	2.472	8.632	0.6250	Medical Test
Distant metastasis	2.307×10^{-3}	0.3846	0.4865	0.6842	0.4648	Medical Test
White peony*	8.760×10^{-3}	1.115	2.584	2.789	3.894	Herb
Shortness of breath*	8.999×10^{-3}	0.6346	0.8555	1.105	0.9676	Symptom
Chinese cinnamon*	9.302×10^{-3}	0.03846	0.2747	0.4211	1.091	Herb
A. asphodeloides*	9.952×10^{-3}	0.05769	0.4120	1.158	3.273	Herb

tract meaningful relationships. Pamidronate disodium injections are well-known treatments for patients with bone metastases [63], which was also a significant feature (p-value $= 1.294 \times 10^{-6}$). Atelectasis and shortness of breath are common symptoms among lung cancer patients. White peony roots have been shown to inhibit tumor growth in non-small-cell lung cancer patients [64]. Lastly, a study has shown *A. asphodeloides* to have anti-tumor effects [65]. As with the SCC patients, our method captured meaningful herb-symptom relationships while the baselines could not. We attribute this to the external information integrated into HEMnet.

3.5.3 KPS Feature Significance

Because Karnofsky performance status had the lowest p-value for both cancer types with our method, we further explored its significance. Performance status scores are assigned to cancer patients in attempts to quantify their overall health and well-being. It was therefore reasonable that KPS was a discriminative feature for patients of both cancer subtypes. Specifically, the Karnofsky score runs from 100 to 0, where 100 is relatively perfect health and 0 is death [60]. In this dataset, Karnofsky performance scores were assigned in standard intervals of 10.

To test whether KPS could accurately classify patients by itself, we placed patients into c_{long} if they had a Karnofsky score greater than 60 and the rest into c_{short} (Figure 3.3). Although using only KPS significantly separated non-SQ NSCLC patients (p-value $= 5.571 \times 10^{-4}$), the survival curves were not significantly different in SCC patients (p-value $= 0.3655$). Thus, we concluded that KPS alone could not accurately predict a patient's survival rate, and that the HEMnet-enhanced profile matrices leveraged a variety of features to achieve



(a) SCC (p-value = 0.3655)

(b) Non-SQ NSCLC (p-value = 5.571×10^{-4})

Figure 3.3: A comparison of survival functions when separating patients only by their Karnofsky performance score.

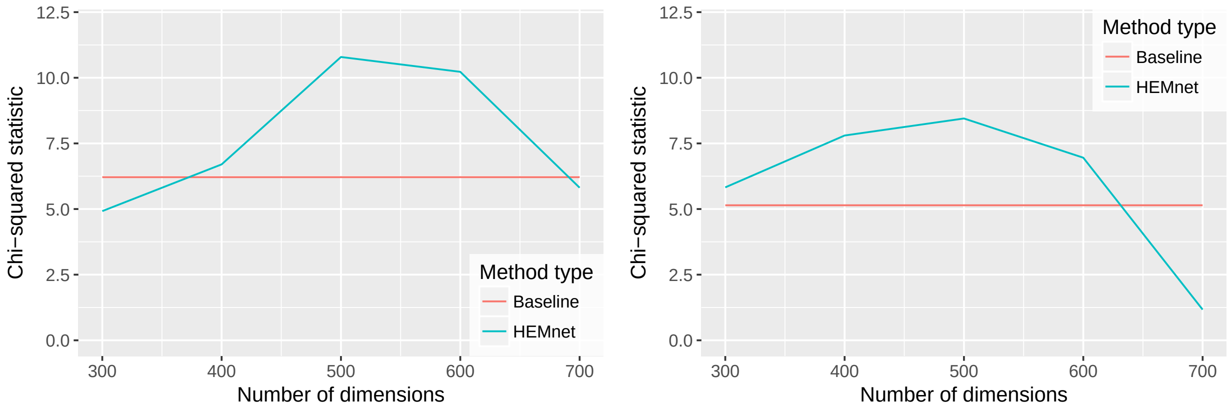
better clustering results.

3.5.4 Parameter Tuning

In our framework, we needed to tune the number of dimensions of each embedding vector, which was recommended to be 500 in the original ProSNet paper. Tuning this parameter, we saw that the optimal results were also achieved around 500 dimensions (Figure 3.4). Furthermore, 400 and 600 dimensions produced statistically significant results for both cancer subtypes.

3.6 CONCLUSIONS AND FUTURE WORK

In this chapter, we integrated EMRs with molecular interaction networks and domain knowledge using a heterogeneous medical record network, HEMnet, to solve two challenges in analyzing EMRs, i.e., missing data and semantic mismatching. By extracting knowledge from HEMnet, we allowed for the training of accurate embedding vectors. We showed how we can use these vectors to enhance EMR databases, and then evaluated their performance based on survival prediction. Our method was able to perform better than the baseline profile matrix with and without mean imputation by successfully splitting patients of both cancer subtypes (squamous-cell lung carcinoma and non-squamous-cell non-small-cell lung carcinoma). We showed this via quantitative evaluation by performing survival analysis on the resulting cluster pairs of each cancer subtype. Lastly, we verified the effects of HEMnet



(a) SCC

(b) Non-SQ NSCLC

Figure 3.4: Results of tuning the number of dimensions of the embedding vectors to observe parameter sensitivity.

with qualitative analysis, studying the features and their differences between healthy and unhealthy clusters.

One limitation that we noted is the fact that the HEMnet is unweighted. If we were to add low-confidence edges (e.g., text mining relationships), then we must incorporate edge weights. One possible method is to sample paths in ProSNet with a probability proportional to the edge weight. This would allow us to rely more on high-confidence neighbors when computing embedding vectors.

Another concern is that the knowledge graph might introduce noise to a complete patient record. A possible solution is to regularize the influence of the knowledge graph by weighting similarity scores with the sparsity of the patient record. Thus, the knowledge graph has less influence on more complete records.

It is also unclear how to intuitively interpret the enhanced patient record. Even though we may view the non-zero values filled in for some fields as predicted values for those fields, there is no guarantee that the values are in the valid range, such as in the case of medical tests. However, our proposed strategy can be implemented in other ways to improve interpretability. For example, we could utilize a probabilistic approach to fill in missing features. We can normalize each row in the similarity matrix such that each feature can be interpreted as a probability distribution over all the other features. We can then use an algorithm like expectation-maximization to maximize the probability of certain missing values taking on existing values from similar features. This would be a very interesting future direction to explore.

For future work, we can also implement tasks for clinical prediction of survival (CPS).

Given the framework proposed in this work, we can attempt to predict a new patient's probability of survival by first grouping him or her into the most similar cluster, and then extrapolating the survival rate from the neighboring patient records (similar to the k-nearest neighbors algorithm). Furthermore, HEMnet allows for even more sophisticated methods of feature discovery, including novel drug-drug interactions. This feature discovery would follow along the lines of our qualitative evaluation.

CHAPTER 4: INTEGRATING EXTERNAL KNOWLEDGE INTO ELECTRONIC MEDICAL RECORD VISUALIZATION

Despite advances in computational methods, such as those discussed in the previous chapters, EMR systems can still greatly benefit from human interpretation, which allows for exploratory analysis and more control over decision-making [66]. However, human interaction with EMR systems has been hindered. In a previous study, 37% of participants reported that interacting with their EMR databases was too time consuming [67]. Another study showed that when using EMRs, nurses face challenges that can threaten quality and safety of care [68]. Both of these shortcomings can be addressed with information visualization, which can aid doctors in processing and understanding complex, high-dimensional EMR data.

In particular, EMR visualization in a two-dimensional space is useful for observing and interpreting patient clusters. Coherent clusters may elucidate a patient's most significant characteristics by visualizing his or her proximity to successfully diagnosed patients [69]. For example, thoracic aortic dissections are commonly misdiagnosed as acute myocardial infarctions (MIs). Misdiagnosis in these cases is extremely harmful, as patients with aortic dissections treated for MIs have mortality rates similar to that of untreated patients. Despite this risk, patients with aortic dissections have a misdiagnosis rate of 39% [70]. Fortunately, the most telltale signs of aortic dissection (age, onset of pain, and syncope) are readily available in EMRs [71, 72]. An effective visualization would utilize these features to place an undiagnosed aortic dissection patient near similar patients, reducing the chance of misdiagnosis.

Unfortunately, designing an effective visualization system is a complicated task, as EMRs are high-dimensional sources of data that consist of thousands of features. Furthermore, the prevalence of missing data and semantic mismatching in EMRs, as discussed in the previous chapter, make it even more challenging to correctly group together similar patients, leading to poor or even misleading visualizations.

To address these challenges, I developed **Vis**ualization **A**ssisted by Knowledge **G**raph **E**nrichment (VisAGE), a method that enhances patient records with a knowledge graph built from external databases, building upon PaReCat and HEMnet. These databases included protein-protein interactions, genomic data, and drug-chemical associations. I continued with the idea that performing network embedding on the knowledge graph allows for the inference of associations among different types of data, which can alleviate data sparsity in EMRs. A major novelty of VisAGE was that it was the first to use all of these data sources in EMR visualization. In the rest of this chapter, I describe the dataset, the details of VisAGE, and the evaluation process.

4.1 DATASET DESCRIPTION

While VisAGE is a general method that can be applied to any set of EMRs, we chose the Parkinson's Progression Markers Initiative (PPMI) dataset [73] for the evaluation process. The PPMI dataset contained a mix of Parkinson's disease (PD) patients as well as control patients suffering from other diseases. We chose this dataset for two reasons: (1) it contained many feature types, and (2) Parkinson's disease is a complicated disorder the causes of which have been attributed to complex combinations of genetic and environmental factors [74]. We only considered the 1,579 patients with baseline visits. This dataset included 6,013 biospecimen, genetic, drug, symptom, diagnosis, medical test, and demographic features, which fit our statement that EMRs are high-dimensional. Feature types included binary, numerical, and categorical features. We binarized the categorical features. On average, each patient only had 261 of the 6,013 available features, which supported our previous assertion that EMRs are typically sparse.

4.2 PATIENT PROFILE MATRIX

From the PPMI dataset, we generated an $n \times m$ patient profile matrix, denoted by P as in the previous chapters, where n was the number of patients and m was the number of features. Here, $n = 1,579$ and $m = 6,013$. Existing visualization methods use P directly as input. However, as previously stated, P is typically sparse, and thus suboptimal for visualization. The main idea of VisAGE is to enhance the profile matrix before visualization by leveraging associations inferred from a knowledge graph. The enhanced profile matrix, denoted by P', then replaces P as input to any visualization method. We later show that P' gives better visualizations in several applications on our dataset.

4.3 VISAGE OVERVIEW

Our proposed framework for VisAGE consists of three steps (Figure 4.1). The first constructs a knowledge graph with external data sources and EMRs. The second performs embedding on the constructed graph to learn a similarity matrix, denoted by M. Lastly, the third step multiplies the patient profile matrix, P, by the similarity matrix, M, to obtain the enhanced patient profile matrix, P'.

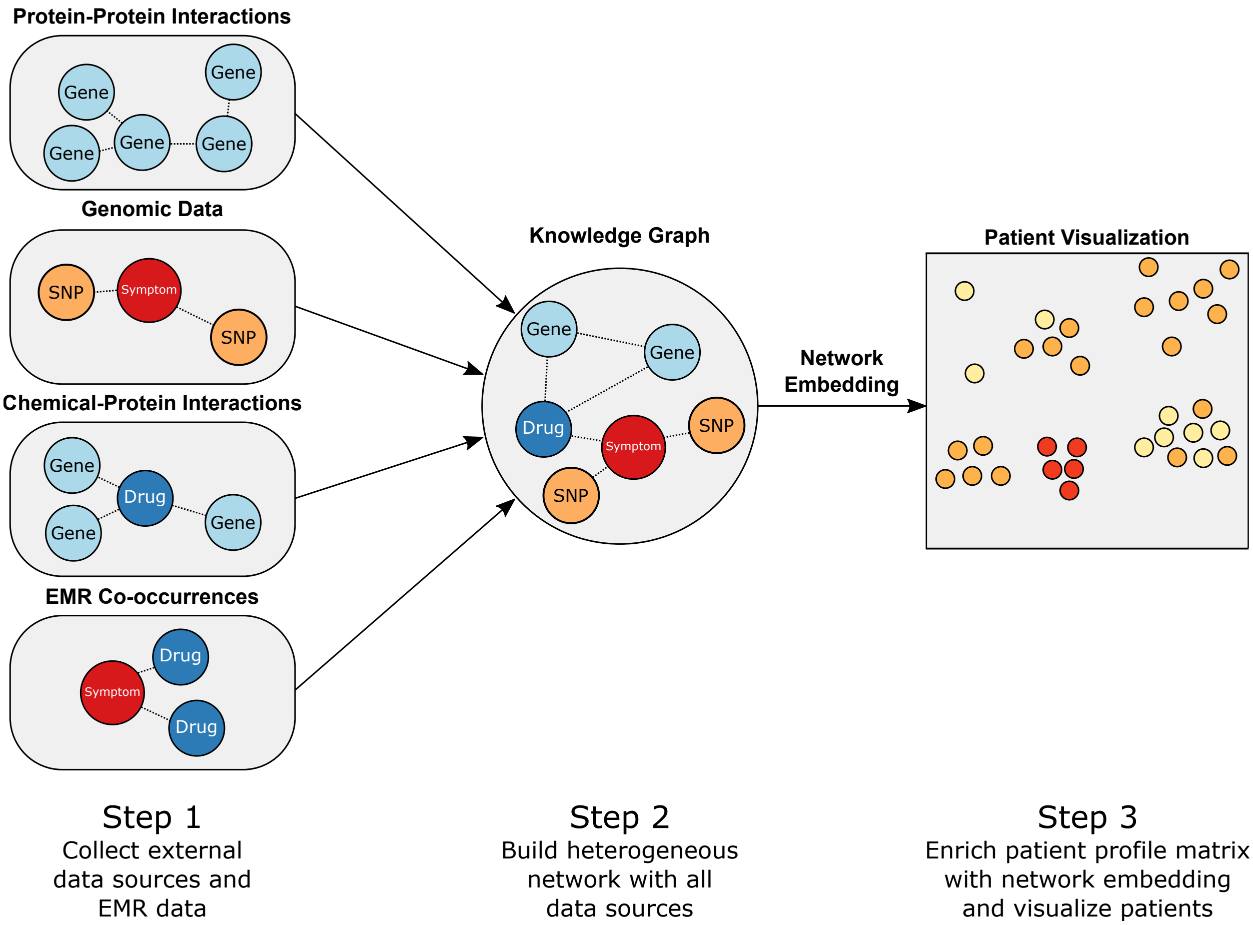

Protein-Protein Interactions
Gene
Gene
Gene
Gene
Gene
Genomic Data
SNP
Symptom
SNP
Chemical-Protein Interactions
Gene
Drug
Gene
Gene
EMR Co-occurrences
Drug
Symptom
Drug
Knowledge Graph
Gene
Gene
Drug
Symptom
SNP
SNP
Network Embedding
Patient Visualization
Step 1
Collect external data sources and EMR data
Step 2
Build heterogeneous network with all data sources
Step 3
Enrich patient profile matrix with network embedding and visualize patients

Figure 4.1: The VisAGE pipeline. It first creates a knowledge graph from multiple data sources. It then performs network embedding on this knowledge graph, enhances the patient profile matrix, and then visualizes each patient in a two-dimensional space.

4.3.1 Knowledge Graph Construction

The knowledge graph is a heterogeneous network containing edges from four data sources. It builds upon the previous chapter's HEMnet, adding genomic information. However, we used a different protein-protein interaction (PPI) database, added genomic data, and changed the herb-protein network to a drug-protein network, since the PPMI dataset contained only western drugs.

1. **Protein-protein interaction network.** We used the inBioMap database [75] of protein-protein interaction (PPI) edges. For a functional linkage between two proteins p_1 and p_2, we created a node for p_1, a node for p_2, and an undirected edge $\{p_1, p_2\}$ in the network. There were 17,327 proteins and 606,194 edges in this network.

2. **Single-nucleotide polymorphism enrichment.** We integrated genomic data in

the form of single-nucleotide polymorphisms (SNPs), which are single variations in the human genome. We identified SNPs that were highly enriched in PD patients in the dataset by using a one-sided Fisher's exact test [76]. Overall, we found 3,900 SNPs with p-values < 0.05. We then selected the nonsynonymous SNPs and determined if specific symptoms were enriched in SNPs with another one-sided Fisher's exact test. For each PD-enriched SNP g, we created a node for g and an edge $\{g, s\}$ if s was significantly enriched in g with a p-value < 0.01. There were 34,324 SNP-symptom edges.

3. **Chemical-protein interaction network.** We used STITCH, a database of known and predicted interactions between chemicals and proteins [77]. STITCH included computationally predicted associations in addition to those aggregated from other databases. For each drug d in the EMR data, if d's active ingredient interacted with a protein p in the STITCH database, then we created a node for d, a node for p, and an undirected edge $\{d, p\}$. There were 7,218 drug-protein edges in this network.

4. **Electronic medical records.** We directly added co-occurrence edges from each medical record. For example, if a patient was diagnosed with symptom s and prescribed a drug d, then we created a node s, a node d, and an undirected edge $\{s, d\}$. We repeated this for all elements in each patient's medical record.

The resulting network contained 23,886 nodes and 17,108,116 edges. The purpose of the knowledge graph was to again utilize the idea of guilt by association: although many related medical concepts may not have directly co-occurred in any medical records, they may have indirectly shared neighbors in the knowledge graph through the protein-protein, SNP-symptom, and drug-protein edges.

4.3.2 Similarity Matrix Learning

We again used ProSNet to infer relationships among entities in the knowledge graph [46]. After generating the embedding vectors, we enhanced the patient profile matrix. Again, we constructed an $m \times m$ similarity matrix, denoted by M, where m was the number of features and each entry was the cosine similarity between the corresponding features' low-dimensional vectors.

Afterwards, we multiplied the raw profile matrix, P, with M to retrieve the enhanced patient profile matrix, denoted by P' (Equation 2.1).

4.4 RELATED WORK

A previous study visualized high-dimensional data with a technique called LargeVis. However, it built a k-NN network directly from the data, and then reduced the network to two dimensions without using external information [7]. Another study built upon LargeVis to visualize single cells, but still also directly computed embeddings from a k-NN network without utilizing external data [78]. Marlin *et al.* visualized a pattern discovery model's clustering parameters in the context of EMR analysis [79]. However, they focused on longitudinal data and predicting mortality outcomes rather than patient clustering. Gotz *et al.* performed interactive visualization of EMR data, but worked with time series data to analyze patterns over time [80]. The Dynamic Icons (DICON) system clusters EMRs that are similar to a given patient, visualizing the clusters. However, it does not utilize molecular interaction networks or genomic data to compute similarities between EMRs [81]. Lastly, Perer *et al.* developed Care Pathway Explorer to visualize EMR data to investigate correlations with patient outcomes [82]. However, their system uses sequential pattern mining, which relies on historical EMR data to extract patterns.

4.5 RESULTS AND DISCUSSION

We wished to determine whether the enhanced profile matrix, P', would lead to better visualization results than the original patient profile matrix, P. Thus, we compared them in practical downstream visualization applications. Specifically, following previous studies [7, 78], we used t-distributed stochastic neighbor embedding (t-SNE) [83], an algorithm that can efficiently model high-dimensional objects as two-dimensional points, which made it especially well-suited for visualizing our dataset. We generated our visualizations by running t-SNE with default settings on P for the baseline and P' for VisAGE. For both methods, this created a new $n \times 2$ matrix, so that each patient was reduced to two dimensions. We plotted this matrix as a set of points. We now discuss our results when visualizing the raw and enhanced patient profile matrices in various applications.

4.5.1 Two-dimensional visualization with UPDRS

Using the two-dimensional representations of patient records, we labeled each record according to its unified Parkinson's disease rating scale (UPDRS) scores. The UPDRS consists of six sections, each containing survey questions that evaluate a patient's physical and mental condition [84]. The questions deal with topics ranging from anxiety to sleeping problems, with scores scaled from 0 to 4. A higher score indicates more severe impairment or disability. As in previous work, we labeled each patient with the sum of his or her UPDRS scores [85].

The main difference is that moderately impaired patients (orange circles) on the right side of the plots were less structured in the baseline visualization (Figure 4.2a), but were clustered more distinctly in the VisAGE visualization (Figure 4.2b). The most severe Parkinson's disease patients are marked by red squares, and were also clustered more tightly in the VisAGE visualization than in the baseline. We attribute the baseline's worse performance to the data sparsity of the EMRs.

4.5.2 Qualitative evaluation: drug and symptom enrichment

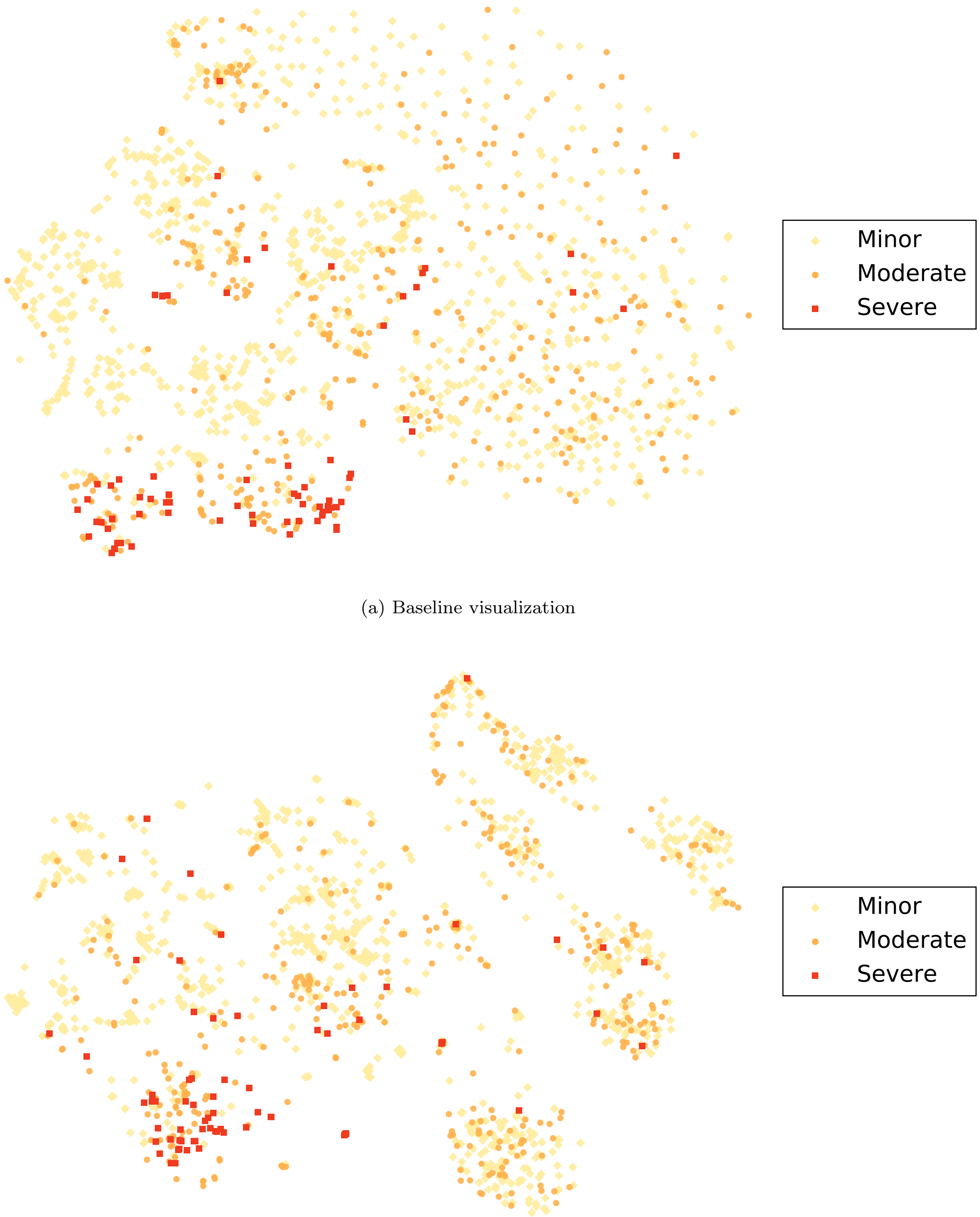
We qualitatively evaluated the visualization results by computing drug and symptom enrichments for each cluster. We used symptoms and drugs because they are strongly connected to patient statuses and diagnoses. Thus, if a cluster was highly enriched in a symptom or drug, then doctors could interpret the patients more coherently. Recall that in the previous chapter, we used HEMnet to obtain two clusters within each cancer type to perform survival analysis. Here, we analyzed an arbitrary number of clusters to view PD patients that require special treatment.

We first clustered the two-dimensional patient representations with DBSCAN [86], which is robust to outliers and does not need to specify the number of clusters. Because the PPMI dataset contained control patients to simulate noise, DBSCAN's robustness to outliers was especially desirable. Additionally, not having to specify the number of clusters *a priori* was useful for our application, as we did not know the exact number of PD clusters that require special treatment.

For the DBSCAN parameters, we set $\epsilon = 1$ and `minPts` = 10. In the baseline method, patients were placed into 10 clusters. With VisAGE, patients were placed into 18 clusters. For each cluster c and each symptom or drug b, we computed Fisher's exact test to determine if c was significantly enriched in b. We only used symptoms and drugs with binary values to avoid medical tests for which all patients had non-zero values (e.g., the Epworth sleepiness scale [87]).

4.5.3 VisAGE discovered more interpretable patient clusters

We show the two-dimensional plot of patients in Figures 4.3 and 4.4, with each color-shape combination corresponding to a unique cluster generated by DBSCAN. We also show the two most enriched symptoms for each cluster in the legends. With the baseline, many of the



(a) Baseline visualization

(b) VisAGE visualization

Figure 4.2: The two-dimensional representations of patient records, plotted with color labels determined by each record's UPDRS scores. VisAGE's visualization identified more clusters for moderately impaired patients, and more tightly grouped severely impaired patients.

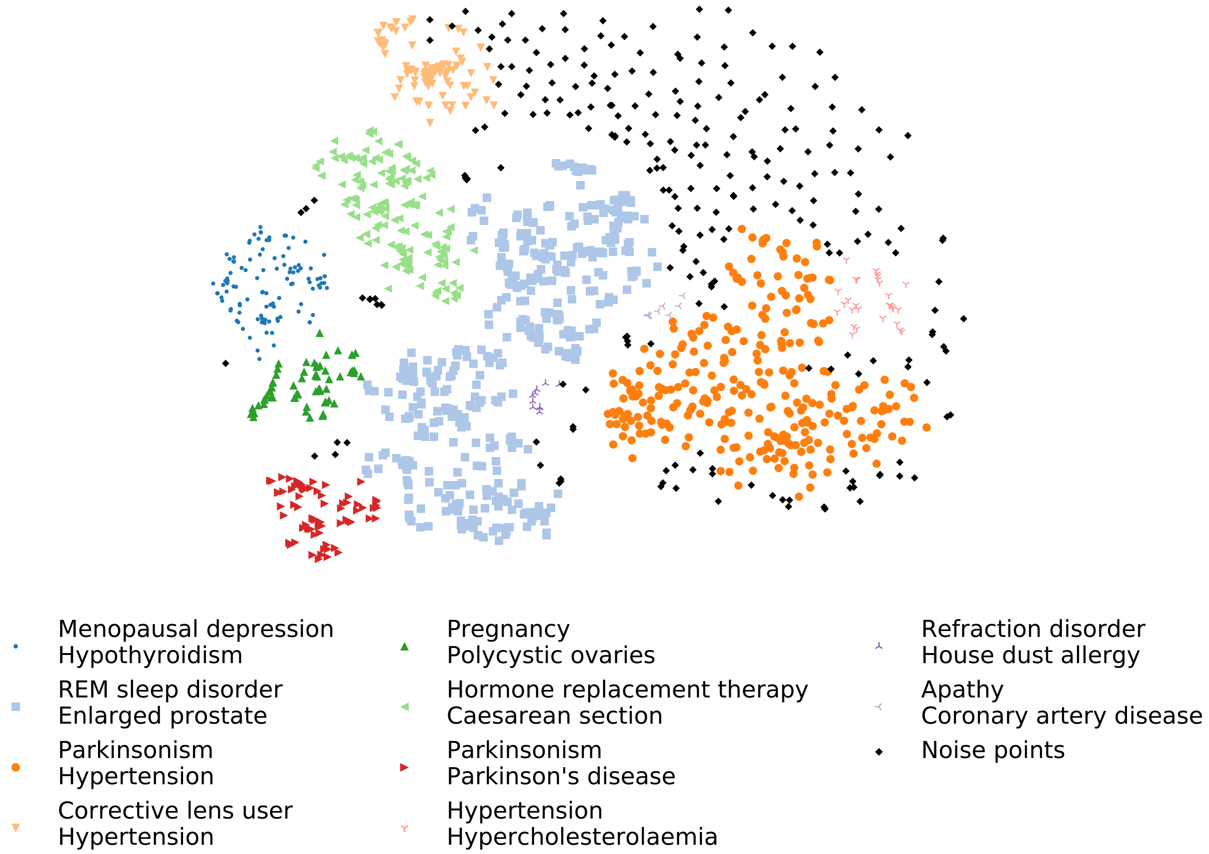


Figure 4.3: The baseline's two-dimensional representation of patient records, with color-shape combinations determined by the DBSCAN clustering. We show each cluster's two most enriched symptoms to indicate a PD cluster requiring special treatment.

points in the upper right quadrant of the plot were determined to be noise (black circles). As a result, no patients could be deemed to be similar to these noise points. On the other hand, VisAGE was able to properly classify many of these patients into distinct clusters.

We saw that both methods mostly grouped together the patients with the highest UPDRS scores (Figure 4.2). The corresponding DBSCAN clusters that overlapped with these high-UPDRS patients were most enriched in parkinsonism and Parkinson's disease, as expected.

Both methods identified a cluster of patients enriched in parkinsonism and hypertension (orange circles in the baseline and dark green triangles pointing up in VisAGE). Indeed, hypertension is commonly known to be prevalent in PD patients [88]. However, VisAGE identified four additional clusters that were significantly enriched in PD/parkinsonism and another informative symptom. On the other hand, the baseline method mixed these clusters into larger ones, losing information in the process. We interpreted these additional clusters as PD clusters that required special treatment.

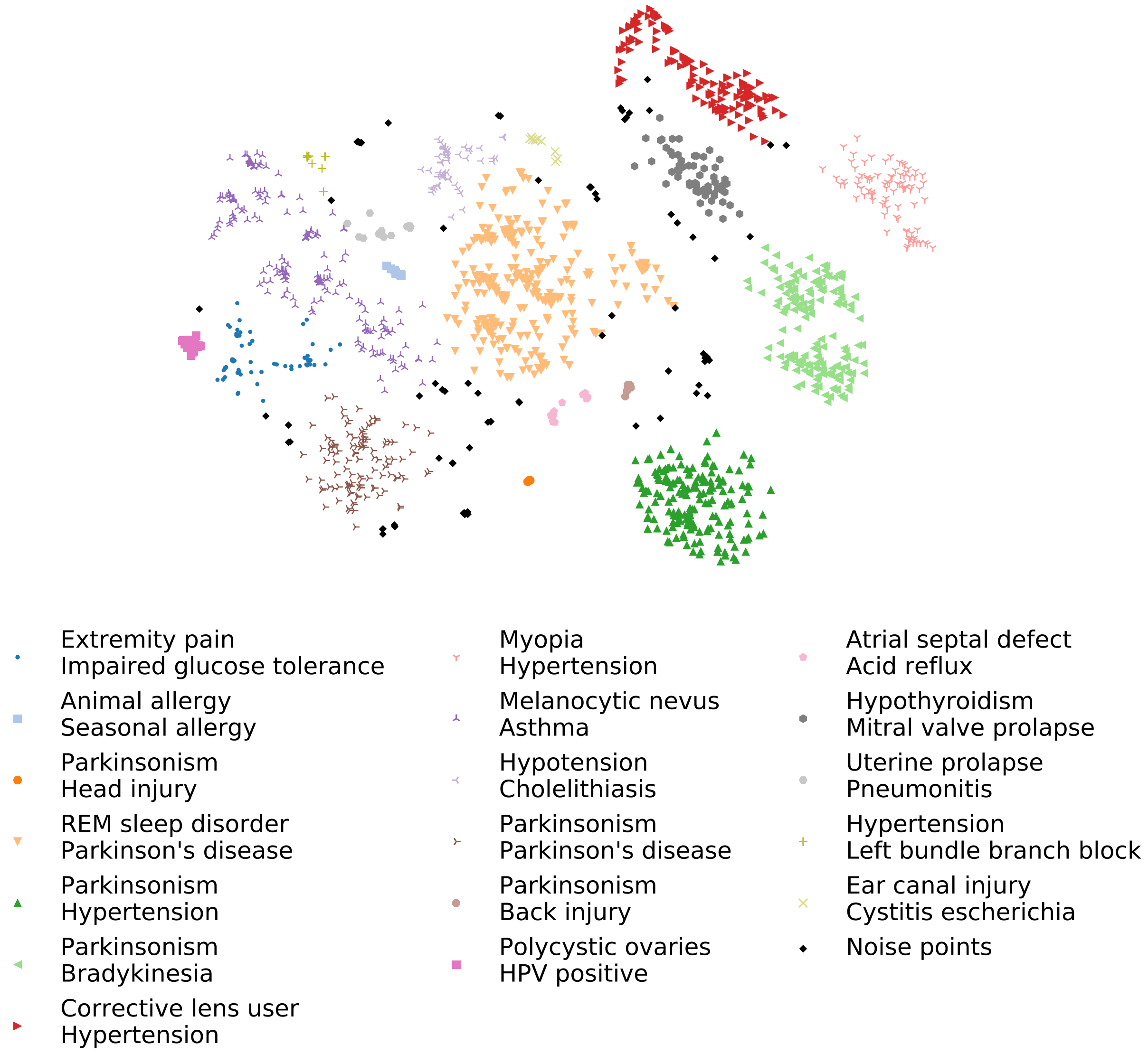


Figure 4.4: VisAGE's two-dimensional representations of patient records, with color-shape combinations determined by the DBSCAN clustering. We show each cluster's two most enriched symptoms to indicate a PD cluster requiring special treatment.

We now discuss these four clusters.

1. **Parkinsonism and head injury.** The cluster of dark orange circles contained 16 patients, and was enriched in parkinsonism and head injury with p-values of 3.110×10^{-4} and 0.01013, respectively. This is consistent with previous work, as head trauma is one of the most common candidates for PD causes [89]. This cluster was highly enriched in entacapone, levodopa, and carbidopa with p-values of 1.085×10^{-25}, 9.030×10^{-10}, and 7.099×10^{-5}, respectively. While levodopa/carbidopa (LC) is the most common drug prescribed to PD patients, entacapone is often prescribed as a supplementary drug to improve the efficacy of LC [90]. As expected, these patients are also labeled as *Severe* in Figure 4.2, which explains the need for this supplement. Furthermore, entacapone has been proposed as a possible treatment for traumatic brain injury [91]. In the baseline, this group of patients was lost in the cluster most enriched in parkinsonism and hypertension.

2. **REM sleep disorders and Parkinson's disease**. The cluster of light orange, down-pointing triangles contained 292 patients, and was enriched in rapid eye movement (REM) sleep behavior disorder, which is most often associated with PD (p-values of 1.477×10^{-5} and 7.246×10^{-3}, respectively) [92]. In addition to the standard levodopa prescription (p-value $= 9.787 \times 10^{-23}$), the cluster was also highly enriched in clonazepam (p-value $= 0.004377$). Clonazepam administered with levodopa at bedtime has been shown to reduce REM sleep disorder symptoms [93]. In the baseline, the corresponding cluster contained nearly twice as many patients (458), and was not highly enriched in Parkinson's disease.

3. **Parkinsonism and bradykinesia.** In VisAGE's visualization, the cluster of light green, left-pointing triangles contained 159 patients, and was enriched in parkinsonism and bradykinesia with p-values of 1.526×10^{-8} and 3.974×10^{-8}, respectively. As expected, bradykinesia is a key symptom of parkinsonism [94]. Additionally, this cluster was highly enriched in ropinirole with a p-value of 1.246×10^{-7}. Ropinirole stimulates mesolimbic D_3 receptors, which alleviates bradykinesia [95]. In the baseline, this group of patients was mixed with patients exhibiting parkinsonism and hypertension.

4. **Parkinsonism and back injury**. The cluster of light brown circles contained 17 patients, and was enriched in parkinsonism and back injury with p-values of 1.320×10^{-5} and 1.092×10^{-4}, respectively. A previous study showed that spinal cord injuries are associated with increased risk of PD [96]. In addition to the standard levodopa/carbidopa prescription, this cluster was significantly enriched in amantadine (p-value $= 6.09 \times$

10^{-5}). Amantadine is not only an antiparkinsonian agent, but has also been shown to act as a non-competitive N-Methyl-D-aspartate (NMDA) receptor antagonist [97]. NMDA receptor antagonists have been shown to treat acute spinal cord injuries [98]. Like in the cluster that was enriched in parkinsonism and head injury, this cluster contained many patients with severe UPDRS scores. In the baseline, these patients were again mixed with the cluster enriched in parkinsonism and hypertension.

4.5.4 Quantitative evaluation: false discovery rate

For each method, we compared the number of clusters highly enriched in drugs and symptoms. To this end, we excluded drugs and symptoms from both patient profile matrices. Additionally, we excluded these features from VisAGE's knowledge graph in order to limit data leakage. We then recomputed the enrichments for drugs and symptoms, taking the drug or symptom with the lowest p-value to represent each cluster. With these p-values, we counted the number of clusters that were significantly enriched in at least one drug or symptom.

To create a fair comparison, we used the Benjamini-Hochberg procedure [99] to control the false discovery rate at different levels of α (Figure 4.5). We saw that VisAGE identified more enriched clusters than the baseline at every level of α, which was consistent with our earlier observation that the baseline method was incapable of distinguishing among patients with less severe symptoms. Thus, we concluded that VisAGE also performed better quantitatively.

4.6 CONCLUSIONS AND FUTURE WORK

In this chapter, we presented VisAGE, a method of improving EMR visualization by enhancing EMRs with external knowledge sources. Evaluations on a PD patient dataset showed that VisAGE can generate visualizations such that similar patients are clustered together more tightly than in a baseline that does not alter the original database. We also evaluated our visualization with enrichments of drugs and symptoms, and showed that VisAGE can produce a higher quantity of fine-grained partitions of PD patients.

One limitation of this work is that the evaluation was done on only one dataset, which was mainly due to the necessity of expensive patient annotations. In the future, it is important to further evaluate the proposed enhancement method on more datasets as they become available. We can also build software that can implement our visualization in real application environments. Since VisAGE is a general method, the software would serve as a framework

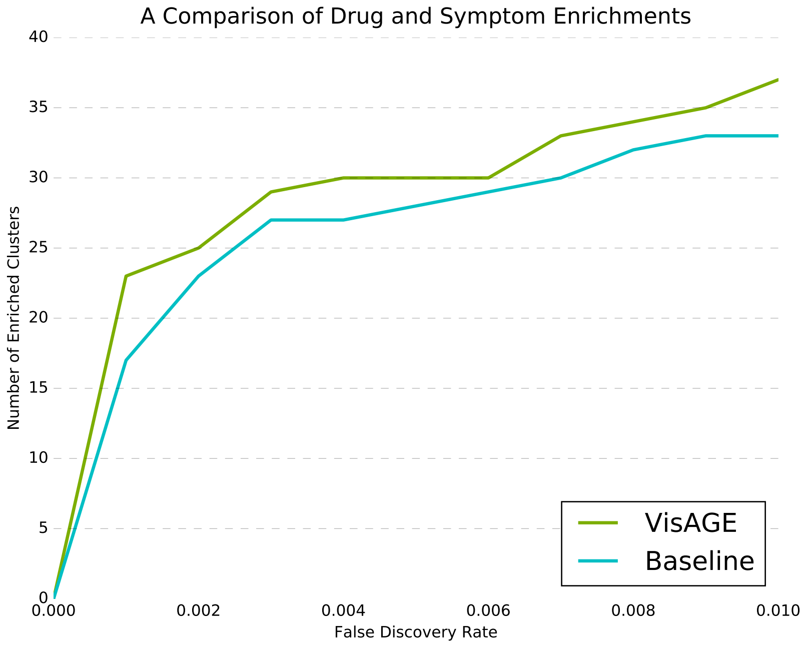


Figure 4.5: A plot comparing the baseline and VisAGE. VisAGE dominated the baseline in the number of clusters enriched with at least one drug or symptom at each level of α.

for an interactive component that can enhance any EMR database. For example, in a clinical setting, previously treated patients can serve as guidelines for doctors treating new patients. Doctors can identify these similar, previously treated patients in the two-dimensional space using the visualization tool and optimize treatment for the current patient.

It would also be informative to examine the clusters in our visualization that did not have obvious interpretations based on prior medical knowledge. This could allow us to observe previously unknown, but meaningful patterns. For example, one cluster that our method identified was most enriched in uterine prolapse and pneumonitis. While the connection between these two symptoms is unclear, the statistical significance of the enrichments would make it interesting to further examine this cluster, as it might help discover new medical knowledge.

CHAPTER 5: FRAMING ELECTRONIC MEDICAL RECORDS AS POLYLINGUAL DOCUMENTS IN QUERY EXPANSION

As EMR databases are further introduced into daily usage, retrieval systems are increasing in importance to aid doctors in processing and understanding the large amounts of data. In particular, one important application is to efficiently parse medical records and identify those that are most relevant to a new patient. Standard information retrieval systems (e.g., web search engines) receive string queries as input, compute numeric scores that determine how well each database document matches the query, and output a ranked list of documents. Ideally, a doctor can query a system with a new patient's symptoms and receive a set of relevant patient records. These records can serve as an informative baseline to prescribe suitable treatments for the new patient.

However, due to synonyms that occur in the medical vocabulary, the original search query may not be complete, and thus may not retrieve optimal results. These synonyms contribute to the semantic mismatching problem discussed in Chapter 3. There are three categories of synonyms of interest to medical record retrieval:

1. **Semantic synonyms** are medical terms that have identical meanings. For instance, halitosis and fetor oris are semantic synonyms because they are different terms that refer to the same symptom. Because doctors will typically record only one of these, queries that contain "halitosis" will not properly match records that contain "fetor oris", and vice versa. Semantic synonyms can be mined with natural language processing techniques and straightforward statistical measures [100].

2. **Treatment synonyms** are drug-symptom pairs in which the drug treats the symptom. For example, ibuprofen and fever are treatment synonyms. Treatment synonyms can be obtained by mining medical publications [101].

3. **Functional synonyms** are terms that are not identical, but co-occur more frequently than random. Arthritis and hypertension are functional synonyms due to their comorbid relationship (a decade-long study showed that nearly half of elderly arthritis patients also suffered from hypertension) [102]. Thus, if a hypothetical query consists of only the term "arthritis" to describe an elderly arthritis patient with hypertension, the retrieved patient records may not contain treatments optimally suited for the query patient. Functional synonyms can be inferred from treatment synonyms.

If a query consists entirely of symptoms, then the relevant semantic and functional synonyms are also symptoms. We show that augmenting an original patient query with all three synonym types can capture relevant but mismatched documents, thus improving retrieval performance.

In prior work, Zeng *et al.* performed query expansion in a similar medical record retrieval setting using synonyms and topic models [103]. Their synonym-based query expansion utilized the Unified Medical Language System (UMLS), which is a compendium of biomedical vocabularies, to map query terms to their semantic synonyms. We refer to this method as dictionary-based query expansion to avoid confusion. On the other hand, their topic-model-based query expansion trained on patient records to jointly find all three synonym types. However, using this standard topic model, symptoms and treatments were grouped together with no distinction in the medical records, which may have decreased performance.

Rather than indiscriminately mine these three types of synonyms, we separately modeled the symptom and treatment synonyms. Our approach was based on traditional monolingual topic models, but instead viewed the symptoms and treatments of a medical record as generated by distinct languages. Thus, output topics were aligned across the two languages and contained synonyms of all three types. This is because symptoms in the same topic were likely to be semantic or functional synonyms, while symptoms and treatments in aligned topics were likely to be treatment synonyms. We then augmented the original query with synonyms of the query symptoms during retrieval. Our proposed method was the first to model symptoms and treatments as separate languages in electronic medical records. We also compared with two embedding methods that jointly mine all three synonyms.

We evaluated our approach on the same traditional Chinese medicine (TCM) medical record collection as in Chapter 2. We chose this dataset because comorbid symptoms and functionally similar herbs are prevalent in the TCM field. Thus, if our method could improve retrieval performance on this dataset, it would also work well for EMR datasets in other domains. We showed that our method performed better than baseline methods as well as state-of-the-art embedding methods in query expansion experiments.

5.1 RETRIEVAL PROBLEM FORMULATION

Given a database of patient records $R = \{r_1, \ldots, r_n\}$, the ith patient record r_i consists of a set of diseases D_i, a set of symptoms S_i, and a set of treatments T_i. From this database, we wish to retrieve the set of patient records most relevant to some new patient, p_{new}, who is not in the database. To achieve this, we first reformulate p_{new}'s symptoms as a query. Thus, given p_{new}'s set of symptoms $S_{new} = \{s_1, \ldots, s_j\}$,

$$Q_{new} = S_{new} = \{s_1, \ldots, s_j\} \tag{5.1}$$

By performing query expansion on Q_{new}, we add query terms to better match relevant records and thus improve the retrieval performance:

$$Q'_{new} = \{s_1, \dots, s_j, q_1, \dots, q_l\} \tag{5.2}$$

Here, $\{q_1, \dots, q_l\}$ is the set of expansion terms that are added to the original query. Although the original query Q_{new} only contains symptoms, the expansion terms can contain both symptoms and treatments. Expansion terms can be obtained with a variety of methods, which we discuss in the next section.

We hypothesized that the expanded query, Q'_{new}, would retrieve more relevant documents because in practice, Q_{new} is usually not comprehensive. In our medical setting, this is analogous to situations in which the list of symptoms that a doctor identifies in a new patient is incomplete, which may be due to a combination of two major factors.

1. The doctor uses one of many possible synonyms, including semantic, treatment, and functional synonyms, to describe a patient's condition.

2. The database is incomplete, so a query symptom may simply not appear in existing medical records, resulting in poor query matches.

We expected the first factor to have a larger impact on query quality, particularly due to unique variations of symptoms that are prevalent in TCM.

5.2 METHODS

With each technique that we used in our experiments, we conducted query expansion to improve retrieval performance. We augmented queries with synonym terms selected by each method, and then performed the retrieval on the existing database of medical records.

Overall, we used five different methods of query expansion. First, we addressed two baselines used in previous work [103]: dictionary-based query expansion and topic-model-based query expansion. In our dataset, the dictionary-based method incorporated an external treatment-symptom knowledge graph (Chapter 2.2.1) to add manually curated treatment synonyms. The topic-model-based method trained topics on the patient record database to add expansion terms that co-occur in the same topics as the given query. Although the previous study used a third method, predicate-based query expansion, we did not utilize this method due to a lack of high-quality TCM ontology databases. Furthermore, the predicate-based method did not outperform the topic-model-based method in prior work.

Next, we explored two network embedding techniques: Med2Vec and diffusion component analysis (DCA). Med2Vec is an embedding algorithm that learns efficient representations of medical records and concepts by using EMR datasets. We explored DCA in Chapter 2. The key difference between these two methods is that DCA depends on external medical knowledge.

Lastly, we discuss our method, which mines semantic, treatment, and functional synonyms by considering symptoms and treatments to be separate languages in a topic model.

5.2.1 Dictionary-Based Query Expansion

Dictionary-based query expansion utilizes a ground-truth, treatment-symptom TCM dictionary. We constructed a knowledge graph, in which an undirected edge $\{t, s\}$ indicated that a treatment t treated a symptom s in the dictionary. There were 1,995 treatments, 1,635 symptoms, and 27,824 treatment relations in the dictionary, which translated to a total of 3,630 nodes and 27,824 edges in the resulting knowledge graph. There were no treatment-treatment or symptom-symptom edges. To perform query expansion on a query Q_{new}, we added all treatments from the knowledge graph that were directly connected to at least one symptom in Q_{new}.

5.2.2 Topic-Model-Based Query Expansion

In prior work, topic-model-based query expansion performed the best in a similar medical record retrieval task in terms of recall and F-measure [103]. Specifically, the authors used latent Dirichlet allocation (LDA) [104] to derive topics from their database of EMRs. With LDA, each document is characterized by a mixture of topics. In turn, each topic consists of mixtures of words. In our study, we also used LDA to train topics from the dataset.

After training k topics, from topic i's per-word distribution, ϕ_i, we referred to the set of 100 words with the highest probabilities as H_i. For a query Q_{new}, we then performed the following multiplication:

$$\phi_i' = |Q_{new} \cap H_i| \cdot \phi_i \tag{5.3}$$

With this operation, we scaled each word's probability in ϕ_i by the number of query terms that were in the top 100 words of ϕ_i. Finally, we summed each word's probabilities across the scaled distributions, $\sum_{i=1}^{k} \phi_i'$, and received a new weight for each word. We empirically chose to identify five topics from our dataset. The top five words with the highest weights were designated expansion terms.

5.2.3 Med2Vec-Based Query Expansion

Med2Vec is a state-of-the-art embedding method designed specifically for EMRs [12]. It discovers efficient representations of medical codes (symptoms and treatments, in the case of our dataset). Med2Vec's optimization function is similar to that of word embedding methods that use the skip-gram model, such as word2vec [105]. The authors stated three major reasons for word2vec's failure to accommodate medical data:

1. Healthcare datasets have unique structures in which the visits are temporally ordered, but the medical codes within a visit form an unordered set.
2. Learned representations should be interpretable.
3. The algorithm should be scalable to handle real-world datasets of millions of patients.

In particular, the first reason was of greatest relevance to our experiment setting. Med2Vec maximizes the likelihood of observing a medical code (symptom or treatment) given the codes in the same visit. In other words, a medical code's vector representation predicts its neighboring medical codes. By obtaining vector representations of all medical codes as well as computing their pairwise similarities, Med2Vec can jointly discover semantic, treatment, and functional relationships.

We ran Med2Vec on our training corpus and obtained a low-dimensional vector representation for each symptom and treatment in the dataset. Given a query $Q_{new} = \{s_1, \ldots, s_j\}$, we computed the cosine similarity between each query term s_i's Med2Vec representation and every non-query term's Med2Vec representation. Thus, for each candidate expansion term, we summed j similarity scores, one for each query term. We took the five terms with the highest score sums as expansion terms.

5.2.4 DCA-Based Query Expansion

Like Med2Vec, DCA can jointly mine all three synonym types. We used the network constructed in dictionary-based query expansion as the input to DCA. After learning vector representations for each node in the network, we computed cosine similarity scores as in Med2Vec-based query expansion, again taking the top five terms with the highest score sums as expansion terms.

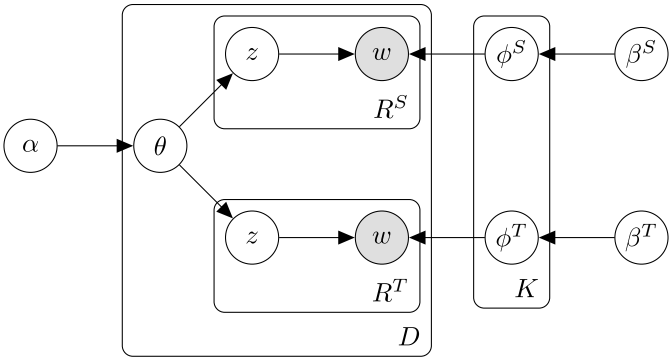


Figure 5.1: The plate notation of our proposed model, framing electronic medical records as bilingual documents. A variable's superscript S indicates symptoms and T indicates treatments. α and β are the parameters of the Dirichlet priors on the per-document topic distributions and the per-topic word distributions, respectively. θ_d is the topic distribution for document d. ϕ_k^S and ϕ_k^T are each language's corresponding word distributions for topic k. z is the latent topic assignment for each observed word w.

5.2.5 BiLDA-Based Query Expansion

In our data, symptoms and treatments were labeled and separated in each patient record. We hypothesized that a topic model that explicitly captures this structure would improve performance over standard, monolingual topic models.

Polylingual topic modeling (PLTM) finds latent cross-lingual topics in a multilingual corpus [106]. These text collections can either be direct translations or theme-aligned [107]. Direct translations occur in sentences of two documents that are translations of one another. An example of a direct translation is *Romeo and Juliet* in English and Chinese. On the other hand, theme alignment occurs in documents that are not necessarily direct translations, but discuss the same topics in similar sections. An example of theme alignment is the Wikipedia pages on *Romeo and Juliet* in English and Chinese.

Our method considers EMRs to consist of two separate "languages": symptoms and treatments. Thus, patient records are theme-aligned in the sense that a patient's symptoms and treatments are generated by the same set of diseases. Furthermore, the symptoms and treatments are varied according to the same syndromes, which are the underlying factors in TCM that we have previously discussed. Standard monolingual topic models are unable to represent these separate "languages", since symptoms and treatments are grouped together. This removes the ability to differentiate, and therefore translate, between the two.

The output of PLTM is a set of cross-lingual topics, including per-document topic distributions and per-topic word distributions in each of the languages. This model assumes that each topic consists of a discrete distribution of words for each language. Thus, there are two language-specific topics Φ^S and Φ^T, each of which is drawn from its own symmetric Dirichlet distribution with parameters β^S and β^T, respectively.

Next, we discuss the generative process. Each EMR is represented as a mixture over topics, and is generated by first sampling from an asymmetric Dirichlet prior with concentration parameter α and base measure m:

$$\theta \sim \mathrm{Dir}(\theta, \alpha m) \tag{5.4}$$

Then, a latent topic assignment is drawn for each word in the corresponding language (i.e., symptoms and treatments).

$$z^S \sim P(z^S \mid \theta) = \prod_r \theta_{z_r^S} \tag{5.5}$$

$$z^T \sim P(z^T \mid \theta) = \prod_r \theta_{z_r^T} \tag{5.6}$$

The individual symptoms and treatments are then drawn using language-specific topic parameters.

$$w^S \sim P(w^S \mid z^S, \Phi^S) = \prod_r \phi^S_{w_r^S \mid z_r^S} \tag{5.7}$$

$$w^T \sim P(w^T \mid z^T, \Phi^T) = \prod_r \phi^T_{w_r^T \mid z_r^T} \tag{5.8}$$

With two languages, PLTM reduces to Bilingual Latent Dirichlet Allocation (BiLDA) (Figure 5.1). We obtained k joint topics that align k symptom topics and k treatment topics. As with monolingual LDA, we experimented with different values of k, finding $k = 5$ to yield the best results. To train topics with BiLDA, we used the **MA**chine **L**earning for **L**anguag**E** **T**oolkit (MALLET) [108], which performs inference with Gibbs sampling. We conducted query expansion the same way we performed LDA-based query expansion, selecting the five terms with the highest weights.

5.3 EVALUATION

We evaluated and compared the five different query expansion methods as well as the baseline with no query expansion by performing retrieval on our dataset. We first describe our dataset, then discuss the evaluation process in the following two sections.

5.3.1 Data Description

We used the gastroenterology dataset from Chapter 2. The same doctor treated all patients in the record, which had the advantage of consistent treatment, but the simultaneous disadvantage of potentially systematic errors or incompleteness. All patients had some variety of stomach illness. Each record contained a detailed list of symptoms, treatments, and diseases. Each patient had an average of 9.08 symptoms and 1.63 diseases. The disease information was used as ground truth labels in the evaluation stage, and was therefore not included when finding query expansion terms. We elected to use only the first visit for each patient to prevent cases in which a patient's query returned other visits of the same patient. This left us with 3,750 patient visits.

5.3.2 Cross-Validation

We split our dataset into 10 random training and test sets as per k-fold cross-validation. Thus, the training records were functionally a database of EMRs. Each held-out test patient was then regarded as a new, unseen patient. For each of the test patients, we retrieved relevant patient documents from the training set.

Each test set contained 375 patient records. We excluded all details from the test set except for symptoms. Using a given test patient's symptom set as a query, we performed each query expansion method in three ways: adding symptoms, adding treatments, and adding both. We refer to these methods as **symptom expansion**, **treatment expansion**, and **mixed expansion**, respectively. The only exception to this was the dictionary-based query expansion, which was only capable of treatment expansion.

5.3.3 Retrieval Tests

For each query in the held-out test set, we performed medical record retrieval. To score a document in the training corpus given a query patient, we used Okapi BM25 as our ranking function, which is one of the most effective retrieval methods [109]. We defined the Okapi BM25 score of a document D given a query patient $Q = \{q_1, \ldots, q_n\}$ as

$$\text{Score}(D, Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)} \tag{5.9}$$

In our experiments, $f(q_i, D)$ was always 1 if q_i appeared in D, since no patient record contained duplicate symptoms or treatments. $|D|$ was the length of document D, and `avgdl` was the average document length in the training corpus. For the symptom expansions and the baseline with no query expansion, D contained only symptoms. For treatment and mixed expansions, D contained all symptoms and treatments of the patient. In the absence of parameter optimization, we chose the default values of $k_1 = 2$ and $b = 0.75$. Additionally, the inverse document frequency was defined as

$$\text{IDF}(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \tag{5.10}$$

where N was the total number of documents in the training corpus, and $n(q_i)$ was the number of training documents containing the term q_i. With this ranking function, we returned a ranked list of retrieved documents given a query Q.

5.3.4 Relevance Measure

To evaluate the performance of each retrieval task, we assigned an objective measure of relevance to a retrieved patient given a query patient. Conveniently, the list of diseases the doctor assigned to each patient was recorded in our dataset. We used these disease lists as ground truth labels for the corresponding patients. Thus, we defined the gain of a document D given a query patient Q to be the following:

$$\text{Gain}(D, Q) = \frac{|D_{disease} \cap Q_{disease}|}{|D_{disease}||Q_{disease}|} \tag{5.11}$$

Here, $D_{disease}$ and $Q_{disease}$ refer to the set of diseases belonging to D and Q, respectively. In the traditional vector space model, this gain is the cosine similarity between the document and query vectors, which is a useful metric for determining similarity between two documents [55]. Thus, we used normalized discounted cumulative gain (NDCG), a standard method of evaluating search engines [110], to compute the quality of our ranked list. The DCG at a particular rank k, for query Q, which returns a ranked list of $D_1, \ldots, D_N$ is defined as

$$\text{DCG@}k = \sum_{i=1}^{k} \frac{\text{Gain}(D_i, Q)}{\log_2(i+1)} \tag{5.12}$$

where $\text{Gain}(D_i, Q)$ was defined in Equation 5.11. NDCG@k is defined as DCG@k divided by the DCG of the ideal ranked list for query Q, thus making it a metric comparable across queries and suitable for our 10-fold framework. We show results for $k \in \{5, 10, 15, 20\}$. We excluded precision, recall, and the F-measure due to their inability to incorporate rankings.

Table 5.1: Retrieval results for various query expansion methods. Bolded values indicate the highest NDCG@k. BiLDA mixed expansion performed best for all choices of k.

Expansion Method	**NDCG@5**	**NDCG@10**	**NDCG@15**	**NDCG@20**
No query expansion	0.1673	0.1675	0.1674	0.1677
Dictionary	0.1633	0.1647	0.1652	0.1659
LDA (symptoms)	0.1686	0.1682	0.1681	0.1690
LDA (treatments)	0.1689	0.1669	0.1667	0.1677
LDA (mixed)	0.1690	0.1671	0.1668	0.1679
Med2Vec (symptoms)	0.1636	0.1637	0.1648	0.1652
Med2Vec (treatments)	0.1682	0.1673	0.1684	0.1677
Med2Vec (mixed)	0.1678	0.1671	0.1685	0.1675
DCA (symptoms)	0.1518	0.1534	0.1556	0.1560
DCA (treatments)	0.1689	0.1702	0.1712	0.1719
DCA (mixed)	0.1510	0.1537	0.1557	0.1565
BiLDA (symptoms)	0.1709	0.1706	0.1713	0.1716
BiLDA (treatments)	0.1684	0.1681	0.1681	0.1679
BiLDA (mixed)	**0.1752**	**0.1739**	**0.1747**	**0.1736**

5.4 RESULTS AND DISCUSSION

The results of the evaluation are shown in Table 5.1. In order to analyze the significance of the NDCG values, we performed the paired t-test on the ranking metrics between pairs of expansion methods.

BiLDA-based mixed query expansion achieved the best retrieval performance among all expansions. For NDCG@5, 10, 15, and 20, it performed better than the baseline with no query expansion with p-values of 2.842×10^{-3}, 4.784×10^{-3}, 6.852×10^{-7}, and 1.929×10^{-6}, respectively. Furthermore, BiLDA mixed expansion performed better than all of the runner-up methods at the 5% significance level.

Mixed expansion was only the best-performing expansion type for BiLDA. This is due to the fact that all other methods do not separately mine symptom and treatment synonyms. On the other hand, BiLDA-based query expansion considers symptoms and treatments to be from separate topics, and therefore it successfully added in mixed query terms.

Dictionary-based expansion's poor performance can be explained by the fact that it added too many treatment synonyms, which diluted the original query's symptoms. On average, dictionary-based expansion nearly doubled each query in size.

We show an example of a mixed query expansion from BiLDA. In the patient query in Table 5.2, the five expansion terms included three symptoms and two herbs. "Fluttering

Table 5.2: Example of an expanded query created by the BiLDA method. Different query terms are separated by semicolons.

Disease Label	Original Query	Expansion Terms
chronic gastritis	yellow, greasy tongue coating; epigastric chills; heartburn; bloating; belching; stomachache; acid reflux; dry mouth	fluttering pulse; dark, red tongue; fullness; bitter orange; crow-dipper

pulse" was an expansion term for this patient, and is indeed an indicative symptom of the patient's disease, chronic gastritis [111]. Dark, red tongue is a functional synonym of yellow, greasy tongue coating, both of which commonly appear in patients with chronic gastritis [112]. Fullness is a semantic synonym of bloating. Bitter oranges are known to treat abdominal bloating [113], and are furthermore known to treat chronic gastritis [114]. Lastly, crow-dippers are also known to treat bloating in chronic gastritis patients [115]. Indeed, crow-dipper was actually prescribed to this particular patient.

5.5 RELATED WORK

Zeng *et al.* performed a study of synonym-, topic-model-, and predicate-based query expansions. They used monolingual LDA as their choice of topic model and determined it to be the best-performing method [103]. Our work built upon their study in the context of traditional Chinese medicine, while also comparing additional methods, showing BiLDA to be even more effective. A major difference between their work and ours is that while both systems aimed to return the most similar patients to a query, each of their experimental queries was a single primary disease (e.g., "PTSD"), while our queries consisted of the complete set of symptoms per patient. Furthermore, we refined the choice of evaluation from traditional measures of precision, recall, and F_1 to the more comparable metric of NDCG@k. Choi *et al.* developed a method of learning efficient representations of medical codes, Med2Vec, which we used as one of this study's expansion methods [12]. Jain *et al.* also performed medical record retrieval with query expansion on a patient's symptoms [116]. However, they used a knowledge base by integrating domain ontologies and automatic semantic relationship learning, similar to Zeng *et al.*'s predicate-based query expansion. Due to the lack of TCM ontologies, this method was unfeasible.

5.6 CONCLUSIONS AND FUTURE WORK

In this work, we studied how query expansion can improve medical record retrieval. Prior work showed topic-model-based query expansion to perform the best [103]. We presented an improved topic model that frames symptoms and treatments as distinct languages.

We performed query expansion on EMR retrieval experiments with a treatment-symptom dictionary, latent Dirichlet allocation (LDA), Med2Vec, diffusion component analysis (DCA), and a polylingual topic model. LDA and dictionary synonyms were studied in prior work and thus served as baselines in this work. Our experimental results showed that our method performed the best by normalized discounted cumulative gain, with significant p-values computed from paired t-tests.

Future work includes experimenting with other methods of query expansion. For instance, pointwise mutual information (PMI) has shown promising treatment-symptom pairings. Another potential method is the Weighted Exclusivity Test (WExT), which computes triples of medical concepts as an extension to PMI [117].

Lastly, a fundamental change to our problem would be to reframe the retrieval task as a treatment recommendation system. Like before, the system would take a test patient's set of symptoms as the input query. However, instead of retrieving patient records relevant to the query, the system would recommend a set of drugs to treat the query symptoms. We can evaluate the new system by counting the number of recommended treatments that match the actual prescribed treatments for the test patient. With this framework, we can skip the step in which the doctor analyzes the set of returned patients in the retrieval task and instead directly recommend treatment. In fact, the embedding and knowledge graph-based methods, in addition to PMI and WExT outputs, can already generate explicit treatment synonyms that would enable this new framework.

CHAPTER 6: COMBINING ALZHEIMER'S AND PARKINSON'S DISEASES DATA TO IDENTIFY PATIENTS AT RISK FOR DEMENTIA

In this chapter, I discuss using heterogeneous data sources to identify patients at risk for dementia. Dementia is an umbrella term that describes a wide range of cognition-related symptoms, ranging from declining memory to impaired language skills. Alzheimer's disease (AD) is a chronic, irreversible neurodegenerative disease [118]. It is a type of dementia, characterized by progressive impairment in memory, judgment, language, and orientation. The primary pathological features of AD are neuronal loss in addition to the accumulation of extracellular plaques containing amyloid β (Aβ) and neurofibrillary tangles (NFT) containing tau in the brain [119]. AD affects an estimated 5.7 million people in the United States [120].

Parkinson's disease (PD), which I discussed in Chapter 4, is also an irreversible, progressive neurological disease that is associated with dementia in most patients long-term [121]. PD initially manifests as a disorder of the motor system, and a typical pathological characteristic is the abnormal accumulation of α-synuclein in the brain, leading to Lewy body deposits [122]. PD patients can eventually develop a type of dementia known as Parkinson's disease dementia (PDD). PD affects one million people in the United States [123].

AD, PD, and other sources of dementia cost roughly $277 billion a year for treatment [124]. Predicting the onset of dementia in patients could potentially mitigate these treatment costs by allowing doctors to employ preventative measures at an early stage [125]. However, prediction models rely on detailed longitudinal patient data, which is expensive and time-consuming to obtain. These datasets can cost tens of millions of dollars over many years for just a few hundred patients. Furthermore, the data collection studies oftentimes lack standardization and statistical power.

In this work, we took the perspective that although AD and PD are clinically distinct entities, the two diseases are closely related in their pathologies (Figure 6.1). For example, Aβ accumulation, an important hallmark of AD pathology, has also been reported to be present in some PD patients [126]. Similarly, typical PD characteristics have also been reported in AD patients, such as Lewy body deposition [127].

AD and PD may also share genetic factors. The APOE and MAPT genes have been linked to the presence of Aβ aggregates and greater tau protein expression, which increase the risk of AD [128]. Variants of these two genes have also been verified to increase the risk of the development of dementia in PD [129].

Additionally, many imaging-based measures have been associated with cognitive performance in both diseases. For example, a baseline AD pattern of brain atrophy, quantified

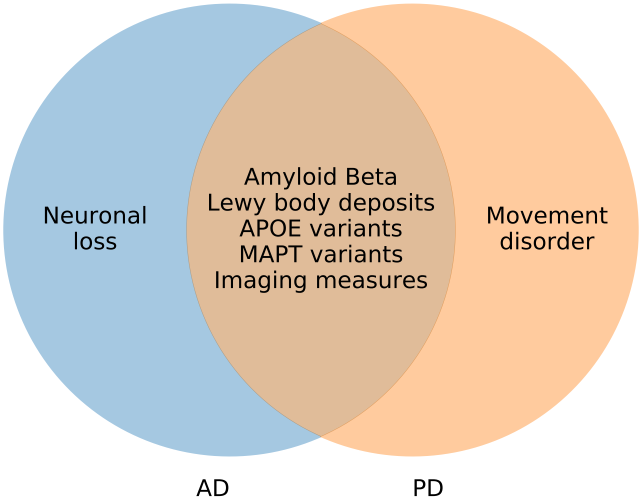


Figure 6.1: Evidence of pathological overlaps, including biospecimen, genetic, and imaging features, can help unify AD and PD patients into a single feature space.

using the Spatial Pattern of Atrophy for Recognition of AD (SPARE-AD) score, predicted long-term global cognitive decline in non-demented PD patients [130].

Despite the discovery of such correlations, research projects that study AD and PD remain distinct, resulting in separate sources of data. This potentially raises the cost of generating datasets that could otherwise be combined. This motivated an aggregated, data-driven study to explore the correlation between the two diseases, which could lead to a clearer understanding of their pathologies.

In this work, we explored biospecimen, genetic, and imaging features that are common to both AD and PD to place patients from both diseases into a common feature space. We showed that we can increase the size of our dataset by supplementing PD data with AD data. We tested this data aggregation in a classification task, in which we predicted the probability of patients developing dementia (either AD or PDD) after their baseline visits. By identifying patients at high risk, we can facilitate the more efficient targeting of available preventive measures than currently possible.

6.1 METHODS

The following sections describe our datasets, classification objective, feature selection process, and the classifier used to accomplish the dementia prediction task.

6.1.1 Dataset Description

Alzheimer's Disease

The AD data used in the preparation of this work was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database[1].

We selected ADNI-2 participants who did not have dementia at their baseline visit. Along with the availability of features, this reduced the dataset to 348 ADNI patients. The baseline and follow-up diagnoses of each patient were provided in the dataset.

Parkinson's Disease

For PD, we again used the PPMI dataset. We utilized Montreal Cognitive Assessment (MoCA) scores [131] to assign dementia labels. They range between 0 and 30, where lower scores indicate higher cognitive impairment. We used the MoCA score to label a patient with respect to his or her cognitive impairment level, as in prior work [132, 133]. A MoCA score > 26 meant the patient was normal with respect to dementia. MoCA scores in the range of $[21, 26]$ implied that the patient had mild cognitive impairment (MCI), and a MoCA score < 21 meant the patient suffered from dementia-level cognitive impairment. Due to the availability of MoCA scores and other features, we were left with 150 PPMI patients.

6.1.2 Objective Overview

In this study, we aimed to predict a given patient's cognitive status in a follow-up visit using only baseline features. The follow-up visit could occur anywhere between two to five years after a patient's baseline visit. We used the last follow-up visit available for any given patient.

For both the baseline and follow-up visits, we labeled patients with three classes: normal cognition (normal CI), mild cognitive impairment (MCI), and dementia. Our methodology for labeling patients in each dataset is outlined in Table 6.1. We further summarize the patient statistics according to disease progression in Table 6.2. Note that since we were only concerned with progression, we did not consider cases in which patients improve (e.g., from MCI to normal cognition). Furthermore, we excluded patients who had dementia at their baseline visits. Our goal was to accurately identify patients who had a follow-up label of dementia.

[1] adni.loni.usc.edu

Table 6.1: The assignment of classification labels to patients in each dataset. ADNI patients were labeled based on their cognitive status diagnoses, while PPMI patients were labeled based on their MoCA scores.

Classification Label	ADNI Criterion	PPMI Criterion
Normal CI	Control patients	MoCA > 26
MCI	MCI patients	MoCA $\in [21, 26]$
Dementia	AD patients	MoCA < 21

Table 6.2: Label statistics for the disease progression of patients in the ADNI and PPMI datasets, split into the possible classes of normal CI, MCI, and dementia.

Dataset	Baseline Label	Follow-Up Label	Number of Patients
ADNI	Normal CI	Normal CI	140
	Normal CI	MCI	16
	Normal CI	Dementia	3
	MCI	MCI	128
	MCI	Dementia	61
PPMI	Normal CI	Normal CI	85
	Normal CI	MCI	30
	Normal CI	Dementia	4
	MCI	MCI	26
	MCI	Dementia	5

We acknowledged that there were only nine patients in PPMI that developed dementia by the time of a follow-up visit. This resulted in test sets having very few PDD patients, depending on the number of folds. However, we attempted to address this issue by computing significance tests when comparing different classifiers. Additionally, due to the small number of PDD patients, we hypothesized that augmenting the smaller PPMI dataset with the ADNI dataset could vastly improve PDD prediction. On the other hand, we hypothesized that ADNI dataset was unlikely to greatly benefit from the smaller PPMI dataset for AD prediction. Thus, we were most interested in seeing whether a combined dataset built from ADNI and PPMI could improve the prediction of PDD patients.

6.1.3 Feature Selection

To jointly study AD and PD, we considered biospecimen, genetic, and imaging features known to influence dementia prediction in both diseases [134]. Additionally, we used the sex of each patient, which has been shown to be related to dementia development in both AD

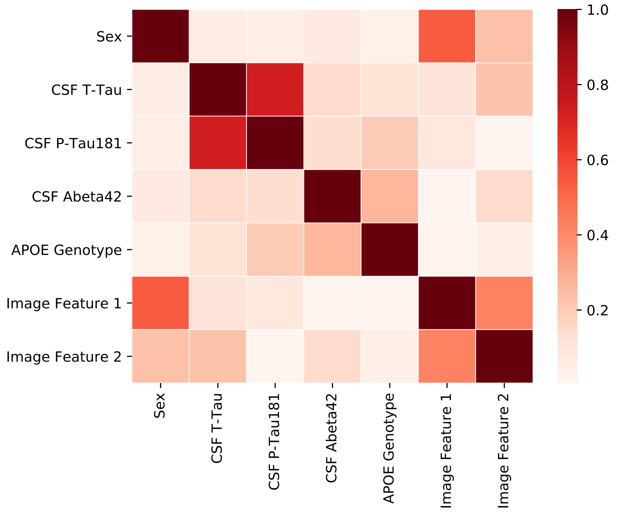


Figure 6.2: Heatmap of the absolute values of correlation coefficients between each pair of features. We found that t-tau and p-tau$_{181}$ were highly correlated with a Pearson correlation coefficient of 0.733. Therefore, we excluded p-tau$_{181}$ based on its lower performance in dementia prediction.

[135] and PD [136]. Our other selected features were also backed by significant literature evidence, and were available in both the ADNI and PPMI datasets. We describe these features in the following sections.

Biospecimen Features

For biospecimen features, we selected cerebrospinal fluid (CSF) concentrations of Aβ42, total tau protein (t-tau), and tau protein phosphorylated at threonine 181 (p-tau$_{181}$). Aβ42 levels have been shown to play an essential role in all forms of AD pathogenesis [137]. Elevated t-tau and p-tau$_{181}$ concentrations have been shown to be associated with neurodegenerative changes in early AD [138, 139]. Each of these biomarkers also has strong prognostic and diagnostic potential in early-stage PD [140].

Not surprisingly, we found that p-tau$_{181}$ was highly correlated with t-tau in our dataset, with a Pearson correlation coefficient of 0.733 and a p-value of 5.59×10^{-85} (Figure 6.2). Thus, as a standard machine learning step, we excluded one of these features. We found that excluding p-tau$_{181}$ yielded better performance than when excluding t-tau (see Section 6.2.4).

We considered using CSF α-synuclein levels as a biospecimen feature [141], but this data was unavailable for ADNI-2/GO participants. In total, we used two numerical biospecimen features.

Table 6.3: Mapping the APOE genotype to a numerical variable. The number of copies of the $\epsilon4$ allele dictated the feature value.

APOE Genotype	**Feature Value**
$\epsilon2/\epsilon2$, $\epsilon2/\epsilon3$, and $\epsilon3/\epsilon3$	0
$\epsilon2/\epsilon4$ and $\epsilon3/\epsilon4$	1
$\epsilon4/\epsilon4$	2

Genetic Feature

We selected a single genetic feature: each patient's APOE genotype. We paid special attention to the $\epsilon4$ allele (APOE4), which is the largest known genetic risk factor for AD in a variety of ethnic groups [142, 143]. Specifically, risk for AD was shown to increase with the number of APOE $\epsilon4$ alleles. Additionally, APOE4 has been associated with cognitive decline in Parkinson's disease [144]. We incorporated the APOE genotype by using the number of copies of the $\epsilon4$ allele as a numerical variable. The full mapping from APOE genotype to numerical variable is shown in Table 6.3.

Imaging Features

To obtain imaging features, we used the FreeSurferV5.1 software, which analyzes structural magnetic resonance imaging (MRI) scans [145]. We processed PPMI MRIs at baseline and screening visits to generate the imaging features. FreeSurfer-generated imaging features were already included in the ADNI dataset.

After obtaining the processed image features, we used principal component analysis (PCA) [146] to reduce the image feature space to two dimensions. We empirically chose to use the two most principal components as numerical features in our classification model because adding a third image feature decreased classification performance.

Alternate dimensionality reduction methods, such as independent component analysis (ICA), did not perform as well as PCA on our dataset.

6.1.4 Experimental Design

Input and Output

To summarize the previous sections, we selected sex, CSF $A\beta42$ levels, CSF t-tau levels, the APOE genotype, and the two imaging features as input to train a classifier. To prevent data leakage, we excluded the baseline cognitive label from the classification input. We performed min-max normalization within each dataset prior to training.

The output was a patient's cognitive disease label (normal CI, MCI, or dementia) in a follow-up visit after the baseline. Again, we were most interested in predicting whether a patient would develop dementia.

We designed experiments by training and testing separately on each dataset. Additionally, we combined the datasets into a single training set by adding the dataset not being tested. For example, if testing on held-out PPMI patients, we trained on all other PPMI patients as well as all ADNI patients.

Cross-Validation

We performed repeated, stratified k-fold cross-validation, splitting the folds on AD and PD patients independently, for $k \in \{3, 5\}$. Results were insensitive to the choice of k, with no significant difference between $k = 3$ and $k = 5$ (p-value > 0.05). We recombined the folds for training to avoid imbalances between the two diseases. Thus, each fold contained equal proportions of AD and PD patients as well as equal proportions of normal cognition, MCI, and dementia patients. We report the AUC values along with the 95% confidence intervals for dementia prediction from the 3-fold cross-validation to maximize the number of PD patients with dementia in the test set.

Classifier

We used logistic regression with the L-BFGS solver [147], which outperformed random forest classifier, support vector machines, and multi-layer perceptrons at both 3- and 5-fold cross-validation.

6.2 RESULTS AND DISCUSSION

6.2.1 ADNI and PPMI patients had overlapping feature values

From strong evidence in prior work, we saw the role that each of these features play in the onset of dementia in both AD and PD patients. To further support this, we observed overlapping features values between ADNI and PPMI patients (Figure 6.3).

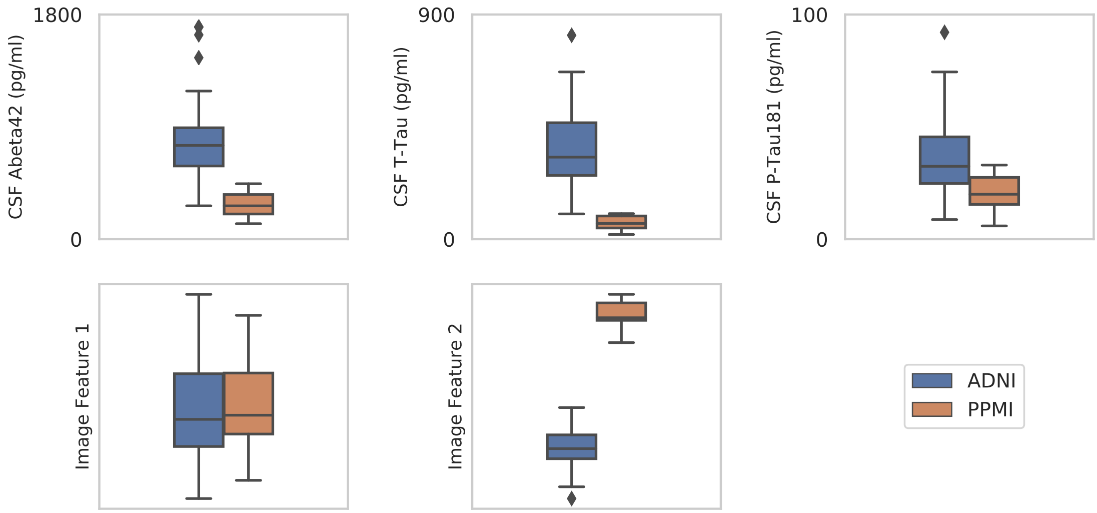


Figure 6.3: Feature value comparisons between ADNI and PPMI patients with dementia, prior to min-max normalization. Image Features 1 and 2 refer to the two most principal components of the FreeSurfer-generated image features.

6.2.2 Visualization yielded clusters of dementia patients

We plotted the two datasets in the same visualization using our selected features. As in Chapter 4, we used t-SNE to reduce our six features to two dimensions. We saw that patients with dementia, regardless of dataset, clustered together in similar areas (Figure 6.4). Furthermore, the resulting embedding vectors were correlated with cognitive status with Pearson correlation coefficients of 0.217 and 0.257 and p-values of 1.05×10^{-6} and 5.98×10^{-9}, respectively. Thus, our features can be useful in identifying patients with dementia.

6.2.3 The combined dataset improved classification performance

Table 6.4 summarizes the performance of our classifier. To determine whether changes in AUC were statistically significant, we computed paired t-tests.

When training and testing on ADNI, we achieved an AUC value for dementia prediction of 0.855 (95% CI $[0.848, 0.861]$). When training on the combined dataset, we observed a statistically significant improvement with an AUC value of 0.862 (95% CI $[0.856, 0.868]$, p-value $= 7.21 \times 10^{-4}$). As expected, the absolute increase in AUC was low, likely due to the small number of PDD patients.

However, we saw a major improvement when using the combined dataset for predicting dementia in PD patients. When training and testing only on PPMI, we achieved an AUC

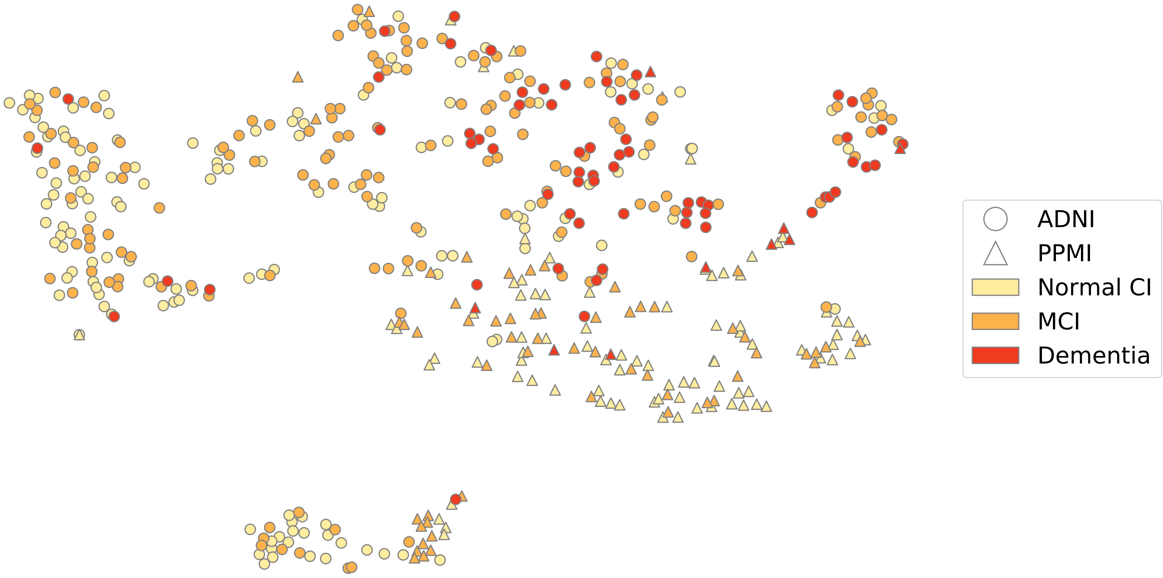


Figure 6.4: Two-dimensional visualization of the ADNI and PPMI patients using our six selected features. ADNI patients are indicated by circles, and PPMI patients are indicated by triangles. The color of each point denotes the cognitive level of the corresponding patient, where red means that patients developed dementia by the time of their follow-up visits. We saw that patients with dementia (points in red) tended to cluster together, regardless of dataset.

value for dementia prediction of 0.797 (95% CI $[0.783, 0.811]$). When training on the combined dataset, we achieved an AUC value of 0.883 (95% CI $[0.873, 0.892]$), a statistically significant improvement with a p-value of 5.05×10^{-21}.

Furthermore, we compared with the setting of training on the ADNI dataset and testing on PPMI, a baseline replicated from a previous study [134]. This setting achieved an AUC of 0.840. Training on the combined dataset performed better with a p-value of 4.24×10^{-15}.

Overall, using the combined dataset to train the classifier improved the prediction of dementia. The main limitation of our model was the small amount of PDD patients in our dataset. We attempted to address this by utilizing 3-fold cross-validation to maximize the number of PDD patients in each test set. Additionally, we showed that AD and PD data can be combined to alleviate this issue. Note that the augmentation of the PPMI dataset with the ADNI dataset was not intended to be a long-term solution to the lack of data. Rather, we hope this study reveals that AD and PD patients share similarities that can facilitate the prediction of the onset of dementia.

Table 6.4: Classification performance for each of our experimental settings for 3-fold cross-validation. When testing on either ADNI or PPMI, we saw improvements when training on the combined dataset versus when training on the original dataset.

Training Set	**Test Set**	**Dementia AUC**
ADNI	ADNI	0.855
Combined	ADNI	**0.862**
PPMI	PPMI	0.797
ADNI	PPMI	0.840
Combined	PPMI	**0.883**

6.2.4 Choosing between CSF total tau and p-tau$_{181}$

Recall that CSF t-tau and p-tau$_{181}$ concentrations were highly correlated in our dataset. We performed leave-one-out experiments to select one feature between the two, holding the rest of the features constant.

We found that including t-tau versus including p-tau$_{181}$ did not yield a statistically significant difference in AD prediction AUC. However, when using both features, the AUC dropped from 0.862 to 0.861 (95% CI $[0.855, 0.868]$), a statistically significant decrease with a p-value of 4.77×10^{-3}.

On the other hand, we found that including p-tau$_{181}$ yielded a PDD prediction AUC of 0.853 (95% CI $[0.841, 0.864]$), a decrease from 0.883 when using t-tau. This was statistically significant with a p-value of 1.89×10^{-13}. The AUC dropped to 0.877 (95% CI $[0.867, 0.888]$) when using both features, a decrease with a p-value of 9.84×10^{-3}. Due to the improved performance of dementia prediction when using t-tau over p-tau$_{181}$, we chose to keep only t-tau while excluding p-tau$_{181}$ from the classifier.

6.2.5 Each of our selected features contributed to the classifier

We illustrated the importance of each feature used in our classifier trained on the combined dataset by computing their standardized model weights. We found that most features had high importance weights (Figure 6.5). However, the APOE genotype had a surprisingly low relative feature importance (mean of 22.1), requiring further analysis.

When leaving the APOE genotype out of the classifier, the prediction AUC for AD dropped from 0.862 to 0.861 (95% CI $[0.854, 0.867]$) with a p-value of 0.0232. The prediction AUC for PDD dropped from 0.883 to 0.877 (95% CI $[0.867, 0.887]$) with a p-value of 3.23×10^{-9}. As a result, we concluded that the APOE genotype can play an important role in dementia prediction.

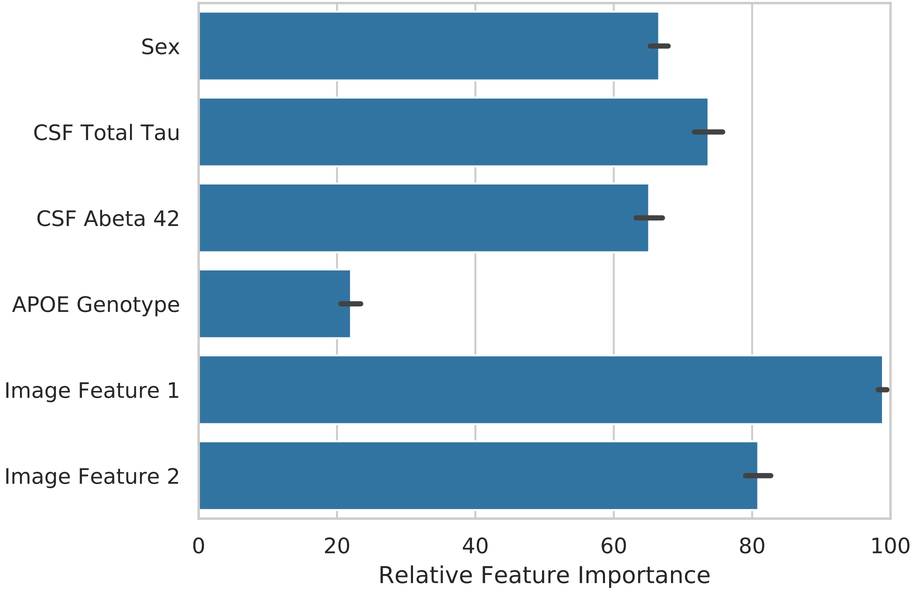


Figure 6.5: The relative importance for each feature when training on the combined dataset. Most features had a high importance in the logistic regression classifier. However, the APOE genotype required further analysis (mean relative feature importance of 22.1).

Although including the APOE genotype increased prediction performance, this did not necessarily mean that there was a strong association between the $\epsilon 4$ allele and the onset of dementia in either disease. Indeed, there are still ongoing debates on this topic, particularly in Parkinson's disease [148, 149, 150, 151].

To firmly establish a causal relationship, we would require further experimental analysis to elucidate how apolipoprotein E directly affects brain function over time. In the meantime, we have shown that identifying a patient's APOE genotype can help doctors predict the development of dementia in both AD and PD.

6.3 RELATED WORK

Our study was motivated by many other works that attempted to study AD and PD from various points of view. Calderone *et al.* analyzed AD and PD from a network perspective to quantify functional and topological similarities between the two pathologies [152]. They used a network community discovery algorithm, InfoMap, to perform this task. However, they did not predict disease progression. Berlyand *et al.* performed a study deriving biomarker signatures from AD to identify PD patients with dementia [134]. However, they did not combine the AD and PD datasets. Several other studies used similar features as ours to predict progression of AD and PD patients individually [141, 153, 144], but never together in the same dataset.

6.4 CONCLUSIONS AND FUTURE WORK

In this work, we attempted to identify AD and PD patients who had high risk of developing dementia after their baseline visits. We framed this as a classification task in which we predicted the patient's cognitive status in a follow-up visit after the baseline. This task is of high clinical importance, potentially allowing doctors to employ preventative measures at an early stage in treatment.

Our method involved identifying features that have been shown in literature evidence to drive both AD and PD pathologies. With these features, we built a combined dataset from the ADNI and PPMI datasets to use as training input of a classifier. We showed that our method yielded statistically significant improvements when predicting the dementia status of patients from both datasets.

The main limitations of our work were associated with the PD dataset. Specifically, our data lacked a high number of PDD patients. However, our solution attempted to address this by supplementing the PD data with AD data. Additionally, when testing on the PD data, we utilized paired t-tests to show that our improvements were statistically significant.

We hope that this work can open future opportunities for collaboration between the AD and PD research communities. For future work, we can replicate these results on more datasets to corroborate our claims, especially since one of our major limitations is the lack of PD patients with dementia. Furthermore, we can include more features that are common to AD and PD patients, such as SNPs and other genetic data.

CHAPTER 7: CONCLUSIONS AND FUTURE DIRECTIONS

In this thesis, I studied the problem of data source heterogeneity in the analysis of EMRs. Using network embedding methods to integrate external data sources and enhance EMRs, I improved a variety of downstream applications. PaReCat uses an external herb-symptom dictionary to mine higher quality patient subcategorizations. This can lead to a better understanding of diseases, and can also facilitate precision medicine. HEMnet includes molecular interaction networks and domain knowledge to better perform survival analysis. Improved survival analysis can help doctors understand the features that may result in higher survival rates. VisAGE adds genomic data to provide more interpretable visualizations. These high-quality visualizations can help doctors identify interpretable patient clusters.

Additionally, I performed query expansion by framing symptoms and treatments as separate languages. This resulted in retrieving EMRs more relevant to an input query, which can help build EMR search engines. Lastly, I utilized heterogeneous data sources to combine Alzheimer's and Parkinson's diseases datasets. This allowed me to accurately predict the risk of a given patient developing dementia. This project also affirmed that these two diseases can be studied jointly, which can lead to cost-saving measures during data collection.

The results of these works highlight the potential of incorporating different data sources in the computational healthcare domain. However, there is still much to be done before clinical translation can be fully achieved.

For example, as EMR databases become more standardized, the problems associated with semantic synonyms will diminish. However, treatment and functional synonyms will still result in semantic mismatching because they are medical relationships that do not disappear with more consistent recordkeeping. This highlights the need for creating higher quality knowledge graphs.

Although an increase in data availability as well as computational power could solve a handful of issues, great strides can still be taken with the resources currently available.

7.1 TEXT MINING IN CLINICAL NOTES

The research in this thesis dealt primarily with the different types of data that are typically available in EMRs. However, one ubiquitous EMR data type that I omitted was clinical notes. These pieces of unstructured text data are crucial to differentiating patients, and offer additional insight into a patient's condition. Due to their unstructured nature, I excluded their direct usage in the analyses performed in this thesis.

There have been many attempts at text mining directly from medical records [154, 155, 156, 157], but many public datasets lack the capacity for a large-scale undertaking. Further work needs to be done such that clinical notes become more readily available in conjunction with the plethora of heterogeneous data types already online. This would enable us to enlarge our heterogeneous network and find more paths between related entities. These relationships would initially have to be mined from the internal EMR data, but could eventually generalize to online databases.

7.2 EXPANDING KNOWLEDGE NETWORKS WITH INFORMATION FROM MULTIPLE DISEASES

As discussed in Chapter 6, many diseases are actually related, and the developments and pathologies of ailments from even different body parts could be interconnected. In my studies that utilized knowledge networks, I was able to link similar medical entities using network embeddings. However, I only dealt with a single disease at a time due to external database constraints.

On the other hand, in order to study Alzheimer's and Parkinson's diseases jointly, we hand-selected common features using literature evidence. In the future, it is possible that a knowledge network can enable the automatic mining of this information, and many types of diseases and relationships can be studied from a single network. Heavy collaboration between doctors and computer scientists would most certainly be required. If successful, this type of knowledge network could pave the way for further joint analyses of complex diseases, allowing for a deeper understanding of how genetic and environmental factors combine to affect the human body.

However, it is not necessarily advantageous to always use as large a knowledge graph as possible (global knowledge graph). A network tailored for a disease might be better suited in some cases. One possible method of building a tailored knowledge graph is to select interactions obtained from experiments on patients suffering from the disease of interest. Another possibility is to perform text mining on publications that concern the disease and adding the mined relationships to the global network. In the future, it would be important to use both a global knowledge graph and a tailored one. In such a case, an important question to address is how to regulate the relative weights. One solution is to empirically set the weights using a training data set by optimizing a meaningful metric.

7.3 UTILIZING GENE EXPRESSION DATA

Finally, we can expand on our usage of genomic data with gene expression profiles. In our work, the genomic data was limited to SNPs, which are mid-level descriptions of mutations. Gene expression profiles have the potential to be the next crucial step in the role of genomics in computational healthcare. We have seen the promising results of using gene expression profiling in drug response prediction [158, 159, 160, 161], mining association rules [162, 163], and more. I hypothesize that by using gene expression data, we can access the low-level details, which can increase the quality of the connections between patients and medical entities.

REFERENCES

[1] R. Hillestad, J. Bigelow, A. Bower, F. Girosi, R. Meili, R. Scoville, and R. Taylor, "Can electronic medical record systems transform health care? Potential health benefits, savings, and costs," *Health affairs*, vol. 24, no. 5, pp. 1103–1117, 2005.

[2] D. W. Bates, L. L. Leape, D. J. Cullen, N. Laird, L. A. Petersen, J. M. Teich, E. Burdick, M. Hickey, S. Kleefield, B. Shea et al., "Effect of computerized physician order entry and a team intervention on prevention of serious medication errors," *Jama*, vol. 280, no. 15, pp. 1311–1316, 1998.

[3] E. W. Huang, S. Wang, R. Zhang, B. Liu, X. Zhou, and C. Zhai, "PaReCat: Patient record subcategorization for precision traditional Chinese medicine," in *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 2016, pp. 443–452.

[4] E. W. Huang, S. Wang, B. Li, R. Zhang, B. Liu, R. Zhang, J. Liu, X. Zhou, H. Lin, and C. Zhai, "HEMnet: Integration of electronic medical records with molecular interaction networks and domain knowledge for survival analysis," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 2017, pp. 378–387.

[5] E. W. Huang, S. Wang, and C. Zhai, "VisAGE: Integrating external knowledge into electronic medical record visualization," in *Pacific Symposium on Biocomputing 2018*, vol. 23. World Scientific, 2018, pp. 578–89.

[6] M. Wang, Z. Geng, M. Wang, F. Chen, W. Ding, and M. Liu, "Combination of network construction and cluster analysis and its application to traditional Chinese medicine," in *International Symposium on Neural Networks*. Springer, 2006, pp. 777–785.

[7] J. Tang, J. Liu, M. Zhang, and Q. Mei, "Visualizing large-scale and high-dimensional data," in *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2016, pp. 287–297.

[8] X. Yu, Q. Gu, M. Zhou, and J. Han, "Citation prediction in heterogeneous bibliographic networks," in *Proceedings of the 2012 SIAM International Conference on Data Mining*. SIAM, 2012, pp. 1119–1130.

[9] D. Zhou, S. A. Orshanskiy, H. Zha, and C. L. Giles, "Co-ranking authors and documents in a heterogeneous network," in *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*. IEEE, 2007, pp. 739–744.

[10] Y. Li and J. C. Patra, "Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network," *Bioinformatics*, vol. 26, no. 9, pp. 1219–1224, 2010.

[11] L. Liu, J. Tang, J. Han, M. Jiang, and S. Yang, "Mining topic-level influence in heterogeneous networks," in *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010, pp. 199–208.

[12] E. Choi, M. T. Bahadori, E. Searles, C. Coffey, M. Thompson, J. Bost, J. Tejedor-Sojo, and J. Sun, “Multi-layer representation learning for medical concepts,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 2016, pp. 1495–1504.

[13] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, “Deep patient: An unsupervised representation to predict the future of patients from the electronic health records,” *Scientific reports*, vol. 6, p. 26094, 2016.

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.

[15] E. W. Huang, S. Wang, D. J.-L. Lee, R. Zhang, B. Liu, X. Zhou, and C. Zhai, “Framing electronic medical records as polylingual documents in query expansion,” in *AMIA Annual Symposium Proceedings*, vol. 2017. American Medical Informatics Association, 2017, p. 940.

[16] P. N. Robinson, “Deep phenotyping for precision medicine,” *Human mutation*, vol. 33, no. 5, pp. 777–780, 2012.

[17] S. Bent, “Herbal medicine in the United States: Review of efficacy, safety, and regulation,” *Journal of general internal medicine*, vol. 23, no. 6, pp. 854–859, 2008.

[18] E. Chan, M. Tan, J. Xin, S. Sudarsanam, and D. E. Johnson, “Interactions between traditional Chinese medicines and western,” *Current Opinion in Drug Discovery and Development*, vol. 13, no. 1, pp. 50–65, 2010.

[19] National Research Council, *Toward precision medicine: Building a knowledge network for biomedical research and a new taxonomy of disease.* National Academies Press, 2011.

[20] J. M. McGinnis, L. Stuckhardt, R. Saunders, M. Smith et al., *Best care at lower cost: The path to continuously learning health care in America.* National Academies Press, 2013.

[21] Chinese Pharmacopoeia Commission, “Pharmacopoeia of the People’s Republic of China,” 2015.

[22] H. Cho, B. Berger, and J. Peng, “Diffusion component analysis: Unraveling functional topology in biological networks,” in *International Conference on Research in Computational Molecular Biology.* Springer, 2015, pp. 62–64.

[23] S. Wang, H. Cho, C. Zhai, B. Berger, and J. Peng, “Exploiting ontology graph for predicting sparsely annotated gene function,” *Bioinformatics*, vol. 31, no. 12, pp. i357–i364, 2015.

[24] “Preparation process for yinianjin capsule,” June 25 2014, cN Patent App. CN 201,210,561,504. [Online]. Available: http://www.google.com/patents/CN103877311A

[25] F. S. Roque, P. B. Jensen, H. Schmock, M. Dalgaard, M. Andreatta, T. Hansen, K. Søeby, S. Bredkjær, A. Juul, T. Werge et al., "Using electronic patient records to discover disease correlations and stratify patient cohorts," *PLoS computational biology*, vol. 7, no. 8, 2011.

[26] Q. He, X. Zhou, Z. Zhou, M. Cui, and Z. Wu, "Efficacy-based clustering analysis of traditional Chinese medicinal herbs," *Chin J Inf TCM*, vol. 11, no. 7, pp. 561–562, 2004.

[27] Q. Shi, H. Zhao, J. Chen, X. Ma, Y. Yang, C. Zheng, and W. Wang, "Study on TCM syndrome identification modes of coronary heart disease based on data mining," *Evidence-Based Complementary and Alternative Medicine*, 2012.

[28] A. Ellis and N. Wiseman, *Fundamentals of Chinese medicine.* Paradigm publications, 1995.

[29] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.

[30] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *Journal of Machine Learning Research*, vol. 11, no. Oct, pp. 2837–2854, 2010.

[31] D. Bertsimas, M. V. Bjarnadóttir, M. A. Kane, J. C. Kryder, R. Pandey, S. Vempala, and G. Wang, "Algorithmic prediction of health-care costs," *Operations Research*, vol. 56, no. 6, pp. 1382–1392, 2008.

[32] Y. Zhao, Q. Cao, H. Liu, K. Wang, A. Yan, and Z. Hu, "Determination of baicalin, chlorogenic acid and caffeic acid in traditional Chinese medicinal preparations by capillary zone electrophoresis," *Chromatographia*, vol. 51, no. 7-8, pp. 483–486, 2000.

[33] Z. Gao, J. Shao, H. Sun, W. Zhong, W. Zhuang, and Z. Zhang, "Evaluation of different kinds of organic acids and their antibacterial activity in Japanese Apricot fruits," *African Journal of Agricultural Research*, vol. 7, no. 35, pp. 4911–4918, 2012.

[34] D. L. McKay and J. B. Blumberg, "A review of the bioactivity and potential health benefits of peppermint tea (Mentha piperita L.)," *Phytotherapy Research: An International Journal Devoted to Pharmacological and Toxicological Evaluation of Natural Product Derivatives*, vol. 20, no. 8, pp. 619–633, 2006.

[35] J. S. Nimitz, "Antiviral supplement compositions and methods of use," Apr. 11 2017, uS Patent 9,616,124.

[36] T. Lahans, "Integrating Chinese and conventional medicine in colorectal cancer treatment," *Integrative cancer therapies*, vol. 6, no. 1, pp. 89–94, 2007.

[37] S. Dharmananda, *Treatment of Gallstones with Chinese Herbs and Acupuncture.* ITM, 2001.

[38] F. Xie, M. Zhang, C.-F. Zhang, Z.-T. Wang, B.-Y. Yu, and J.-P. Kou, "Anti-inflammatory and analgesic activities of ethanolic extract and two limonoids from Melia toosendan fruit," *Journal of ethnopharmacology*, vol. 117, no. 3, pp. 463–466, 2008.

[39] R. Harpaz, W. DuMouchel, N. H. Shah, D. Madigan, P. Ryan, and C. Friedman, "Novel data-mining methodologies for adverse drug event discovery and analysis," *Clinical Pharmacology & Therapeutics*, vol. 91, no. 6, pp. 1010–1021, 2012.

[40] S. Gripp, S. Moeller, E. Bolke, G. Schmitt, C. Matuschek, S. Asgari, F. Asgharzadeh, S. Roth, W. Budach, M. Franz et al., "Survival prediction in terminally ill cancer patients by clinical estimates, laboratory tests, and self-rated anxiety and depression," *Journal of Clinical Oncology*, vol. 25, no. 22, pp. 3313–3320, 2007.

[41] P. Liu, L. Lei, J. Yin, W. Zhang, W. Naijun, and E. El-Darzi, "Healthcare data mining: Prediction inpatient length of stay," in *Intelligent Systems, 2006 3rd International IEEE Conference on.* IEEE, 2006, pp. 832–837.

[42] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou et al., "STRING v10: Protein–protein interaction networks, integrated over the tree of life," *Nucleic acids research*, vol. 43, no. D1, pp. D447–D452, 2014.

[43] L. Hirschman, G. A. Burns, M. Krallinger, C. Arighi, K. B. Cohen, A. Valencia, C. H. Wu, A. Chatr-Aryamontri, K. G. Dowell, E. Huala et al., "Text mining for the biocuration workflow," *Database*, vol. 2012, 2012.

[44] S. Oliver, "Proteomics: Guilt-by-association goes global," *Nature*, vol. 403, no. 6770, p. 601, 2000.

[45] I. Lee, U. M. Blom, P. I. Wang, J. E. Shim, and E. M. Marcotte, "Prioritizing candidate disease genes by network-based boosting of genome-wide association data," *Genome research*, pp. gr–118 992, 2011.

[46] S. Wang, M. Qu, and J. Peng, "ProSNet: Integrating homology with molecular networks for protein function prediction," in *Pacific Symposium on Biocomputing 2017.* World Scientific, 2017, pp. 27–38.

[47] C. A. McCarty, R. L. Chisholm, C. G. Chute, I. J. Kullo, G. P. Jarvik, E. B. Larson, R. Li, D. R. Masys, M. D. Ritchie, D. M. Roden et al., "The eMERGE network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies," *BMC medical genomics*, vol. 4, no. 1, p. 13, 2011.

[48] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, "Pathsim: Meta path-based top-k similarity search in heterogeneous information networks," *Proceedings of the VLDB Endowment*, vol. 4, no. 11, pp. 992–1003, 2011.

[49] K. Caballero and R. Akella, "Dynamic estimation of the probability of patient readmission to the ICU using electronic medical records," in *AMIA Annual Symposium Proceedings*, vol. 2015. American Medical Informatics Association, 2015, p. 1831.

[50] S. Clavey, *Fluid physiology and pathology in traditional Chinese medicine.* Churchill Livingstone, 1995.

[51] D. G. Beer, S. L. Kardia, C.-C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek, L. Lin, G. Chen, T. G. Gharib, D. G. Thomas et al., "Gene-expression profiles predict survival of patients with lung adenocarcinoma," *Nature medicine*, vol. 8, no. 8, p. 816, 2002.

[52] Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer-van Gelder, J. Yu et al., "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *The Lancet*, vol. 365, no. 9460, pp. 671–679, 2005.

[53] P. C. Burger and S. B. Green, "Patient age, histologic features, and length of survival in patients with glioblastoma multiforme," *Cancer*, vol. 59, no. 9, pp. 1617–1625, 1987.

[54] A. J. Nixon, D. Neuberg, D. F. Hayes, R. Gelman, J. L. Connolly, S. Schnitt, A. Abner, A. Recht, F. Vicini, and J. R. Harris, "Relationship of patient age to pathologic features of the tumor and prognosis for patients with stage I or II breast cancer." *Journal of Clinical Oncology*, vol. 12, no. 5, pp. 888–894, 1994.

[55] A. Singhal et al., "Modern information retrieval: A brief overview," *IEEE Data Eng. Bull.*, vol. 24, no. 4, pp. 35–43, 2001.

[56] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American statistical association*, vol. 53, no. 282, pp. 457–481, 1958.

[57] N. Mantel, "Evaluation of survival data and two new rank order statistics arising in its consideration," *Cancer Chemother Rep*, vol. 50, pp. 163–170, 1966.

[58] R. G. Miller Jr, *Survival Analysis.* John Wiley & Sons, 2011, vol. 66.

[59] J. R. Egner, "AJCC cancer staging manual," *JAMA*, vol. 304, no. 15, pp. 1726–1727, 2010.

[60] D. A. Karnofsky, W. H. Abelmann, L. F. Craver, and J. H. Burchenal, "The use of the nitrogen mustards in the palliative treatment of carcinoma. With particular reference to bronchogenic carcinoma," *Cancer*, vol. 1, no. 4, pp. 634–656, 1948.

[61] T. H. Kim, J. S. Kim, Z. H. Kim, R. B. Huang, and R. S. Wang, "Khz (fusion of ganoderma lucidum and polyporus umbellatus mycelia) induces apoptosis in A549 human lung cancer cells by generating reactive oxygen species and decreasing the mitochondrial membrane potential," *Food Science and Biotechnology*, vol. 23, no. 3, pp. 859–864, 2014.

[62] Y. Zhang, Q. Wang, T. Wang, H. Zhang, Y. Tian, H. Luo, S. Yang, Y. Wang, and X. Huang, "Inhibition of human gastric carcinoma cell growth in vitro by a polysaccharide from Aster tataricus," *International journal of biological macromolecules*, vol. 51, no. 4, pp. 509–513, 2012.

[63] D. Glover, A. Lipton, A. Keller, A. A. Miller, S. Browning, R. J. Fram, S. George, K. Zelenakas, R. S. Macerata, and J. J. Seaman, "Intravenous pamidronate disodium treatment of bone metastases in patients with breast cancer. A dose-seeking study," *Cancer: Interdisciplinary International Journal of the American Cancer Society*, vol. 74, no. 11, pp. 2949–2955, 1994.

[64] J.-Y. Hung, C.-J. Yang, Y.-M. Tsai, H.-W. Huang, and M.-S. Huang, "Antiproliferative activity of paeoniflorin is through cell cycle arrest and the Fas/Fas ligand-mediated apoptotic pathway in human non-small cell lung cancer A549 cells," *Clinical and Experimental Pharmacology and Physiology*, vol. 35, no. 2, pp. 141–147, 2008.

[65] W. Bao, H. Pan, M. Lu, Y. Ni, R. Zhang, and X. Gong, "The apoptotic effect of sarsasapogenin from Anemarrhena asphodeloides on HepG2 human hepatoma cells," *Cell biology international*, vol. 31, no. 9, pp. 887–892, 2007.

[66] A. Rind, P. Federico, T. Gschwandtner, W. Aigner, J. Doppler, and M. Wagner, "Visual analytics of electronic health records with a focus on time," in *New Perspectives in Medical Records*. Springer, 2017, pp. 65–77.

[67] A. Rind, T. D. Wang, W. Aigner, S. Miksch, K. Wongsuphasawat, C. Plaisant, B. Shneiderman et al., "Interactive information visualization to explore and query electronic health records," *Foundations and Trends® in Human–Computer Interaction*, vol. 5, no. 3, pp. 207–298, 2013.

[68] M. Ozkaynak, B. Reeder, L. Hoffecker, M. B. Makic, and K. Sousa, "Use of electronic health records by nurses for symptom management in inpatient settings: A systematic review," *CIN: Computers, Informatics, Nursing*, vol. 35, no. 9, pp. 465–472, 2017.

[69] S. Ebadollahi, J. Sun, D. Gotz, J. Hu, D. Sow, and C. Neti, "Predicting patient's trajectory of physiological data using temporal trends in similar patients: A system for near-term prognostics," in *AMIA annual symposium proceedings*, vol. 2010. American Medical Informatics Association, 2010, p. 192.

[70] M. S. Hansen, G. J. Nogareda, and S. J. Hutchison, "Frequency of and inappropriate treatment of misdiagnosis of acute aortic dissection," *The American journal of cardiology*, vol. 99, no. 6, pp. 852–856, 2007.

[71] P. G. Hagan, C. A. Nienaber, E. M. Isselbacher, D. Bruckman, D. J. Karavite, P. L. Russman, A. Evangelista, R. Fattori, T. Suzuki, J. K. Oh et al., "The international registry of acute aortic dissection (IRAD): New insights into an old disease," *Jama*, vol. 283, no. 7, pp. 897–903, 2000.

[72] B. K. Nallamothu, R. H. Mehta, S. Saint, A. Llovet, E. Bossone, J. V. Cooper, U. Sechtem, E. M. Isselbacher, C. A. Nienaber, K. A. Eagle et al., “Syncope in acute aortic dissection: Diagnostic, prognostic, and clinical implications,” *The American journal of medicine*, vol. 113, no. 6, pp. 468–471, 2002.

[73] K. Marek, D. Jennings, S. Lasch, A. Siderowf, C. Tanner, T. Simuni, C. Coffey, K. Kieburtz, E. Flagg, S. Chowdhury et al., “The Parkinson progression marker initiative (PPMI),” *Progress in neurobiology*, vol. 95, no. 4, pp. 629–635, 2011.

[74] W. Poewe, K. Seppi, C. M. Tanner, G. M. Halliday, P. Brundin, J. Volkmann, A.-E. Schrag, and A. E. Lang, “Parkinson disease,” *Nature reviews Disease primers*, vol. 3, p. 17013, 2017.

[75] T. Li, R. Wernersson, R. B. Hansen, H. Horn, J. Mercer, G. Slodkowicz, C. T. Workman, O. Rigina, K. Rapacki, H. H. Stærfeldt et al., “A scored human protein–protein interaction network to catalyze genomic interpretation,” *Nature methods*, vol. 14, no. 1, p. 61, 2017.

[76] R. A. Fisher, “On the interpretation of $\chi 2$ from contingency tables, and the calculation of P,” *Journal of the Royal Statistical Society*, vol. 85, no. 1, pp. 87–94, 1922.

[77] D. Szklarczyk, A. Santos, C. von Mering, L. J. Jensen, P. Bork, and M. Kuhn, “STITCH 5: Augmenting protein–chemical interaction networks with tissue and affinity data,” *Nucleic acids research*, vol. 44, no. D1, pp. D380–D384, 2015.

[78] J. Kim, N. Russell, and J. Peng, “Scalable visualization for high-dimensional single-cell data,” in *Pacific Symposium on Biocomputing 2017*. World Scientific, 2017, pp. 623–634.

[79] B. M. Marlin, D. C. Kale, R. G. Khemani, and R. C. Wetzel, “Unsupervised pattern discovery in electronic health care data using probabilistic clustering models,” in *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*. ACM, 2012, pp. 389–398.

[80] D. Gotz, J. Sun, N. Cao, and S. Ebadollahi, “Visual cluster analysis in support of clinical decision intelligence,” in *AMIA Annual Symposium Proceedings*, vol. 2011. American Medical Informatics Association, 2011, p. 481.

[81] N. Cao, D. Gotz, J. Sun, and H. Qu, “DICON: Interactive visual analysis of multidimensional clusters,” *IEEE transactions on visualization and computer graphics*, vol. 17, no. 12, pp. 2581–2590, 2011.

[82] A. Perer, F. Wang, and J. Hu, “Mining and exploring care pathways from electronic medical records with visual analytics,” *Journal of biomedical informatics*, vol. 56, pp. 369–378, 2015.

[83] L. v. d. Maaten and G. Hinton, “Visualizing data using t-SNE,” *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[84] C. Ramaker, J. Marinus, A. M. Stiggelbout, and B. J. Van Hilten, "Systematic evaluation of rating scales for impairment and disability in Parkinson's disease," *Movement disorders: Official journal of the Movement Disorder Society*, vol. 17, no. 5, pp. 867–876, 2002.

[85] C. Shi, Z. Zheng, Q. Wang, C. Wang, D. Zhang, M. Zhang, P. Chan, and X. Wang, "Exploring the effects of genetic variants on clinical profiles of Parkinson's disease assessed by the unified Parkinson's disease rating scale and the Hoehn-Yahr stage," *PloS one*, vol. 11, no. 6, p. e0155758, 2016.

[86] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN," *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 3, p. 19, 2017.

[87] M. W. Johns, "A new method for measuring daytime sleepiness: The Epworth sleepiness scale," *sleep*, vol. 14, no. 6, pp. 540–545, 1991.

[88] A. A. Ejaz, I. S. Sekhon, and S. Munjal, "Characteristic findings on 24-h ambulatory blood pressure monitoring in a series of patients with Parkinson's disease," *European journal of internal medicine*, vol. 17, no. 6, pp. 417–420, 2006.

[89] S. M. Goldman, C. M. Tanner, D. Oakes, G. S. Bhudhikanok, A. Gupta, and J. W. Langston, "Head injury and Parkinson's disease risk in twins," *Annals of neurology*, vol. 60, no. 1, pp. 65–72, 2006.

[90] R. A. Hauser, M. Panisset, G. Abbruzzese, L. Mancione, N. Dronamraju, A. Kakarieka, and F.-S. S. Group, "Double-blind trial of levodopa/carbidopa/entacapone versus levodopa/carbidopa in early Parkinson's disease," *Movement Disorders*, vol. 24, no. 4, pp. 541–550, 2009.

[91] R. D. Zafonte, J. Lexell, and N. Cullen, "Possible applications for dopaminergic agents following traumatic brain injury: Part 2," *The Journal of head trauma rehabilitation*, vol. 16, no. 1, pp. 112–116, 2001.

[92] J. J. Gugger and M. L. Wagner, "Neurology: Rapid eye movement sleep behavior disorder," *Annals of Pharmacotherapy*, vol. 41, no. 11, pp. 1833–1841, 2007.

[93] M. Stacy, "Sleep disorders in Parkinson's disease," *Drugs & aging*, vol. 19, no. 10, pp. 733–739, 2002.

[94] M. Hallett and S. Khoshbin, "A physiological mechanism of bradykinesia," *Brain*, vol. 103, no. 2, pp. 301–314, 1980.

[95] W. H. Jost and D. Angersbach, "Ropinirole, a non-ergoline dopamine agonist," *CNS drug reviews*, vol. 11, no. 3, pp. 253–272, 2005.

[96] T. Yeh, Y. Huang, H. Wang, and S. Pan, "Spinal cord injury and Parkinson's disease: A population-based, propensity score-matched, longitudinal follow-up study," *Spinal cord*, vol. 54, no. 12, p. 1215, 2016.

[97] J. Kornhuber, G. Quack, W. Danysz, K. Jellinger, W. Danielczyk, W. Gsell, and P. Riederer, "Therapeutic brain concentration of the NMDA receptor antagonist amantadine," *Neuropharmacology*, vol. 34, no. 7, pp. 713–721, 1995.

[98] A. I. Faden, J. Ellison, and L. Noble, "Effects of competetive and non-competetive NMDA receptor antagonists in spinal cord injury," *European journal of pharmacology*, vol. 175, no. 2, pp. 165–174, 1990.

[99] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300, 1995.

[100] P. D. Turney, "Mining the web for synonyms: PMI-IR versus LSA on TOEFL," in *European Conference on Machine Learning.* Springer, 2001, pp. 491–502.

[101] E. S. Chen, G. Hripcsak, H. Xu, M. Markatou, and C. Friedman, "Automated acquisition of disease–drug knowledge from biomedical and clinical documents: An initial study," *Journal of the American Medical Informatics Association*, vol. 15, no. 1, pp. 87–98, 2008.

[102] G. E. Caughey, E. N. Ramsay, A. I. Vitry, A. L. Gilbert, M. A. Luszcz, P. Ryan, and E. E. Roughead, "Comorbid chronic diseases, discordant impact on mortality in older people: A 14-year longitudinal population study," *Journal of Epidemiology & Community Health*, pp. jech–2009, 2010.

[103] Q. T. Zeng, D. Redd, T. Rindflesch, and J. Nebeker, "Synonym, topic model and predicate-based query expansion for retrieving clinical documents," in *AMIA Annual Symposium Proceedings*, vol. 2012. American Medical Informatics Association, 2012, p. 1050.

[104] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[105] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[106] D. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum, "Polylingual topic models," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2.* Association for Computational Linguistics, 2009, pp. 880–889.

[107] I. Vulić and M.-F. Moens, "Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, vol. 2, 2015, pp. 719–725.

[108] A. K. McCallum, "MALLET: A machine learning for language toolkit," 2002, http://mallet.cs.umass.edu.

[109] C. Zhai and S. Massung, *Text data management and analysis: A practical introduction to information retrieval and text mining.* Morgan & Claypool, 2016.

[110] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Transactions on Information Systems (TOIS)*, vol. 20, no. 4, pp. 422–446, 2002.

[111] T. Williamson, "An attempt to estimate some of the characteristic marks, by which to judge of the cause of perforations of the stomach," *Dublin Journal of Medical Science (1836-1845)*, vol. 19, no. 2, pp. 191–220, 1841.

[112] J. Gao, J. Zhang, W. He et al., "Research of clinical distribution law for rheumatoid arthritis in traditional Chinese medicine syndrome type," *Hebei Journal of Traditional Chinese Medicine*, vol. 9, p. 006, 2012.

[113] J. A. S. Suryawanshi, "An overview of Citrus aurantium used in treatment of various diseases," *African Journal of Plant Science*, vol. 5, no. 7, pp. 390–395, 2011.

[114] T. M. Moraes, H. Kushima, F. C. Moleiro, R. C. Santos, L. R. M. Rocha, M. O. Marques, W. Vilegas, and C. A. Hiruma-Lima, "Effects of limonene and essential oil from Citrus aurantium on gastric mucosa: Role of prostaglandins and gastric mucus secretion," *Chemico-Biological Interactions*, vol. 180, no. 3, pp. 499–505, 2009.

[115] X. Ji, B. Huang, G. Wang, and C. Zhang, "The ethnobotanical, phytochemical and pharmacological profile of the genus Pinellia," *Fitoterapia*, vol. 93, pp. 1–17, 2014.

[116] H. Jain, C. Thao, and H. Zhao, "Enhancing electronic medical record retrieval through semantic query expansion," *Information systems and e-business management*, vol. 10, no. 2, pp. 165–181, 2012.

[117] M. D. Leiserson, M. A. Reyna, and B. J. Raphael, "A weighted exact test for mutually exclusive mutations in cancer," *Bioinformatics*, vol. 32, no. 17, pp. i736–i745, 2016.

[118] R. S. Wilson, E. Segawa, P. A. Boyle, S. E. Anagnos, L. P. Hizel, and D. A. Bennett, "The natural history of cognitive decline in Alzheimer's disease." *Psychology and aging*, vol. 27, no. 4, p. 1008, 2012.

[119] C. Ballard, S. Gauthier, A. Corbett, C. Brayne, D. Aarsland, and E. Jones, "Alzheimer's disease," *The Lancet*, vol. 377, no. 9770, Mar 2011.

[120] L. E. Hebert, J. Weuve, P. A. Scherr, and D. A. Evans, "Alzheimer disease in the United States (2010–2050) estimated using the 2010 census," *Neurology*, vol. 80, no. 19, pp. 1778–1783, 2013.

[121] S. Sveinbjornsdottir, "The clinical symptoms of Parkinson's disease," *Journal of neurochemistry*, vol. 139, pp. 318–324, 2016.

[122] W. R. Galpern and A. E. Lang, “Interface between tauopathies and synucleinopathies: A tale of two proteins,” *Annals of neurology*, vol. 59, no. 3, pp. 449–458, 2006.

[123] T. Vos, C. Allen, M. Arora, R. M. Barber, Z. A. Bhutta, A. Brown, A. Carter, D. C. Casey, F. J. Charlson, A. Z. Chen et al., “Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: A systematic analysis for the global burden of disease study 2015,” *The Lancet*, vol. 388, no. 10053, pp. 1545–1602, 2016.

[124] A. Association et al., “2018 Alzheimer’s disease facts and figures,” *Alzheimer’s & Dementia*, vol. 14, no. 3, pp. 367–429, 2018.

[125] M. Kivipelto, T. Ngandu, T. Laatikainen, B. Winblad, H. Soininen, and J. Tuomilehto, “Risk score for the prediction of dementia risk in 20 years among middle aged people: A longitudinal, population-based study,” *The Lancet Neurology*, vol. 5, no. 9, pp. 735–741, 2006.

[126] P. A. Kempster, S. S. Osullivan, J. L. Holton, T. Revesz, and A. J. Lees, “Relationships between age and late progression of Parkinson’s disease: A clinico-pathological study,” *Brain*, vol. 133, no. 6, pp. 1755–1762, 2010.

[127] Y. Compta, L. Parkkinen, S. S. O’sullivan, J. Vandrovcova, J. L. Holton, C. Collins, T. Lashley, C. Kallis, D. R. Williams, R. de Silva et al., “Lewy- and Alzheimer-type pathologies in Parkinson’s disease dementia: Which is more important?” *Brain*, vol. 134, no. 5, pp. 1493–1505, 2011.

[128] T. Polvikoski, R. Sulkava, M. Haltia, K. Kainulainen, A. Vuorio, A. Verkkoniemi, L. Niinistö, P. Halonen, and K. Kontula, “Apolipoprotein E, dementia, and cortical deposition of β-amyloid protein,” *New England Journal of Medicine*, vol. 333, no. 19, pp. 1242–1248, 1995.

[129] T. L. Edwards, W. K. Scott, C. Almonte, A. Burt, E. H. Powell, G. W. Beecham, L. Wang, S. Züchner, I. Konidari, G. Wang et al., “Genome-wide association study confirms SNPs in SNCA and the MAPT region as common risk factors for Parkinson disease,” *Annals of human genetics*, vol. 74, no. 2, pp. 97–109, 2010.

[130] D. Weintraub, N. Dietz, J. E. Duda, D. A. Wolk, J. Doshi, S. X. Xie, C. Davatzikos, C. M. Clark, and A. Siderowf, “Alzheimer’s disease pattern of brain atrophy predicts cognitive decline in Parkinson’s disease,” *Brain*, vol. 135, no. 1, pp. 170–180, 2011.

[131] Z. S. Nasreddine, N. A. Phillips, V. Bédirian, S. Charbonneau, V. Whitehead, I. Collin, J. L. Cummings, and H. Chertkow, “The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment,” *Journal of the American Geriatrics Society*, vol. 53, no. 4, pp. 695–699, 2005.

[132] J. Dalrymple-Alford, M. MacAskill, C. Nakas, L. Livingston, C. Graham, G. Crucian, T. Melzer, J. Kirwan, R. Keenan, S. Wells et al., “The MoCA: Well-suited screen for cognitive impairment in Parkinson disease,” *Neurology*, vol. 75, no. 19, pp. 1717–1725, 2010.

[133] S. Hoops, S. Nazem, A. Siderowf, J. Duda, S. Xie, M. Stern, and D. Weintraub, "Validity of the MoCA and MMSE in the detection of MCI and dementia in Parkinson disease," *Neurology*, vol. 73, no. 21, pp. 1738–1745, 2009.

[134] Y. Berlyand, D. Weintraub, S. X. Xie, I. A. Mellis, J. Doshi, J. Rick, J. McBride, C. Davatzikos, L. M. Shaw, H. Hurtig et al., "An Alzheimer's disease-derived biomarker signature identifies Parkinson's disease patients with dementia," *PloS one*, vol. 11, no. 1, p. e0147319, 2016.

[135] A. Jorm, A. Korten, and A. Henderson, "The prevalence of dementia: A quantitative integration of the literature," *Acta psychiatrica scandinavica*, vol. 76, no. 5, pp. 465–479, 1987.

[136] E. Cereda, R. Cilia, C. Klersy, C. Siri, B. Pozzi, E. Reali, A. Colombo, A. L. Zecchinelli, C. B. Mariani, S. Tesei et al., "Dementia in Parkinson's disease: Is male gender a risk factor?" *Parkinsonism & related disorders*, vol. 26, pp. 67–72, 2016.

[137] S. G. Younkin, "The role of Aβ42 in Alzheimer's disease," *Journal of Physiology-Paris*, vol. 92, no. 3-4, pp. 289–292, 1998.

[138] P. A. Thomann, E. Kaiser, P. Schönknecht, J. Pantel, M. Essig, and J. Schröder, "Association of total tau and phosphorylated tau 181 protein levels in cerebrospinal fluid with cerebral atrophy in mild cognitive impairment and Alzheimer disease," *Journal of psychiatry & neuroscience: JPN*, vol. 34, no. 2, p. 136, 2009.

[139] L. M. Shaw, H. Vanderstichele, M. Knapik-Czajka, C. M. Clark, P. S. Aisen, R. C. Petersen, K. Blennow, H. Soares, A. Simon, P. Lewczuk et al., "Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects," *Annals of neurology*, vol. 65, no. 4, pp. 403–413, 2009.

[140] J.-H. Kang, D. J. Irwin, A. S. Chen-Plotkin, A. Siderowf, C. Caspell, C. S. Coffey, T. Waligórska, P. Taylor, S. Pan, M. Frasier et al., "Association of cerebrospinal fluid β-amyloid 1-42, T-tau, P-tau181, and α-synuclein levels with clinical features of drug-naive patients with early Parkinson disease," *JAMA neurology*, vol. 70, no. 10, pp. 1277–1287, 2013.

[141] D. J. Irwin, M. Grossman, D. Weintraub, H. I. Hurtig, J. E. Duda, S. X. Xie, E. B. Lee, V. M. Van Deerlin, O. L. Lopez, J. K. Kofler et al., "Neuropathological and genetic correlates of survival and dementia onset in synucleinopathies: A retrospective analysis," *The Lancet Neurology*, vol. 16, no. 1, pp. 55–65, 2017.

[142] E. H. Corder, A. M. Saunders, W. J. Strittmatter, D. E. Schmechel, P. C. Gaskell, G. Small, A. D. Roses, J. Haines, and M. A. Pericak-Vance, "Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families," *Science*, vol. 261, no. 5123, pp. 921–923, 1993.

[143] S. Sadigh-Eteghad, M. Talebi, and M. Farhoudi, “Association of apolipoprotein E epsilon 4 allele with sporadic late onset Alzheimer’s disease,” *A meta-analysis. Neurosciences (Riyadh)*, vol. 17, no. 4, pp. 321–326, 2012.

[144] D. Aarsland, B. Creese, M. Politis, K. R. Chaudhuri, D. Weintraub, C. Ballard et al., “Cognitive decline in Parkinson disease,” *Nature Reviews Neurology*, vol. 13, no. 4, p. 217, 2017.

[145] B. Fischl, “FreeSurfer,” *Neuroimage*, vol. 62, no. 2, pp. 774–781, 2012.

[146] K. Pearson, “LIII. on lines and planes of closest fit to systems of points in space,” *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[147] D. C. Liu and J. Nocedal, “On the limited memory BFGS method for large scale optimization,” *Mathematical programming*, vol. 45, no. 1-3, pp. 503–528, 1989.

[148] F. N. Emamzadeh, “Role of apolipoproteins and α-synuclein in Parkinson’s disease,” *Journal of Molecular Neuroscience*, vol. 62, no. 3-4, pp. 344–355, 2017.

[149] N. Pankratz, L. Byder, C. Halter, A. Rudolph, C. W. Shults, P. M. Conneally, T. Foroud, and W. C. Nichols, “Presence of an APOE4 allele results in significantly earlier onset of Parkinson’s disease and a higher risk with dementia,” *Movement disorders*, vol. 21, no. 1, pp. 45–49, 2006.

[150] M. Federoff, B. Jimenez-Rolando, M. A. Nalls, and A. B. Singleton, “A large study reveals no association between APOE and Parkinson’s disease,” *Neurobiology of disease*, vol. 46, no. 2, pp. 389–392, 2012.

[151] I. F. Mata, J. B. Leverenz, D. Weintraub, J. Q. Trojanowski, H. I. Hurtig, V. M. Van Deerlin, B. Ritz, R. Rausch, S. L. Rhodes, S. A. Factor et al., “APOE, MAPT, and SNCA genes and cognitive performance in Parkinson disease,” *JAMA neurology*, vol. 71, no. 11, pp. 1405–1412, 2014.

[152] A. Calderone, M. Formenti, F. Aprea, M. Papa, L. Alberghina, A. M. Colangelo, and P. Bertolazzi, “Comparing Alzheimer’s and Parkinson’s diseases networks using graph communities structure,” *BMC systems biology*, vol. 10, no. 1, p. 25, 2016.

[153] J. D. Jones, T. P. Kuhn, and S. M. Szymkowicz, “Reverters from PD-MCI to cognitively intact are at risk for future cognitive impairment: Analysis of the PPMI cohort,” *Parkinsonism & related disorders*, vol. 47, pp. 3–7, 2018.

[154] M. C. Tremblay, D. J. Berndt, S. L. Luther, P. R. Foulis, and D. D. French, “Identifying fall-related injuries: Text mining the electronic medical record,” *Information Technology and Management*, vol. 10, no. 4, p. 253, 2009.

[155] D. T. Heinze, M. L. Morsch, and J. Holbrook, “Mining free-text medical records.” in *Proceedings of the AMIA Symposium.* American Medical Informatics Association, 2001, p. 254.

[156] R. M. Anholt, J. Berezowski, I. Jamal, C. Ribble, and C. Stephen, "Mining free-text medical records for companion animal enteric syndrome surveillance," *Preventive veterinary medicine*, vol. 113, no. 4, pp. 417–422, 2014.

[157] M. Kushima, K. Araki, M. Suzuki, S. Araki, and T. Nikama, "Text data mining of in-patient nursing records within electronic medical records using KeyGraph," *IAENG International Journal of Computer Science*, vol. 38, no. 3, pp. 215–224, 2011.

[158] R. Takata, T. Katagiri, M. Kanehira, T. Tsunoda, T. Shuin, T. Miki, M. Namiki, K. Kohri, Y. Matsushita, T. Fujioka et al., "Predicting response to methotrexate, vinblastine, doxorubicin, and cisplatin neoadjuvant chemotherapy for bladder cancers through genome-wide gene expression profiling," *Clinical cancer research*, vol. 11, no. 7, pp. 2625–2636, 2005.

[159] B. M. Ghadimi, M. Grade, M. J. Difilippantonio, S. Varma, R. Simon, C. Montagna, L. Füzesi, C. Langer, H. Becker, T. Liersch et al., "Effectiveness of gene expression profiling for response prediction of rectal adenocarcinomas to preoperative chemoradiotherapy," *Journal of clinical oncology: Official journal of the American Society of Clinical Oncology*, vol. 23, no. 9, p. 1826, 2005.

[160] J. C. Chang, E. C. Wooten, A. Tsimelzon, S. G. Hilsenbeck, M. C. Gutierrez, R. Elledge, S. Mohsin, C. K. Osborne, G. C. Chamness, D. C. Allred et al., "Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer," *The Lancet*, vol. 362, no. 9381, pp. 362–369, 2003.

[161] P. Geeleher, N. J. Cox, and R. S. Huang, "Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines," *Genome biology*, vol. 15, no. 3, p. R47, 2014.

[162] C. Becquet, S. Blachon, B. Jeudy, J.-F. Boulicaut, and O. Gandrillon, "Strong-association-rule mining for large-scale gene-expression data analysis: A case study on human sage data," *Genome Biology*, vol. 3, no. 12, pp. research0067–1, 2002.

[163] C. Creighton and S. Hanash, "Mining gene expression databases for association rules," *Bioinformatics*, vol. 19, no. 1, pp. 79–86, 2003.