MACHINE LEARNING OF STRUCTURED AND UNSTRUCTURED HEALTHCARE DATA

By

SUHAO CHEN

Bachelor of Management in Information Management and Systems
Nanjing University
Nanjing, China
2007

Master of Management in Corporate Management
Shanghai Jiao Tong University
Shanghai, China
2010

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
July, 2022

MACHINE LEARNING OF STRUCTURED AND UNSTRUCTURED HEALTHCARE DATA

Dissertation Approved:

Dr. Tieming Liu

Dissertation Advisor

Dr. Sunderesh S. Heragu

Dr. Bing Yao

Dr. Zheyu Jiang

ACKNOWLEDGMENTS

I would like to express my boundless gratitude to my Ph.D. advisor Dr. Tieming Liu for his academic mentoring, comprehensive guidance, constant encouragement, and full support during my study. I would not be able to complete this journey without his tremendous help.

I would like to thank my committee members, Dr. Sunderesh S. Heragu, Dr. Bing Yao, and Dr. Zheyu Jiang for their insightful suggestions and unconditional support.

My appreciation also goes to my collaborators including Dr. Bing Yao, Dr. Zhuqi Miao, Dr. Thanh Thieu, Dr. Yajun Lu, Zekai Wang, Tuan-Dung Le, and Dr. Lin Guo for their help and inspiration.

I am always grateful to be a part of the School of Industrial Engineering and Management (IEM) and deeply appreciate the help I have received from all the extraordinary professors and staff members in the IEM family.

I acknowledge IEM and the Center for Health Systems Innovation for financial support, the Cerner Corporation for sharing the Health Facts® EHR database, and the High Performance Computing Center at Oklahoma State University for computing support.

My friends and lab-mates at Oklahoma State University are kind and helpful. I have a lot of sweet memories in Oklahoma because of them. I thank them for the friendship.

Lastly, I am thankful for my family for their understanding, patience, and encouragement.

Acknowledgments reflect the views of the author and are not endorsed by committee members or Oklahoma State University.

Name: SUHAO CHEN

Date of Degree: JULY, 2022

Title of Study:

MACHINE LEARNING OF STRUCTURED AND UNSTRUCTURED HEALTHCARE DATA

Major Field: INDUSTRIAL ENGINEERING AND MANAGEMENT

Abstract: The widespread adoption of Electronic Health Records (EHR) systems in healthcare institutions in the United States makes machine learning based on large-scale and real-world clinical data feasible and affordable. Machine learning of healthcare data, or healthcare data analytics, has achieved numerous successes in various applications. However, there are still many challenges for machine learning of healthcare data both structured and unstructured. Longitudinal structured clinical data (e.g., lab test results, diagnoses, and medications) have an enormous variety of categories, are collected at irregularly spaced visits, and are sparsely distributed. Studies on analyzing longitudinal structured EHR data for tasks such as disease prediction and visualization are still limited. For unstructured clinical notes, existing studies mostly focus on disease prediction or cohort selection. Studies on mining clinical notes with the direct purpose to reduce costs for healthcare providers or institutions are limited. To fill in these gaps, this dissertation has three research topics.

The first topic is about developing state-of-the-art predictive models to detect diabetic retinopathy using longitudinal structured EHR data. Major deep-learning-based temporal models for disease prediction are studied, implemented, and evaluated. Experimental results on a large-scale dataset show that temporal deep learning models outperform non-temporal random forests models in terms of AUPRC and recall.

The second topic is about clustering temporal disease networks to visualize comorbidity progression. We propose a clustering technique to outline comorbidity progression phases as well as a new disease clustering method to simplify the visualization. Two case studies on Clostridioides difficile and stroke show the methods are effective.

The third topic is clinical information extraction for medical billing. We propose a framework that consists of two methods, a rule-based and a deep-learning-based, to extract patient history information directly from clinical notes to facilitate the Evaluation and Management Services (E/M) billing. Initial results of the two prototype systems on an annotated dataset are promising and direct us for potential improvements.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

1.1 Motivation

The healthcare industry in the United States faces many challenges including astonishing total expenditures, low quality, and high disparities [Sethi, 2009]. It is estimated that in 2020, health care spending in the US reached $4.1 trillion, a 9.7 percent increase from 2019 and representing almost 19.7 percent of the gross domestic product of the country in that year [Hartman et al., 2022]. However, the quality of healthcare is not satisfactory, and disparities are pervasive in the United States [Kelley et al., 2005]. Take vision care for example, 24% of US counties had no ophthalmologists or optometrists in 2011 [Gibson, 2015], making it hard for many patients to have quality eye care services. Therefore, it is of great importance to promote studies of healthcare data analytics to assist in preventive medicine, predictive medicine, or precision medicine so that we can reduce healthcare costs, improve patient care quality, and reduce healthcare disparities.

The widespread adoption of electronic health records (EHR) [Bardhan and Thouin, 2013, Huerta et al., 2013] and the increasing emphasis on the use of clinical decision support systems (CDSS) [Bright et al., 2012, Gupta and Sharda, 2013, Grout et al., 2018] have been two of the most remarkable outcomes of healthcare reform in the U.S. during the past decade. The adoption rate of basic EHR systems among U.S. hospitals has surged from 9.4% in 2008 to 83.8% in 2015 [Henry et al., 2016]. The ubiquitous adoption of EHR has generated an unprecedented amount of health data, which provide the longitudinal picture of patients' journeys, treatment pathways, and care outcomes [Moores, 2012]. A CDSS refers to "any electronic system designed to aid directly in clinical decision making, in which characteristics of individual patients are used to generate patient-specific assessments or recommendations that are then presented to clinicians for consideration [Kawamoto et al., 2005]".

The abundance and comprehensiveness of EHR data, in conjunction with recent advances in CDSS, has offered researchers and practitioners an ideal platform to mine actionable insights to improve clinical decision-making for better healthcare outcomes [Gupta and Sharda, 2013, Johnson et al., 2014, Fichman et al., 2011]. Specific applications include test ordering [Zhuang et al., 2013], therapy management [Yet et al., 2013], improving care delivery and access [Barjis et al., 2013, Li et al., 2017], detecting and predicting health conditions [Piri et al., 2017, Topuz et al., 2018], and medication evaluation [Van Valkenhoef et al., 2013].

However, there are still many challenges in analyzing healthcare data. One of the critical challenges is analyzing longitudinal or temporal structured healthcare data. Structured data refers to data stored as tables in relational databases that can be easily queried and processed. Healthcare data are inherently longitudinal, e.g., many diseases develop gradually and there may be crucial temporal correlations between health conditions for the progression. By using temporal reasoning or visualization in healthcare data analytics we can discover more hidden patterns or knowledge [Combi and Shahar, 1997]. However, longitudinal healthcare data are heterogeneous, are collected at irregularly spaced visits, and are sparsely distributed, making it a challenging task to analyze or visualize these temporal data. Existing studies on analytics and visualization of temporal healthcare data are still relatively limited.

Another challenge is analyzing unstructured healthcare data, especially clinical notes. Clinical notes usually contain more information (e.g., patient lifestyle, social status, and family history) and subtle descriptions of patient conditions or symptoms, making them an invaluable information source for healthcare data analytics. However, clinical notes are heterogeneous and require sophisticated text mining efforts. With the help of fast advancements in the field of natural language processing (NLP) and deep learning in recent years, the number of studies on mining clinical notes has increased significantly. However, existing studies mostly focus on disease prediction, cohort selection, or patient care. Studies on mining clinical notes with the direct purpose to reduce costs for healthcare institutions are quite limited.

This dissertation aims to develop methods to better analyze and visualize temporal structured EHR data as well as extract patient information from clinical notes for medical billing with the ultimate purpose to improve healthcare quality, reduce healthcare disparities, or reduce medical costs.

1.2 Problem Statements

Research Problem 1: How to build a temporal predictive model to analyze longitudinal structured EHR data to better predict the onset of chronic diseases (specifically, diabetic retinopathy)?

Longitudinal healthcare data may contain many crucial hidden insights and improve the performance of healthcare data analytics. Take the task of disease prediction for instance. Most chronic diseases develop over years. There are many risk factors that gradually contribute to the development of the diseases. A longitudinal observation of these factors may unfold more insights about the trajectory of the disease progression, thus making the disease prediction more accurate.

However, building a temporal disease prediction model is challenging due to the following data characteristics. First, there is a significant disparity in the number of health records for different patients in EHR systems. For example, for a specific lab test, some patients may have hundreds of results while some others may only have several or none. Second, for a single patient, temporal healthcare data are collected at irregularly spaced visits and irregular frequencies. For example, an ICU visit often has intensive lab tests, a routine well-being check only has limited lab tests while a consultation visit usually does not have lab tests. What's more, many healthcare data have a variety of categories, are sparsely distributed, and are imbalanced.

This dissertation specifically focuses on the prediction of diabetic retinopathy (DR), a major complication of diabetes [CDC, 2021] (with blood vessel damage in the retina illustrated in Figure 1.1). DR is a leading cause of blindness in working-age adults globally [Yau et al., 2012, Ting et al., 2016]. If diagnosed at an early stage, DR can be effectively treated or even cured with intensive therapy. However, the compliance rate of DR screening remains low due to the hurdles of current DR screening methods.

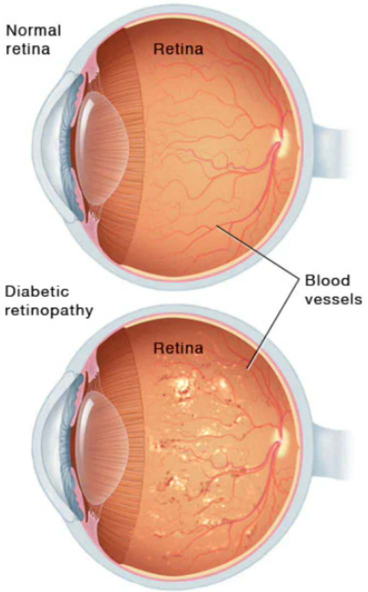


Figure 1.1: Blood Vessel Damage in the Retina

Figure source: Mayo Clinic, www.mayoclinic.org/diseases-conditions/diabetic-retinopathy/symptoms-causes/syc-20371611

There are two major DR screening methods, fundus exams, and fundus photography. Fundus exams are performed by an ophthalmologist or optometrist using a binocular indirect ophthalmoscope or a slit lamp, which requires ophthalmic skills. Dilation, which uses specialized eye drops to enlarge pupils, is often needed for such exams. A comprehensive eye exam can cost between $170-200 without insurance [Kraff, 2020]. The requirement for ophthalmic skills and the high costs of exams make this kind of screening difficult to access for patients in rural, remote, or other medically underserved communities.

Another screening method is fundus photography, which detects retinal abnormalities by taking fundus images. Fundus photography is generally performed by trained physicians or certified technicians using expensive equipment such as optical coherence tomography (OCT) machines and digital fundus cameras. The images are later sent to ophthalmologists for examinations. In addition to high equipment costs, certified ophthalmologists or specialists are still needed to analyze and interpret the fundus images either at point-of-care or remotely.

Recently, artificial intelligence (AI) algorithms, particularly deep learning algorithms, have been extensively studied to analyze fundus photographs [Gulshan et al., 2016, Ting et al., 2017]. Since 2018, the Food and Drug Administration has approved a few AI-based DR screening systems including IDx-DR [Abràmoff et al., 2018] and EyeArt [Bhaskaranand et al., 2019]. Such systems utilize deep learning techniques to automatically analyze fundus images and certified specialists are no longer needed. This relaxes the need for ophthalmologists or certified specialists, and thus expands DR screening to more healthcare settings. However, expensive fundus cameras are still needed for retinal imaging, which limits the use of this screening approach to only well-funded healthcare providers. In addition, retinal imaging is technically challenging, potentially hindering its adoption in resource-limited settings, like rural primary care clinics, where providers have limited experience with ophthalmic imaging.

Therefore, there is an urgent need to develop non-image-based DR screening tools that only use common Electronic Health Records (EHR) data (e.g., patient demographics, diagnoses, lab tests, medications, and procedures) and are accessible to all healthcare settings.

Research Problem 2: How to construct comorbidity networks to group patients into phases more reasonably and better visualize temporal comorbidity progression?

Like disease prediction, healthcare data visualization is another task that we would like to utilize temporal analytics. A cross-sectional visualization of healthcare data can help us identify some patterns or associations. For example, the percentage of male patients in the age range of 50 to 65 who are diagnosed with diabetes may be higher than that of female patients in the same age range. However, such a cross-sectional visualization neglects the longitudinal nature of healthcare data and cannot effectively explain many important clinical issues such as comorbidity progression.

Temporal disease networks are commonly used to visualize longitudinal data. For example, researchers use disease networks to visualize comorbidity progression. General practice is that patients are grouped into different progression phases based on some kind of time attribute (e.g., length of stay in hospital), comorbidity networks are constructed for these phases separately, and then the networks across different phases are compared.

However, there are still some challenges, such as how to outline the phases. Existing studies assign phases either by the same length of time or the same number of patients within each phase. Whereas such simple methods are arbitrary and may not effectively find the boundaries of phases and thus may not reveal insights about comorbidity progression. Another challenge is that the number of comorbidities is enormous, which makes comorbidity networks constructed still too complicated for visualization.

Research Problem 3: How to extract patient history information from clinical notes to help automate the medical billing process to reduce billing costs?

The last research problem in this dissertation is about analyzing clinical notes. Healthcare data are heterogeneous. Besides structured database tables, healthcare data are also stored in unstructured formats such as clinical notes, patient questionnaires, radiology images, and even speeches and videos. In fact, researchers estimate that about 80% of medical data are unstructured [Kong, 2019, Assale et al., 2019], often in the form of free-text notes [Meystre et al., 2008]. Compared to structured database tables (e.g., lab test tables and diagnosis tables), clinical notes contain more detailed information about patients (e.g., patient living habits, social status, and family history) and subtle descriptions of conditions or symptoms, making them an important data source for healthcare data analytics.

In the past decades, studies mining clinical notes have increased substantially. However, there are some challenges. One challenge is the scarcity of annotated clinical notes for supervised learning. Annotating clinical notes is time-consuming and may expose patient protected health information (PHI). The scarcity of publicly available annotated datasets limits the progress of studies mining clinical notes. Another challenge is that the number of studies mining clinical notes with a direct purpose to reduce costs (e.g., medical billing costs) for healthcare institutions is limited. Most existing studies focus on disease prediction and cohort selection to improve patient care.

In this dissertation, we aim to reduce medical billing costs. Medical billing is one of the heavy burdens facing healthcare institutions in the United States. In the modern healthcare ecosystem, financial intermediaries such as private insurance companies and government programs (e.g., Medicare and Medicaid) serve as payers to providers' health services. To ensure accurate reimbursement and manage the quality of care, paying intermediaries request alphanumeric codes of diagnoses and procedures performed at a patient's visit. As a result, coding for medical claims, also known as medical billing, becomes one of the most important tasks in the healthcare revenue management cycle.

At the core of medical billing are Evaluation and Management (E/M) services codes, which are a category of Current Procedural Terminology (CPT) codes specific for billing purposes. In addition to the Centers for Medicare and Medicaid Services (CMS)'s 1997 guidelines [CMS, 1997], American Medical Association (AMA) recently published simplified guidelines specific for office and outpatient visits [AMA, 2019]. According to the 1997 guidelines, the three key components of documentation needed to support the selection of an appropriate level of E/M services furnished at a patient's visit are *history*, *examination*, and *medical decision making*. The new simplified guidelines also require medically appropriate history and/or examination although the extent of them is not used for the selection of E/M service codes.

Unfortunately, gleaning and coding the billing information is still highly relied on the manual processing by clinical coders to date. Among large healthcare providers, in-house coding professionals are often hired while smaller healthcare organizations commonly outsourced their coding tasks. On the other side of the transaction, it costs paying intermediaries an equivalent amount of manpower to evaluate claims, request clarification and re-submission, and detect and penalize frauds such as up-coding/over-coding and false charges.

The entire process has brought tremendous workload and financial burdens to care providers, staff, professional coders, and payers. Studies have shown that physicians in the United States spent on average from 17% to 43% of their time with EHR systems for documentary tasks [Sinsky et al., 2016, Tai-Seale et al., 2017, Arndt et al., 2017, Woolhandler and Himmelstein, 2014]. Such administrative responsibilities took them away from patients and lowered their career satisfaction [Woolhandler and Himmelstein, 2014]. On the payers' side, such costly human processes have under-met the sheer volume of reimbursement claims, resulting in billing errors that cost U.S. tax and insurance payers a magnitude of billions of dollars [Champagnie, 2019]. For example, the total value of challenged claims was estimated from $11 billion to $54 billion annually [Gottlieb et al., 2018]. The overall billing and insurance-related administrative costs for the whole healthcare revenue management cycle in the United States approximated $471 billion a year [Jiwani et al., 2014]. High cost and error volume undermined the effectiveness of current billing practices.

Despite the critical need for automatic technologies that can accurately recognize billing information from clinical free-text notes, the research in this field remains limited to date. Due to the complexity of E/M billing, this dissertation will focus on the extraction of the first component (i.e., patient history information), annotate a dataset, and develop models and prototype systems using public or academic available resources with the ultimate purpose to facilitate the billing practice.

1.3 Research Objectives

The research objectives of this dissertation are as follows.

- Designing state-of-the-art temporal predictive models to analyze real-world, imbalanced, and longitudinal structured EHR data to predict diabetic retinopathy
- Conceiving a method to model comorbidity progression using temporal disease networks for real-world EHR data and cluster the networks into progression phases
- Creating a method to cluster diseases to simplify the visualization of comorbidity progression
- Establishing methods and prototype systems to extract essential patient history information from clinical notes to facilitate medical billing
- Designing a comprehensive model evaluation metric for named entity recognition including a taxonomy to quantify outcomes for notes with text span overlapping entities

1.4 Expected Contributions

In this dissertation, we have three research topics. The first topic is about designing predictive models to analyze longitudinal structured EHR data to detect patients with diabetic retinopathy. The second topic is about constructing temporal comorbidity networks and clustering these networks and diseases to visualize comorbidity progression. The third topic is about constructing methods and systems to extract essential patient history information from clinical notes for medical billing and comparing the performances of the systems.

For the first topic, the contributions include two aspects.

- In the methodological aspect, to the best of our knowledge, it is the first study building deep learning architectures to analyze longitudinal structured EHR data (e.g., lab tests) for DR prediction. Previous DR prediction models built on structured EHR data only analyze cross-sectional or aggregated data.

- In the application aspect, deep learning models outperformed non-temporal models, which indicates a better alternative DR screening method. Only a small set of common variables are used as input data, making the models accessible and easier to deploy.

For the second topic, this research has the below contributions.

- In the methodological aspect, it proposes a new method to cluster temporal disease networks into consecutive progression phases and a new method to cluster highly associated diseases into groups to simplify the visualization of comorbidity progression.

- In the application aspect, the method to construct disease networks based on Clinical Classifications Software (CCS) categories is easier than conventional diagnosis codes and the system can be integrated into clinical decision support systems to visualize comorbidity progression.

For the third topic, the contributions include two aspects.

- In the methodological aspect, to the best of our knowledge, this is the first study focusing on clinical information extraction for the Evaluation/Management (E/M) medical billing. It proposes a framework of two methods to extract patient history information. A comprehensive metric is also proposed to evaluate named entity recognition performances including an exact-match metric and a novel hierarchical relaxed-match metric suitable for notes with text span overlapping entities.

- In the application aspect, this study provides technical solutions to develop libraries, knowledge extraction rules, and deep learning architectures, and can be helpful in real medical billing settings.

1.5 Organization of the Dissertation

The rest of this dissertation is organized as follows. Chapter 2 presents a literature review on the topics of this dissertation. Chapter 3 builds DR temporal prediction models by analyzing longitudinal structured EHR data and compares the performances of these models. Chapter 4 introduces a mechanism to construct temporal disease networks for comorbidities and two clustering methods for better visualization of comorbidity progression. Chapter 5 constructs a rule-based and a deep-learning-based systems to extract essential history information from clinical notes for E/M medical billing and compares the performances of the two systems. The last chapter summarizes and concludes this dissertation.

CHAPTER II

LITERATURE REVIEW

2.1 Diabetic Retinopathy Prediction Using Longitudinal EHR Data

2.1.1 Diseases Prediction Using Longitudinal EHR Data

Longitudinal healthcare data are considered to be an invaluable source for healthcare machine learning [Moskovitch et al., 2019]. On one hand, longitudinal healthcare data can help us discover insights about potential changes in patient health conditions. For example, an unintentional weight loss may signal diabetes or a more severe illness [Williamson et al., 2000]. On the other hand, long-term temporal dependencies exist ubiquitously in health conditions [Pham et al., 2017]. For example, patients with a longer history of diabetes are more likely to develop diabetic retinopathy (DR) [Klein et al., 1984]. Therefore, analyzing longitudinal data is important for healthcare analytics [Combi and Shahar, 1997]. In fact, literature shows that predictive models built on longitudinal data have better performances [Singh et al., 2015, Gupta et al., 2020, Wang and Yao, 2022].

However, building disease prediction models to analyze longitudinal data is challenging. First, longitudinal healthcare data differ in collecting frequencies. Some data are collected at fixed intervals (e.g., electrocardiogram data) while others are collected at irregular frequencies (e.g., lab tests and diagnoses).

Second, there is a significant disparity in the number of temporal health records for different patients in EHR systems. Part of the disparity results from the difference in visit frequency across patients. Some patients visit hospitals only when they feel really necessary while some others visit hospitals much more frequently. Patients with complicated health conditions also tend to generate more records. Coverage of the dataset also accounts for this disparity. If a patient chooses to visit a new hospital that uses a different EHR system, the new records may not be included in the dataset.

Healthcare data in EHR are also high dimensional, sparsely distributed, and contain many missing values. Take diagnoses for example, there are about 13,000 ICD-9-CM diagnosis codes and about 68,000 ICD-10-CM codes. For most of these diagnosis codes, only a small fraction of the population has records. Therefore, if we use diagnosis codes as input data, the dataset is likely high dimensional and sparsely distributed with many missing values.

Due to the above characteristics of longitudinal healthcare data, conventional time series analysis methods which rely on the assumption of regularly sampled data are not suitable for temporal healthcare data analytics. Traditional machine learning algorithms are also not designed to analyze longitudinal data. Hence, some researchers proposed the knowledge-based temporal abstraction approach [Shahar, 1997]. There are three basic steps for this approach [Moskovitch et al., 2019]. The first step is to transform raw, irregularly time-stamped variables into symbolic time intervals using some kind of knowledge abstraction method (e.g., statistical mean or median or counts). The second step is to exact temporal patterns or relations in these symbolic intervals using techniques such as Allen's interval algebra [Allen, 1983]. The last step is to induce classifiers based on these patterns or relations for various machine learning tasks such as prediction.

We have many studies following this approach in the literature. An algorithm called "IEMiner" (Interval-based Event Miner) was proposed to discover frequent temporal patterns from interval-based events and was evaluated on hepatitis classification [Patel et al., 2008]. A framework named "KarmaLegoSification (KLS)" that can efficiently extract relations from symbolic time intervals for clinical event prediction was also proposed [Moskovitch and Shahar, 2015]. The time abstraction approach was also integrated with hidden Markov models for sepsis prediction [Gupta et al., 2020]. However, the process of temporal abstraction for temporal intervals has the problem of information loss, and extracting the temporal relations between time intervals is also demanding.

In the past several years, deep neural networks, or deep learning techniques, have been increasingly utilized for longitudinal healthcare data analytics due to their superior capacities to model sequential information with great flexibility [Xie et al., 2022]. Recurrent neural networks (RNN) [Rumelhart et al., 1986], particularly the sub-types of long short-term memory (LSTM) [Hochreiter and Schmidhuber, 1997] and gated recurrent units (GRU) [Cho et al., 2014], is one of the most popular deep learning techniques used. A reverse-time attention mechanism based on RNN was proposed and evaluated for the prediction of heart failure [Choi et al., 2016]. An end-to-end system named “DeepCare” which uses LSTM achieved better prediction performances than plain RNN and non-temporal models on datasets of diabetes and mental health [Pham et al., 2017]. Attention mechanism was introduced into RNN, and multi-task prediction layers were added to the architecture which demonstrated better prediction performances for bone diseases and cardiovascular disease [Suo et al., 2017]. Interested readers may refer to [Xie et al., 2022] for a comprehensive survey.

Recent studies show that temporal convolutional network (TCN) [Lea et al., 2016] which uses dilated causal convolutions has superior performances in modeling sequential data. An empirical evaluation found that generic TCN model outperforms canonical RNN models for a variety of temporal modeling tasks [Bai et al., 2018]. TCN was also employed to predict clinical events for ICU patients and achieved better performances than LSTM models [Catling and Wolff, 2020]. TCN has been successfully employed for the prediction of diseases including diabetes [Xie and Wang, 2020], influenza-like illness [Lee et al., 2021], depression [Du et al., 2019], and sepsis [Moor et al., 2019, Kok et al., 2020, Wang and Yao, 2022].

Although more and more studies are utilizing deep learning temporal models for disease prediction, the number of such studies is still limited, particularly for diseases with multiple types of risk factors (e.g., diabetes and its complications). Machine learning models trained and tested on real-world large-scale datasets for these diseases are even scarce.

2.1.2 Diabetic Retinopathy Prediction Using EHR Data

There are extensive clinical studies on DR risk factors in the literature [Mohamed et al., 2007, Stitt et al., 2016], which provide clinical insights for machine learning feature selection and thus significantly facilitate healthcare data analytics on DR prediction. Widely recognized risk factors of DR include patient demographics [Klein et al., 1984], duration of diabetes [Klein et al., 1984], lab tests [Olsen et al., 2000, Irace et al., 2011], and comorbidities (e.g., neuropathy [Candrilli et al., 2007], nephropathy [Cruickshanks et al., 1993], hypertension [Van Leiden et al., 2002], obesity [Van Leiden et al., 2002], and cardiovascular disease [Van Hecke et al., 2005]), among many others.

Meanwhile, machine learning researchers have constructed various models using common EHR data for DR prediction in the past decades. Cox's proportional hazard model was used to analyze patient demographics, blood tests, and some comorbidity variables for DR prediction [Semeraro et al., 2011]. Decision tree models were built to include more DR risk factors and performances of two ensemble approaches were compared [Ogunyemi and Kermah, 2015]. Multiple machine learning models (decision tree, random forests, logistic regress, and artificial neural networks) were utilized to analyze patient demographics and a small set of lab tests for DR prediction [Piri et al., 2017]. A DR risk index was recently proposed with ten demographics and lab test variables and a DR predictive model was then developed based on the risk index [Wang et al., 2021]. However, all these models are built on either cross-sectional or aggregated data and do not leverage the temporal information in the longitudinal EHR data for DR prediction.

2.1.3 Intellectual Gaps

Through a thorough literature review, we found that in the area of DR prediction, there has been limited work to:

- *Leverage state-of-the-art deep learning techniques to analyze structured longitudinal EHR data.* Existing research either built traditional machine learning models to analyze cross-sectional or aggregated data or employed deep learning techniques to analyze retinal fundus images.
- *Incorporate clinical domain knowledge into deep learning disease prediction models.* There is a lack of clinical domain knowledge in developing deep-learning-based temporal representation models [Xie et al., 2022]. Domain knowledge can be used to select those potentially important variables and make the models easier for deployment.

2.2 Visualization Analytics and Temporal Disease Network

2.2.1 Visualization Analytics for Comorbidity

Visual analytics (VA) can reduce the information overload on memory and cognition, and leverage the power of human perception [Caban and Gotz, 2015, Simpao et al., 2014]. Nowadays, it has become an integral component of clinical decision support systems [Mane et al., 2012, Simpao et al., 2015a, Simpao et al., 2015b, Rind et al., 2013, Nelson et al., 2019]. For example, through VA, large volumes of data and complex ideas in healthcare settings can be presented with clarity, accuracy, and efficiency in visual diagrams [Nadj et al., 2020, Kamsu-Foguem et al., 2012]. Furthermore, VA dashboards allow real-time monitoring and tracking of healthcare information, such as hospital-specific antibiograms [Simpao et al., 2018], adverse drug events [Sorbello et al., 2017], and departmental performance metrics [Karami and Safdari, 2016]. It also has been reported that visualized data improved recall of important clinical information [Tscholl et al., 2018].

An emerging and important direction of VA in clinical decision-making is to visualize and mine comorbidity progression patterns [Hidalgo et al., 2009, Warner et al., 2013, Krishnamurthy et al., 2018, Wang et al., 2020, Hossain et al., 2020]. Comorbidity refers to one or more other health conditions coexisting with a particular index disease under investigation [Feinstein, 1970]. Comorbidity has been increasingly prevalent [Divo et al., 2014] and consistently challenging healthcare practice and research by leading to worse health outcomes, complicating diagnostics and treatments, and misleading medical statistics [Feinstein, 1970, Gijsen et al., 2001]. As a result, great efforts have been devoted to exploring effective methodology to handle comorbidity to improve clinical decision-making during the past few decades [De Groot et al., 2003, Capobianco and Lio, 2013, Zolbanin et al., 2015].

Network modeling represents an intuitive and useful approach to investigating comorbidities and their progression patterns [Cramer et al., 2010, Barabási et al., 2011, Brunson and Laubenbacher, 2018] mainly for the following advantages

- *User-friendly presentation of disease associations.* By modeling comorbidities as nodes and their pairwise associations as edges, network models can present comorbidity visually. The nodes and edges can further carry attributes to express specific features of diseases and disease associations. Examples of such attributes include node size for disease prevalence and edge weight for association strength [Divo et al., 2015, Warner et al., 2015]. Furthermore, edges can be directed to represent the dynamic (e.g., causal or sequential) interactions among diseases [Jensen et al., 2014, Wang et al., 2020].

- *Support for disease progression analysis.* By discretizing the entire time frame of the index disease into different time windows, modeling comorbidity within each window as a disease network (hereafter referred to as *temporal disease network*, TDN), then comparing the dynamics through the TDN sequence across different windows, researchers are able to show and analyze the progression of the index condition and comorbidities. This approach has been applied to chronic conditions, such as cancer and mental disorders, which often come with a long period and multiple comorbidities [Chen et al., 2009, Chmiel et al., 2014].

- *Capability to incorporate additional biomarkers.* In addition to diseases, other biomarkers, such as genes and symptoms, can also be modeled as nodes and incorporated into the disease network by establishing edges that are representative of associations between the diseases and the biomarkers. For example, "diseasome" networks incorporate genes and/or proteins as nodes, and link them with diseases [Barabási, 2007, Nam et al., 2019], and psychiatric symptom networks include symptoms, drugs, and even adverse effects of drugs in addition to diseases [Cramer et al., 2010, Davazdahemami and Delen, 2018].

2.2.2 Temporal Disease Networks for Comorbidity Progression

Networks are commonly employed in healthcare visualization analytics. In a basic undirected, unweighted network modeling comorbidities, nodes represent comorbidities, while edges manifest the coexistence relationships among diseases in a certain patient cohort. The coexistence relationship is usually evaluated using a statistical measure, such as relative risk [Jeong et al., 2017], Pearson's correlation [Hidalgo et al., 2009], and Salton Cosine Index (SCI) [Chen et al., 2015], among others. Then, a threshold is used to eliminate trivial coexistences and retain the significant ones as edges.

Comorbidity networks are often large, dense, and complicated. To facilitate the analysis and visualization of complex comorbidity networks, graph clustering methods have been used to detect comorbidity patterns [Guo et al., 2019, Shu et al., 2019] and reduce network complexity [Schäfer et al., 2014]. A common network clustering model is the clique, i.e., a complete graph, in which all nodes are pairwise interconnected [Sokolova et al., 2017, Peleg et al., 2009]. For instance, in Figure 2.1, the TDN at Window 1 is a clique of three nodes.

Given a sequence of TDNs across different time windows, progression analysis often

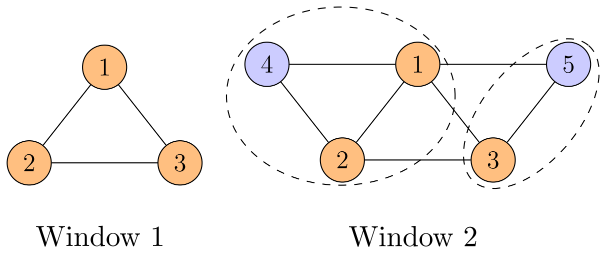


Figure 2.1: Two TDNs in Different Windows

Note: The one at Window 1 is a clique with three nodes. In Window 2, a clique enumeration algorithm may detect two cliques as circled, which is a split of the clique at the earlier window, causing analysts to lose track of the disease cluster implied by the clique.

involves comparing how much the TDNs are dissimilar from each other. There have been abundant approaches proposed in the literature to measure network dissimilarity [Tantardini et al., 2019, Wills and Meyer, 2020]. A majority of these methods summarize the structural features of a network into a vector of statistics, then define the dissimilarity between a pair of networks as the distance (e.g., Euclidean or Manhattan distance) between the two vectors associated with the networks. In addition to basic structures in network theory, e.g., node degree and network diameter, the literature also used many advanced structural features, including cluster coefficient [Berlingerio et al., 2013], graphlet [Pržulj et al., 2004], and graph kernel [Vishwanathan et al., 2010, Ghosh et al., 2018], to name a few.

2.2.3 Intellectual Gaps

Through a thorough literature review, we found that in the area of TDN modeling and analysis, there has been limited work to

- *Outline progression phases.* Most TDN-related studies [Chen et al., 2009, Martel et al., 2016, McElroy et al., 2018] predefined a granularity parameter m, then discretized the entire time frame of the study cohort into m windows of even length or even sample size, without providing algorithms that can detect at which window(s) notable changes of TDNs had occurred. Another issue brought by the simple m-window discretization method is that when the granularity is high, many windows come with very similar

TDNs, which increases the redundancy of visualization, especially at late stages of the time frame, when the number of comorbidities grows to a stable level.

- *Streamline the visualization.* Network clustering methods, such as the clique model, can be used to streamline the visualization of a single network as discussed earlier, but the extension to TDNs across multiple time windows is not straightforward. A confusion is that a clique in one window may be divided in another window, as shown in Figure 2.1. For complex TDNs with large sizes and many windows, the confusion will be much deteriorated, leading to the loss of track of certain disease clusters.

2.3 Clinical Information Extraction Using NLP

Natural language processing (NLP) has become increasingly popular in healthcare analytics over the past decades [Assale et al., 2019]. It has been extensively applied to extract useful clinical information from different types of clinical free text, such as radiology reports [Pons et al., 2016], pathology reports [Burger et al., 2016], medical literature [Huang et al., 2011], and medical social media [Denecke, 2014, Tutubalina et al., 2018]. Functionalities involved in clinical information extraction (CIE) mainly include text classification (e.g., phenotyping, mortality prediction, and severity prediction), named entity recognition (e.g., clinical concepts, de-identification, and negation), relation extraction, and others (e.g., information retrieval, disambiguation, and segmentation) [Wu et al., 2020]. Rule-based and supervised machine learning (ML) approaches are often employed in NLP for CIE, and a multitude of associated computer programs have been created.

2.3.1 Rule-based Approaches

As suggested by the name, rule-based approaches leverage a series of predefined semantic rules to extract the information of specific interest [Wang et al., 2018b]. The rules often consist of a specified lexicon and the search logic for the lexicon. Popular lexicons in healthcare research and applications include Unified Medical Language System (UMLS) for medical terminologies [Bodenreider, 2004], Chronic Conditions Data Warehouse regarding chronic condition categories [HealthAPT, 2022], Phenome-Wide Association Studies (PheWAS) about disease-gene relations [Denny et al., 2010], and DrugBank with respect to drug-gene relations [Wishart et al., 2008], among others. The search logic can be implemented with algorithmic flow control statements (e.g., “if...else...”) in combination with regular expression (a sequence of characters defining the match pattern to locate specific strings embedded in a text [Thompson, 1968]).

Physicians' domain knowledge and experience are generally leveraged to guide the development of rules. By integrating physicians' insights and other medical knowledge bases, rule-based algorithms are generally easy to interpret and can be accurate in many applications. For example, a review study shows that rule-based NLP algorithms and machine learning algorithms have similar performances in case detection [Ford et al., 2016].

However, rule-based algorithms require physicians and engineers to collaborate to manually craft rules, which is usually demanding and time-consuming [Chiticariu et al., 2013]. Another disadvantage is that researchers cannot enumerate all possible rules by heuristic observation inherently, particularly for complex or causal relations [Gevarter, 1984]. In addition, rule-based systems often perform unsatisfactorily if corpora or tasks change. In other words, they have limited portability and generalizability [Wang et al., 2018b].

2.3.2 Supervised Machine Learning

A typical ML procedure for NLP/CIE starts with an annotated corpus. Then, a set of word and context features of the corpus are extracted and used to train machine learning algorithms that can classify and recognize the text information of interest [Nadkarni et al., 2011]. The application of ML models, such as support vector machines [Wright et al., 2013], conditional random fields [Jiang et al., 2011], maximum entropy [Osborne et al., 2016], random forests [Brown and Kachura, 2019], and neural networks [Goldberg, 2016] in CIE has been extensively reported in the literature.

Feature engineering, which extracts a set of informative features and/or aggregates distinct features into new features in order to obtain a representation to enable ML classification, plays an essential role in CIE [Garla and Brandt, 2012, Xu et al., 2012]. Prevalent features in CIE include part-of-speech tags that are categories (e.g. noun, verb) to which words are assigned in accordance with their syntactic functions [Collier and Takeuchi, 2004], bag-of-words that represent a text as a multiset of words in the text disregarding word order [Kolari et al., 2006], n-gram that is a contiguous sequence of n items from a given text [Sidorov et al., 2014], term frequency-inverse document frequency (TF-IDF) that is a value for each word in a document through an inverse proportion of the frequency of the word in a particular document to the percentage of documents containing the word [Ramos et al., 2003], and word embeddings that are trained vector representations of words [Wang et al., 2018a].

It is worth noting that deep learning (DL), an emerging paradigm of artificial neural networks, has recently gained increasing attention in CIE [Wu et al., 2020]. DL shows great promise in many CIE applications thanks to its sophisticated network architectures. For instance, RNN and transformer networks have mechanisms designed to accommodate the sequential nature of texts that allow them to capture and exploit temporal information contained in these texts [Jurafsky and Martin, 2022]. Interested readers may refer to [Jurafsky and Martin, 2022] for a thorough discussion about DL models in NLP.

Compared to rule-based algorithms, ML/DL models are more efficient since they require fewer manual efforts and are automatically trainable [Chiticariu et al., 2013]. They also have a better capability of learning from high-dimensional representations or features of texts [Li, 2018] as well as capturing distant contextual information [Jurafsky and Martin, 2022]. However, in order to create accurate models, the model training process usually requires a substantial amount of labeled data samples, whose annotation can be highly time-consuming [Chiticariu et al., 2013]. In addition, feature engineering for non-DL ML models is complex and demanding [Chiticariu et al., 2013] while word-embeddings that DL models often rely on contain and magnify biases [Papakyriakopoulos et al., 2020].

2.3.3 Integrated NLP Software

By integrating rule-based and ML/DL algorithms, a lot of CIE software has been created in academia and industry. A considerable proportion of the software was designed to handle certain specific tasks in CIE. Examples of such specialized CIE software include NegEx [Chapman et al., 2001] for identifying negated findings and diseases, MedEx [Xu et al., 2010] and MedXN [Sohn et al., 2014] for detecting medications, and MedTime [Lin et al., 2013] that was designed to extract and normalize temporal information. On the other hand, generic CIE software, such as MedLEE [Friedman et al., 2004], MetaMap [Aronson and Lang, 2010], cTAKES [Savova et al., 2010], and CLAMP [Soysal et al., 2018] have emerged as popular integrated clinical NLP tools. These integrated tools have multiple built-in specialized algorithms and are able to recognize a variety of medical information including note headings, diseases, diagnoses, treatments, tests, and medications, among many others.

2.3.4 Intellectual Gaps

Through a thorough literature review, we found that in the area of analyzing clinical notes for E/M medical billing, there has been limited work to

- *Outline the framework or technical solutions for patient information extraction to facilitate E/M billing.* No academic research has been found to extract essential patient information from clinical notes for E/M medical billing.
- *Annotate a dataset for patient history information extraction.* No public datasets are available for training models to extract patient history information.

- *Compare model performances for the purpose of patient history information extraction.* There are a few studies comparing model performances for general named entity recognition but no study discussing performances of models to extract patient history information. In addition, there is no evaluation metric to quantify named entity recognition performance for notes with text span overlapping entities.

CHAPTER III

PREDICTING DIABETIC RETINOPATHY USING LONGITUDINAL ELECTRONIC HEALTH RECORDS DATA[1]

3.1 Introduction

The objective of this research is to build state-of-the-art deep learning temporal models to analyze longitudinal EHR data for diabetic retinopathy (DR) prediction. These models can be implemented as an accessible and low-cost alternative screening method to identify diabetic patients at high risk of DR. To make sure the models are easy for deployment, only a small set of patient demographics, comorbidities, and a few routine blood test variables are selected as input variables.

In Section 2.1, we review related literature in the area of modeling longitudinal EHR data for disease prediction and DR prediction as well as the intellectual gaps that we are addressing in this research. The remainder of this chapter is organized as follows. In Section 3.2, we introduce the dataset and data pre-processing. In Section 3.3, we discuss the models we intend to construct in detail. Model performance comparisons are in Section 3.4. Finally, Section 3.5 includes the discussion and conclusion of this study.

[1] This is a joint work with Zekai Wang, Dr. Bing Yao, and Dr. Tieming Liu. A manuscript based on this study has been accepted and will be published at IEEE CASE 2022 conference. Chen, S., Wang, Z., Yao, B. and Liu, T.(2022). Prediction of Diabetic Retinopathy Using Temporal Electronic Health Records. IEEE International Conference on Automation Science and Engineering (CASE), Mexico City, August 20-24, 2022.

3.2 Dataset and Data Pre-processing

The dataset used in this study is retrieved from the 2018 Cerner Health Facts® database, which is one of the largest real-world, de-identified, and HIPAA-compliant EHR databases in America. The database includes clinical records dated back to the late 1990s covering more than 63 million patients across the entire country. Clinical information such as patient demographics, admissions and discharges, lab tests, diagnoses, medications, procedures, and medical events are stored in the database with detailed timestamps, making Health Facts® an invaluable resource for predictive medicine.

Table 3.1: Diagnosis Codes for Diabetes and Its Complications

Code Type	**Code**	**Code Description**
ICD-9-CM	250.x	Diabetes mellitus (DM)
	362.0x	DR
	250.4x	Nephropathy
	250.6x	Neuropathy
	278.0x	Overweight and obesity
	414.0x	Coronary atherosclerosis
ICD-10-CM	E10.x	Type 1 DM
	E11.x	Type 2 DM
	E10.31x–E10.35x	Type 1 DM with DR
	E11.31x–E11.35x	Type 2 DM with DR
	E10.21	Type 1 DM with diabetic nephropathy
	E11.21	Type 2 DM with diabetic nephropathy
	E10.40	Type 1 DM with diabetic neuropathy
	E11.40	Type 2 DM with diabetic neuropathy
	E66.x	Overweight and obesity
	I25.x	Chronic ischemic heart disease
CCS Category	Hypertension%	Hypertension*

Note: "Hypertension complicating pregnancy; childbirth and the puerperium" and "Hypertension with complications and secondary hypertension" will be matched by the query "Hypertension%".

This study aims to develop state-of-the-art temporal predictive models to identify DR patients from diabetic patients. To retrieve the study cohort, diagnosis codes including International Classification of Diseases Ninth Revision-Clinical Modification (ICD-9-CM), ICD-10-CM, and the aggregated Clinical Classifications Software (CCS) categories were used (Table 3.1). Patients with one or more diabetes mellitus (DM) codes were defined as diabetic patients. Among these diabetic patients, those who have one or more DR codes were considered to be DR patients, and the remaining patients were treated as non-DR diabetic patients. In total, the DR cohort contains 69,354 patients whose admission dates range from December 1999 to June 2017. The non-DR diabetic cohort contains 2,363,051 patients whose admission dates range from October 1998 to September 2017.

After surveying the literature on DR prediction [Yau et al., 2012, Wang et al., 2021], we included 21 routine lab test variables (Table 3.2), 5 comorbidity variables (neuropathy, nephropathy, hypertension, obesity, and cardiovascular disease), 3 demographic variables (age, gender, and race), and the duration of diabetes in years.

Table 3.2: Lab Test Variables Used in the Dataset

Variable Name	Abbr.	Normal Range (Unit)
Alanine Aminotransferase/SGPT	ALT	7-55 U/L
Anion Gap	AG	7-15 mmol/L
Aspartate Aminotransferase/SGOT	AST	8-60 U/L
Bilirubin Total Serum or Plasma Mass/Volume	TSB	$\leq$ 1.2 mg/dL
Blood Urea Nitrogen	BUN	6-24 mg/dL
Hematocrit	Hct	35.5-48.6%
Hemoglobin	Hb	11.6-16.6 g/dL
Hemoglobin A1C (Glycosylated Hemoglobin)	HbA1c	4.0-6.4%
Mean Corpuscular Hemoglobin	MCH	27-31 pg/cell
Mean Corpuscular Hemoglobin Concentration	MCHC	32-36 g/dL
Mean Corpuscular Volume	MCV	78.2-97.9 fL
Red Blood Cell Count	RBC	3.92-5.65 x 10^{12}/L
Serum Albumin	ALB	3.5-5.0 g/dL
Serum Calcium	Ca	8.6-11.0 mg/dL
Serum Chloride	Cl	98-112 mmol/L
Serum Triglyceride	Tgl	$\leq$ 150 mg/dL
Serum Potassium	K	3.6-5.2 mmol/L
Serum Quantitative Creatinine	Cr	0.59-1.35 mg/dL
Serum Sodium	Na	135-145 mmol/L
Serum/Plasma Quantitative Glucose	Glu	70-140 mg/dL
White Blood Cell Count	WBC	3.4-9.6 x 10^9/L

Note: Normal range values (except for MCH and MCHC) are taken from Rochester 2022 Interpretive Handbook by Mayo Clinic Laboratories. Values are mainly for adults, and some are relaxed for gender and age groups if differences exist. Normal range values for MCH and MCHC are from https://www.ucsfhealth.org/medical-tests/rbc-indices.

For a patient in the DR cohort, we search the first encounter when the patient was diagnosed with diabetes as encounter I, the first encounter when the patient was diagnosed of DR as encounter J, and then extract all the longitudinal lab test results ranging from I to J. For non-DR patients, we also set I as the encounter when the patient was first diagnosed with diabetes and extract all longitudinal lab tests from I to the patient's latest encounter J in the EHR database.

Lab test values are numerical and are aggregated in means at the encounter level. Patient race and gender are retrieved from the last encounter and are transformed into numerical with one-hot encoding. Patient age in years is collected at each encounter. Duration of diabetes in years is calculated as the temporal distance of admission time from current encounter to encounter I. The five comorbidity variables are binary. Neuropathy, nephropathy, and cardiovascular disease are generally considered irreversible, so if there exists one diagnosis code for these variables in a patient's retrieved records, the patient will be labeled positive for these variables, otherwise, the patient will be labeled as negative. Obesity and hypertension are considered reversible, so we only check the diagnosis codes in the last encounter. If there is such a diagnosis code for these two variables in the last encounter, we will label the patient positive. Otherwise, we will label the patient negative for these two variables.

Figure 3.1 shows that missing values are prevalent for lab test variables. Inspired by a previous study [Wang and Yao, 2022], we first use forward imputation, then use backward imputation, and impute the remaining missing values with 0.

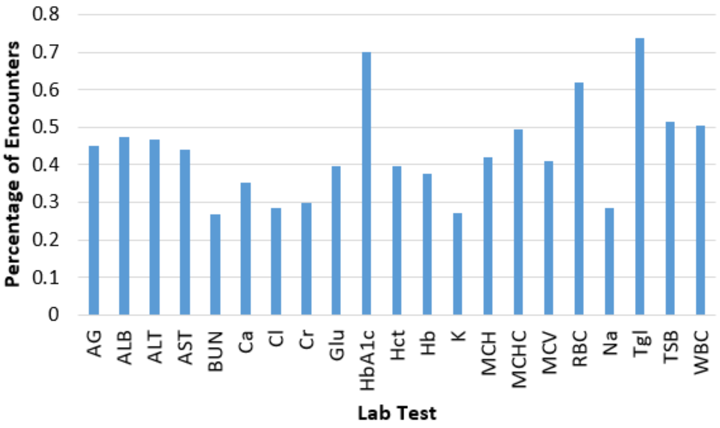


Figure 3.1: Percentages of Encounters with Missing Values

Table 3.3: Encounter Number Statistics of the Original Dataset

Encounter Number Statistic	**DR Cohort**	**Non-DR Cohort**
Min of Encounter Numbers	1	1
Max of Encounter Numbers	236	504
Mean of Encounter Numbers	5.89	5.68
Median of Encounter Numbers	3	3
% of Patients with $\leq$ 25 Encounters	97%	97%

Table 3.3 shows the statistics of encounter numbers in the two cohorts. Most patients have limited numbers of encounters, which is not ideal for building temporal models. Since the objective of this study is to build temporal models for analyzing longitudinal EHR data, we sample a subset as the final dataset containing patients with at least 10 encounters. In the final dataset, there are 414,199 patients in total, and 12,590 of them are DR positive. DR positive rate is 3.04%, close to the 2.85% DR rate in the original whole dataset. To ensure an equal number of encounters for all the patients, we set the maximum time steps as 25 and apply zero-padding techniques. For patients with less than 25 encounters, padding encounters will be added at the beginning of the encounter sequences. For patients with more than 25 encounters, only the latest 25 encounters are kept.

3.3 Methodology

As we discussed in Section 2.1.1, artificial-neural-networks-based deep learning models have been increasingly employed in temporal analysis of longitudinal healthcare data in recent years due to their flexibility in modeling sequential data. RNN (particularly its sub-type of LSTM) and TCN are among the most popular temporal models with many successful applications including disease prediction. Inspired by a recent study on sepsis prediction [Wang and Yao, 2022], we intend to build LSTM and TCN models with a multi-branching mechanism and compare the model performances.

A recent survey finds that there is a lack of clinical domain knowledge in the current deep learning studies modeling healthcare data [Xie et al., 2022]. Incorporating invaluable clinical domain knowledge may improve model performances. In Section 2.1.2, we discussed some of the most important clinical findings about DR in literature including potential DR risk factors. We will incorporate such knowledge of DR risk factors in our model building.

In summary, we intend to build LSTM and TCN models with a multi-branching mechanism for DR prediction. We will compare them with non-temporal random forests models, which have the best performances in literature.

3.3.1 Artificial Neural Networks

A typical artificial neural networks architecture is composed of one input layer, one or more hidden layers, and one output layer. Each layer may contain different numbers of nodes, and nodes are connected from layer to layer to allow information to flow from the input to the output. The interaction between two adjacent hidden layers i and $i-1$ is characterized as

$$Y^i = \sigma(B^i + W^i Y^{i-1})$$

where Y^{i-1} and Y^i denote the outputs of layers $i-1$ and i respectively, W^i and B^i are the weight matrix and bias vector for layer i, and $\sigma(\cdot)$ denotes the nonlinear activation function (e.g., sigmoid, tanh, or ReLu).

By designing sophisticated hidden layer structures, advanced neural network models can be developed to capture hidden information and patterns from different types of data. For example, convolutional neural networks (CNNs) have been developed to investigate spatial correlations for pattern recognition in imaging data [Rawat and Wang, 2017]. Furthermore, RNN and TCN have been proposed to capture the temporal correlations in longitudinal data. In this work, we will engage different temporal network architectures to analyze the longitudinal EHR data for DR prediction.

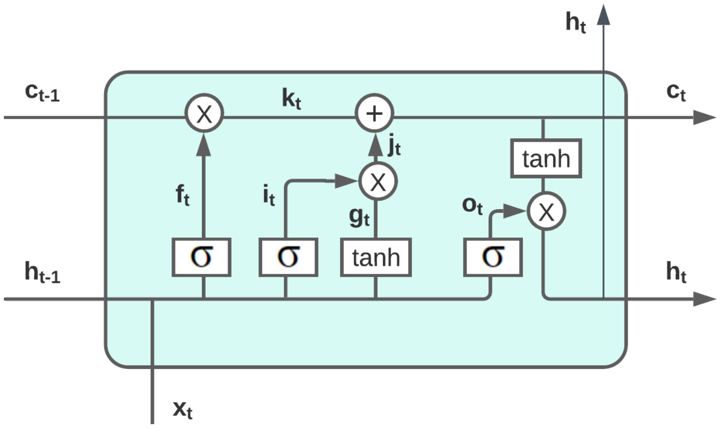


Figure 3.2: Illustration of an LSTM Block

3.3.2 LSTM

LSTM networks have been designed to mitigate the common problem of gradient vanishing and explosion in traditional RNNs by incorporating a gating mechanism in the network structure. LSTM has a wide application to capture the temporal dynamics of sequential data in a variety of areas such as machine translation and speech recognition. The information flow across the LSTM block is achieved by three layers, i.e., the input layer x, the hidden layer h, and a context layer c. As illustrated in Figure 3.2, a classic LSTM block takes the current input vector x_t, and the hidden state and context state (h_{t-1}, c_{t-1}) from the previous block and further controls the information flow through the three gates ("forget" gate f_t, "add" gate i_t, and "output" gate o_t).

The gate f_t is used to delete information from context no longer needed and performs the following operations

$$f_t = sigm(U_f h_{t-1} + W_f x_t)$$

$$k_t = c_{t-1} \otimes f_t$$

where $sigm$ is sigmoid operation, $\otimes$ is element-wise multiplication, and U_f and W_f are coefficient matrices.

The gate i_t is used to add information to the context state and its output is calculated by the following operations

$$g_t = tanh(U_g h_{t-1} + W_g x_t)$$

$$i_t = sigm(U_i h_{t-1} + W_i x_t)$$

$$j_t = g_t \otimes i_t$$

where $tanh$ is hyperbolic tangent operation. The context layer is further updated as

$$c_t = j_t + k_t$$

Finally, the gate o_t decides the information needed for the current hidden state and its operations are as follows

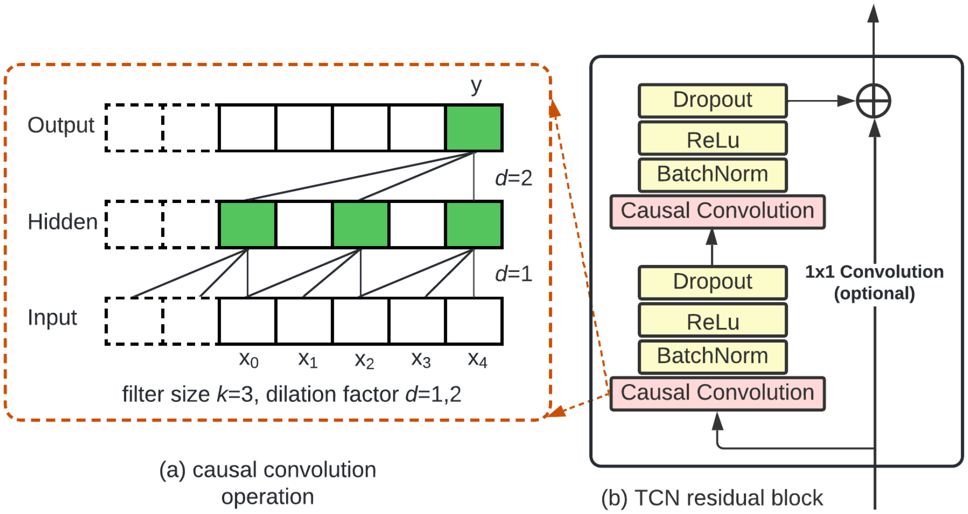
$$o_t = sigm(U_o h_{t-1} + W_o x_t)$$

$$h_t = o_t \otimes tanh(c_t)$$

LSTM is flexible in processing sequential data with irregular duration and further making predictions. LSTM blocks can be stacked to make the networks deeper and possibly learn more complex representations. In this study, we will build an architecture of two stacked LSTMs. The output dimensions of the two LSTM layers are 32 and 64 respectively. Kernel regularizer is l_2-norm of the network parameters and the loss function is binary cross-entropy.

3.3.3 TCN

TCN is an extension of conventional CNN. Figure 3.3 illustrates a typical TCN residual block. Instead of capturing spatial patterns as in traditional CNN, TCN-based models extract temporal dependency in longitudinal data through the mechanism of dilated causal convolutions. A simple dilated causal convolution operation with dilation factors $d = 1, 2$ and filter size $k = 3$ is shown in Figure 3.3(a). Mathematically, the dilated convolutional



(a) causal convolution operation

(b) TCN residual block

Figure 3.3: Dilated Causal Convolution and Residual Block of a TCN Block

operation is given as

$$C(t) = (x_v^p *_d f)(t) = \sum_{a=0}^{k-1} f(a) \cdot x_v^p(t - a \cdot d)$$

where $C(\cdot)$ denotes the convolution output, x_v^p denotes the observations of variable v over time for patient p, $*$ represents the convolution operation, d is the dilation factor, $f(a) : a \in \{0, 1, ...k - 1\}$ is a filter, and k is the kernel size. Note that the operation $(t - a \cdot d)$ ensures only historic information is included in the causal convolution calculation at time t.

The causal convolution guarantees all the historical sequential information in the data is captured in the network without leaks. We can tune dilation factors and the kernel size to modify the model's capabilities of identifying local or distant temporal dependency. Inside the residual block of Figure 3.3(b), there are two layers of causal convolutions, and each causal convolution layer is followed by a batch normalization, a ReLu activation function, and a dropout rate. A residual shortcut is also included in the block which directly passes the input data to the output operation of the block. Note that an additional 1 × 1 convolutional layer is needed for the shortcut connection if the input and output of the causal convolution operations have different dimensions. The output of a residual block $o(x_v^p)$ is derived as

$$o(x_v^p) = activation(\mathcal{F}(x_v^p) + x_v^p)$$

where $\mathcal{F}$ denotes a series of network transformations including the causal convolution, batch normalization, ReLu, and a dropout layer.

In this study, we will build a TCN architecture with 1 stack of a residual block, 64 filters, a kernel size of 3, ReLu activation, and a dropout rate of 0.1.

3.3.4 Multi-Branching Output Mechanism

Another technique used in this study is the multi-branching output mechanism to address imbalanced data issue.

Real-world medical data is often subject to imbalanced data issue. For a classification task, a dataset is called imbalanced if the proportions of the target classes (e.g., positive/negative for a certain disease, mild/moderate/severe symptoms, survived/deceased after a procedure) of the dataset are skewed. Imbalanced data is a common issue in EHRs, as the number of patients with the target disease/symptom is much smaller compared with the population. The class or classes with abundant examples may dominate the machine learning process. In other words, with imbalanced data, the algorithms pursuing classification accuracy tend to ignore the minority class (e.g., people with the disease). Such a classifier is unsatisfactory because the detection of minority examples is crucial. In our study, the proportion of patients with DR in the population is less than 3%. The LSTM and TCN models trained directly on such imbalanced data will yield unsatisfactory prediction performance.

Literature shows that a multi-branching output mechanism can improve model performances due to balanced training datasets and robust prediction [Wang and Yao, 2022]. In this study, we will add z branching outputs as shown in Figure 3.4 after the LSTM and TCN layers to build different architectures. Each branching output consists of a dense layer followed by a sigmoid activation function. During the training process, the training dataset is first transformed into a balanced set by oversampling the minority class (positive for DR) and then divided into equal-size batches. During the training phase, for each balanced training

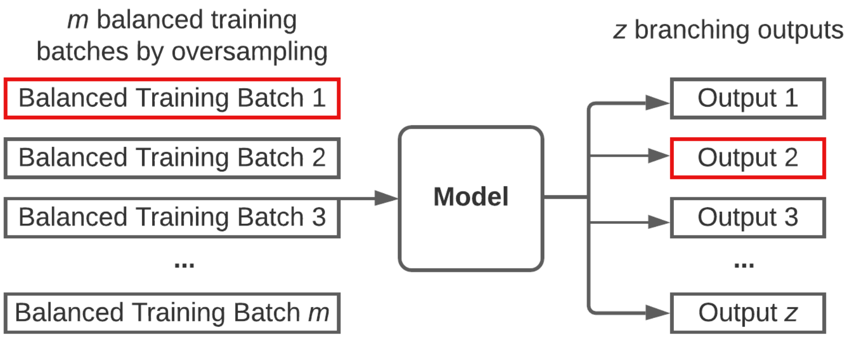

m balanced training batches by oversampling
Balanced Training Batch 1
Balanced Training Batch 2
Balanced Training Batch 3
...
Balanced Training Batch *m*
Model
z branching outputs
Output 1
Output 2
Output 3
...
Output *z*

Figure 3.4: Multi-branching Output Mechanism

batch, a random output is selected for the training. The core TCN or LSTM architectures will eventually be trained on the entire dataset. During the prediction phase, we will take the mean of the z outputs as the probability of the positive case.

Due to computational limits, we will build multi-branching models for LSTM and TCN with $z = 1$, 5, and 10, respectively. When $z = 1$, the models are conventional architectures without the multi-branching output mechanism.

In addition to LSTM and TCN architectures, we will also build random forest (RF) classifiers as non-temporal baseline models and compare the performances of all the models. We use two methods to transform longitudinal data into non-temporal data for RF models - only keeping the lab results of the last encounter ("LastEnct") or aggregating the lab results of all the encounters in statistical means ("MeanEnct").

In addition, to address the missing value issue of temporal EHR data, the framework first generates missing-value masks for the raw data, imputes any missing values in the raw data, and uses both data as input data of the model.

3.4 Results

3.4.1 Model Evaluation Metrics

Choosing the right evaluation metrics is important when we compare performances of classification models. We will discuss the major evaluation metrics in the literature and select the ones suitable for our study.

Accuracy is one of the most popular metrics for classification and is defined as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

where TP (short for “true positive”) is the number of positive cases that are correctly predicted as positive by a model, FP (short for “false positive”) is the number of negative cases that are incorrectly predicted as positive, TN (short for “true negative”) is the number of negative cases that are correctly predicted as negative, and FN (short for “false negative”) is the number of positive cases that are incorrectly predicted as negative. A false positive case is also called a Type I error while a false negative case is called a Type II error.

Although accuracy is simple and has been widely used, it is not a good metric for highly imbalanced datasets or the costs of different types of errors vary significantly [Chawla, 2009]. In our study, the dataset is extremely imbalanced (DR positive rate is less than 3%). A simple guess of non-DR for all patients can achieve high accuracy of more than 97%, but it fails to identify any DR patient. In addition, a Type II error is significantly more costly than a Type I error. A Type II error (i.e., the model fails to identify a DR patient) may delay the intervention or treatment, and thus result in vision damage to the patient while a Type I error (i.e., the model incorrectly classifies a non-DR patient as DR) may only have the extra cost of an ophthalmic exam for confirmation. Based on these two points, accuracy is not an appropriate metric for our study and will not be used in model evaluation.

Specificity and *sensitivity* are one pair of common evaluation metrics for classification models. Specificity represents the ability of a model to correctly identify people without a disease while sensitivity reflects the ability of that model to correctly identify patients with the disease. They are calculated as follows.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}}$$

Note that specificity is also called true negative rate or (1 - false positive rate) while sensitivity is also called true positive rate.

A model is desirable if it can achieve high values of both specificity and sensitivity. Clearly, a simple guess of the majority class non-DR in our study will achieve a high specificity value of 1 but a low sensitivity of 0, an example of why this pair of metrics is better than accuracy. A receiver operating characteristic (ROC) curve is commonly used to plot the true positive rate (i.e., sensitivity) against the false positive rate (i.e., 1 - specificity), and the area under the curve, called AUC or AUROC, is often used as an evaluation metric in practice. AUROC ranges from 0 to 1. The higher the AUROC, the better model we have.

Precision and *recall* are another pair of common evaluation metrics. Precision is the fraction of true positive cases among all the "positive" cases that a model predicts while recall is the fraction of positive cases that are correctly predicted by that model. They are calculated as follows.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$$

Obviously, recall is the same as sensitivity.

We expect a model to achieve high values of both precision and recall. A simple guess of the majority class non-DR for all people will achieve a precision of 1 or N/A (depending on how we define the edge scenarios) but a low recall of 0. A precision-recall curve is used to plot the precision against the recall and the area under the curve, called AUPRC, is often used as an integrated evaluation metric. AUPRC ranges from 0 to 1. The higher the AUPRC, the better model we have.

Different from specificity (based on the negative class) and sensitivity, both precision and recall are based on the positive class, making AUPRC more sensitive to improvements for the positive class which is more important than the negative class in our study. In fact, studies in the literature have claimed that AUPRC is more suitable than AUROC for imbalanced datasets [Bradley et al., 2006, Soleymani et al., 2020]. During our model comparison, we will include both AUROC and AUPRC, but we will consider AUPRC to be a more effective metric based on the literature.

As we discussed that a Type II error may be detrimental (may cause vision loss or blindness to patients), we would like to minimize the Type II error rate, or maximize sensitivity or recall. Therefore, we will check the recall values (denoted as Recall hereafter) as well.

3.4.2 Model Performances

We summarize the prediction performances of our models in Table 3.4 with the three metrics we discussed, AUROC, AUPRC, and Recall.

The RF-LastEnct model outperforms the others in terms of AUROC (0.927) although the performances of TCN and RF-MeanEnct models are very close (0.922). However, TCN-MB-10 model outperforms its peers in terms of AUPRC (0.765). RF models achieve slightly lower AUPRC values (0.757 and 0.751).

Since Recall depends on the threshold we choose for our probabilistic prediction, we arbitrarily choose 0.5 as the threshold and compare the corresponding performances. From the last column in Table 3.4, we can see that temporal models outperform baseline RF models in terms of Recall. We tried other thresholds and the temporal models generally have higher Recall values.

Table 3.4: Model Performances

Model	AUROC	AUPRC	Recall*
RF-LastEnct	**0.927**	0.757	0.650
RF-MeanEnct	0.922	0.751	0.650
LSTM	0.919	0.734	0.777
LSTM-MB-5	0.919	0.740	0.776
LSTM-MB-10	0.917	0.736	0.792
TCN	0.922	0.724	0.708
TCN-MB-5	0.916	0.716	0.722
TCN-MB-10	0.912	**0.765**	**0.886**

*: Recall is short for the recall score of the DR positive class with a threshold of 0.5. RF-LastEnct represents the RF model built on a non-temporal dataset derived by keeping lab tests of the last encounter while RF-MeanEnct is built on a non-temporal dataset derived by aggregating lab tests across all encounters.

3.5 Conclusion

In this study, we implement state-of-the-art temporal models including LSTM and TCN with a multi-branching output mechanism to analyze longitudinal EHR data for DR prediction. Experimental results on a large-scale dataset show that temporal models achieve similar AUROC values compared with baseline RF models but outperform in both AUPRC and Recall scores, exhibiting the benefits of analyzing longitudinal EHR data for DR prediction.

Contributions of this study are twofold. First, to the best of our knowledge, this is the first study using temporal models to analyze longitudinal EHR data for DR prediction. Second, we only use a small set of patient demographics, comorbidities, and routine lab test variables as model input, providing a more accessible approach for DR screening. This can be deployed in medically underserved communities to reduce eye care disparities.

We identified several limitations of the present investigation, which need to be addressed in our future work. First, the numbers of temporal and baseline models examined are limited. Examining more temporal models (e.g., the knowledge-based temporal abstraction approach and the transformers models) and more non-temporal baseline models may help us identify more differences in model performances. Second, more data representation techniques may be needed. Some variables (i.e., patient demographics and comorbidities) are not longitudinal in nature. In this study, they are duplicated to accommodate the sequence of lab tests, which may not be the best method. Differentiating the representation of different types of variables is needed to further improve the model performances. Third, a more detailed study on the association of model performance and the number of multi-branching outputs is desired.

CHAPTER IV

CLUSTERING TEMPORAL DISEASE NETWORKS TO VISUALIZE COMORBIDITY PROGRESSION[1]

4.1 Introduction

The objective of this research is to design a visual analytics (VA) system that can efficiently detect and visualize comorbidity progressions using temporal disease network (TDN) models. The VA system incorporates two novel TDN clustering technologies—*temporal clustering* and *disease clustering*. The temporal clustering identifies notable changes during the comorbidity progression and aggregates windows to phases based on the time of the changes. On the other hand, the disease clustering captures higher-level structures of TDNs by clustering highly coexisting conditions and simplifies the TDN visualization based on the identified structures. The developed VA system can be integrated into clinical decision support systems to provide evidence-based, visual insights regarding the timeline and patterns of comorbidity progression to support the decision-making in healthcare settings.

In Section 2.2, we provide a literature review of related work in the area of TDNs and show the intellectual gaps that we are addressing in this research. The remainder of this chapter is organized as follows. In Section 4.2, the system design and proposed clustering technologies are presented in detail, followed by two case studies on Clostridioides Difficile

[1] This is a joint work with Dr. Yajun Lu, Dr. Zhuqi Miao, Dr. Dursun Delen, and Dr. Andrew Gin. A paper based on this study has been published on Decision Support Systems. Lu, Y., Chen, S., Miao, Z., Delen, D. and Gin, A.(2021). Clustering temporal disease networks to assist clinical decision support systems in visual analytics of comorbidity progression. Decision Support Systems, 148, article number 113583.

(C. Diff) and stroke in Section 4.3. Finally, Sections 4.4 and 4.5 include the discussion and conclusion of this study, respectively.

4.2 Methodology

The proposed VA system consists of four modules as shown in Figure 4.1. Module 1 receives data from clinical data warehouses and prepares the data for subsequent analysis and visualization. Module 2 then builds TDNs with sufficient granularity using the preprocessed data, while Module 3 identifies highly similar TDNs and clusters corresponding windows into phases, followed by Module 4 which visualizes TDNs over the phases. The role of clinical domain experts is to guide the process of modules by determining the initialization parameters and examine the output for validity, while the system eventually provides visual insights regarding comorbidity progression back to the clinical experts to support evidence-based decision making. Since the technological contribution of our work mainly revolves around Modules 2, 3, and 4, the rest of this section will focus on elaborating the methodology we employed to design these modules.

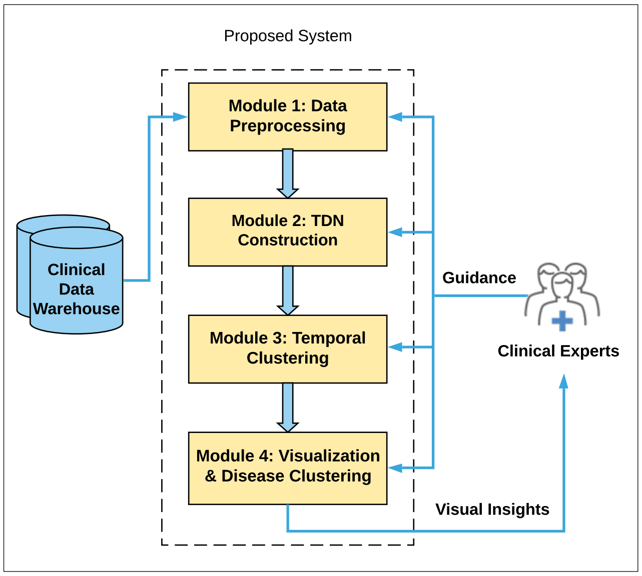

Proposed System
Module 1: Data Preprocessing
Module 2: TDN Construction
Module 3: Temporal Clustering
Module 4: Visualization & Disease Clustering
Clinical Data Warehouse
Guidance
Clinical Experts
Visual Insights

Figure 4.1: Proposed TDN-based VA System for Comorbidity Progressions

4.2.1 TDN Construction

First, we provide basic definitions and notations of a network used in this study. A most fundamental network (i.e., graph) model, denoted by G, is a mathematical structure composed of a set of nodes $V(G)$ that model the objects of interest and a set of undirected edges $E(G)$ representing the pairwise relationships among the objects [West, 1996]. The number of nodes and the number of edges included in a network are called the *order* and the *size* of the network and are denoted by $|V(G)|$ and $|E(G)|$ respectively. Given a subset of nodes $S \subseteq V(G)$, we herein denote by $G[S]$ the subgraph induced by S, i.e., a subgraph obtained by dropping nodes in $V(G) \backslash S$ and their incident edges from G. For a node $v \in V(G)$, the neighbors of v, $N_G(v)$ refers to the set of nodes adjacent to v and its cardinality is called the *degree* of v, denoted by $\deg_G(v)$ herein. In this research, node v and its neighbors $N_G(v)$ together are referred to as the *closed neighborhood* of v, denoted by $N_G[v]$, and its induced subgraph is called the ego network of v, denoted by $ego_G(v)$. In other words, $N_G[v] \coloneqq N_G(v) \cup \{v\}$ and $ego_G(v) \coloneqq G[N_G[v]]$.

To quantify the coexistence relationship among comorbid diseases, we make use of SCI [Chen et al., 2015, Kalgotra et al., 2017], which can be expressed as

$$SCI_{ij} = \frac{n_{ij}}{\sqrt{n_i n_j}} \tag{4.2.1}$$

where n_{ij} represents the number of hospital encounters with the onset of both diseases i and j, while n_i (or n_j) corresponds to the number of encounters with the onset of disease i (or j). When $SCI_{ij} = \gamma\%$, at least one of n_{ij}/n_i and n_{ij}/n_j is no less than $\gamma\%$. It implies that encounters with the onset of both diseases are at least γ percent of all encounters of one disease. *SCI* has been used as an alternative to Pearson's correlation coefficient (PCC) for disease network modeling because it avoids two potential issues of PCC: (i) sample size can have an overly high impact on the PCC strength [Kalgotra et al., 2017], and (ii) PCC may underestimate the coexistence of a pair of diseases, of which one is rare while the other is prevalent [Fotouhi et al., 2018].

Given an SCI threshold θ determined under the clinical experts' guidance, we can then establish an edge between each pair of diseases (nodes) i and j such that $SCI_{ij} \geq \theta$. In addition to θ, the system also needs advice from clinical experts to specify a value for the granularity parameter m to discretize the entire time frame into m windows that are as granular as possible. Re-organization of the windows will be accomplished by the Temporal Clustering Module of the system.

4.2.2 Temporal Clustering

This module involves two techniques: (i) network dissimilarity measurement, and (ii) consecutive p-median clustering for time windows, as elaborated in the following.

Network Dissimilarity Measurement

In this research, we adapted and improved the *NetSimile* method proposed by [Berlingerio et al., 2013] to evaluate the network dissimilarity among different windows. The NetSimile method "quantifies" the structural features of a network G by calculating multiple statistical metrics (including median, mean, standard deviation, skewness, and kurtosis in this study) for a number of features associated with each node $v \in V(G)$. The specific features employed in this study include:

- The degree of v;
- Clustering coefficient of v, defined as $\frac{2}{\deg_G(v)(\deg_G(v)-1)}|E(G[N_G(v)])|$;
- The average degree of the neighbors of v;
- The average clustering coefficient of the neighbors of v;
- The size of the ego network of v;

- The number of edges connecting the nodes in $ego_G(v)$ and nodes not in $ego_G(v)$;
- The number of nodes that are not in $ego_G(v)$ but are neighbors of nodes in $ego_G(v)$.

The process results in a 35-entry vector of statistics that evaluates the structure of a network. The vector is herein referred to as the *signature vector* and denoted by Z_G for a network G.

In the classical NetSimile method, the dissimilarity between a pair of networks, G_i and G_j, was assessed using the Canberra distance of the corresponding signature vectors, i.e.,

$$\delta(G_i, G_j) = \frac{1}{35} \sum_{k=1}^{35} \frac{|Z_{G_i}[k] - Z_{G_j}[k]|}{|Z_{G_i}[k]| + |Z_{G_j}[k]|} \tag{4.2.2}$$

where $Z_{G_i}[k]$ (or $Z_{G_j}[k]$) indicates the kth entry of the vector Z_{G_i} (or Z_{G_j}). The similarity then can be expressed as $1 - \delta(G_i, G_j)$. Nevertheless, the classical NetSimile method does not consider the disparity of node sets, and thus can underestimate the dissimilarity when there are uncommon nodes between two networks. Considering the two TDNs, G_1 and G_2 shown in Figure 4.2, $\delta(G_1, G_2) = 0$ indicating the "identical" edge structure between the two TDNs. However, the structure is based on different node sets (new diseases 4 and 5 are developed from G_1 to G_2, whereas diseases 1 and 2 become absent), so the two TDNs are actually not the same.

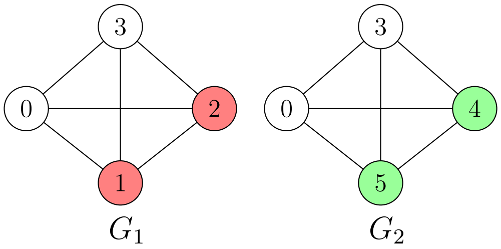


Figure 4.2: Two TDNs that Are Cliques with Different Node Sets

This issue motivates us to introduce an overlapping factor to enhance the NetSimile method to handle the dissimilarity caused by the difference between node sets. The overlapping factor $\omega(G_i, G_j)$ is defined as follows,

$$\omega(G_i, G_j) = \frac{|E(G_i[D])| + |E(G_j[D])|}{|E(G_i)| + |E(G_j)|} \tag{4.2.3}$$

where $D = V(G_i) \cap V(G_j)$. Clearly, $0 \leq \omega(G_i, G_j) \leq 1$. By incorporating $\omega(G_i, G_j)$, the modified dissimilarity $d(G_i, G_j)$ is expressed as

$$d(G_i, G_j) = 1 - \omega(G_i, G_j) \times (1 - \delta(G_i[D], G_j[D])) \tag{4.2.4}$$

The rationale behind the modified formula is straightforward: $1 - \delta(G_i[D], G_j[D])$ evaluates the similarity between the node-overlapping subgraphs of G_i and G_j. Because the rest parts are completely different, we scale down $1 - \delta(G_i[D], G_j[D])$ with the overlapping factor $\omega(G_i, G_j)$ to evaluate the overall similarity between the two entire networks. Re-considering the two networks in Figure 4.2, the dissimilarity between node-overlapping subgraphs $\delta(G_i[D], G_j[D]) = 0$ and $\omega(G_1, G_2) = 1/6$, therefore $d(G_1, G_2) = 5/6$ and the overall similarity between the two networks is 1/6, which is a better evaluation compared with that returned by the classical NetSimile method.

Consecutive p-Median Clustering

As we pointed out in Section 2.2.3, some consecutive windows can come with very similar TDNs, thus providing limited new information about comorbidity progression, and leading to visualization redundancy. In order to address the issue, we propose and solve a *consecutive p-median problem* (CPMP) defined as follows.

Problem: Consecutive p-median problem.
Input: A positive integer p, a collection of m objects $\mathcal{O} := \{O_1, O_2, \ldots, O_m\}$, and the distance between any two objects.
Output: From $\mathcal{O}$, find p objects with indices $\{j_1, j_2, \ldots, j_p\}$ as medians and assign the remaining $m - p$ objects to the medians such that

(i) The total summation of distances from each O_i to its assigned median is minimized,

(ii) When O_i is assigned to median O_{j_q}, if $i \geq j_q$, O_k for all k such that $j_q \leq k < i$ must be assigned to O_{j_q}, otherwise, O_k for all k such that $i < k \leq j_q$ must be assigned to O_{j_q}.

The problem is an extension of the classical p-median problem that has been often used for clustering [Klastorin, 1985, Köhn et al., 2010]. The extension is condition (ii) that imposes the assignment of consecutive objects to medians. For example, if we would like to solve the consecutive 2-median problem on the TDNs shown in Figure 4.3, we cannot assign the TDNs on Window 1 and Window 5 together even though they are identical. By applying CPMP to TDNs, we can group consecutive windows with highly similar TDNs into the same temporal cluster, which can be interpreted as a *phase* of comorbidity progression.

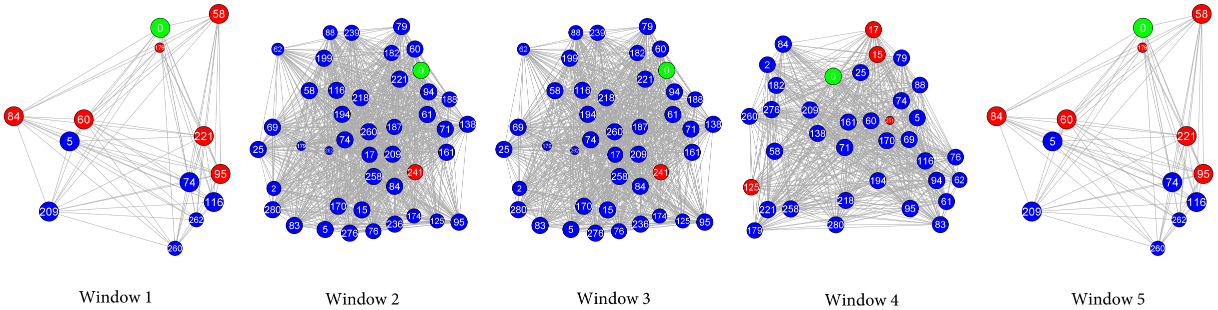


Figure 4.3: A Sequence of TDNs Across Five Time Windows

Note: TDNs on Window 1 and Window 5 are identical and the ones through Window 2 to Window 4 are highly similar.

In this study, we developed a (linear) integer programming (IP) formulation (4.2.5)–(4.2.12) to model and solve the CPMP on a sequence of TDNs, $\mathcal{G} = \{G_1, G_2, \ldots, G_m\}$. In the formulation, the binary variable $x_{ij} = 1$ if and only if TDN G_i is assigned to median G_j, for any $i, j \in \{1, 2, \ldots, m\}$ such that $i \neq j$, otherwise $x_{ij} = 0$. When $x_{jj} = 1$, it indicates that G_j is designated as a median for any $j \in \{1, 2, \cdots, m\}$. The objective function (4.2.5) aims to minimize the total dissimilarity between TDNs and the medians to which the TDNs are assigned across all windows. Constraint (4.2.6) ensures that at most p TDNs are selected as medians. In Constraint (4.2.7), we force each TDN G_i to be assigned to exactly one median. Constraint (4.2.8) guarantees that if G_i is assigned to G_j then G_j must be a median. Constraints (4.2.9) and (4.2.10) make sure that only consecutive TDNs can be grouped into the same cluster. In constraint (4.2.11), τ represents a threshold for *not* clustering. When the dissimilarity between two consecutive TDNs G_i and G_{i+1} is greater than or equal to τ, they will not be grouped into the same cluster. This constraint allows us to avoid clustering highly different TDNs.

$$\min \sum_{i=1}^{m} \sum_{j=1}^{m} d(G_i, G_j) x_{ij} \tag{4.2.5}$$

$$\text{subject to: } \sum_{j=1}^{m} x_{jj} \leq p \tag{4.2.6}$$

$$\sum_{j=1}^{m} x_{ij} = 1 \qquad \forall i \in \{1, 2, \cdots, m\} \tag{4.2.7}$$

$$x_{ij} \leq x_{jj} \qquad \forall i, j \in \{1, 2, \cdots, m\} \mid i \neq j \tag{4.2.8}$$

$$x_{ij} \leq x_{kj} \qquad \forall i \in \{1, 2, \cdots, m-2\}, j \in \{i+2, i+3, \cdots, m\}, k \in \{i+1, i+2, \cdots, j-1\} \tag{4.2.9}$$

$$x_{ij} \leq x_{kj} \qquad \forall i \in \{3, 4, \cdots, m\}, j \in \{1, 2, \cdots, i-2\}, k \in \{j+1, j+2, \cdots, i-1\} \tag{4.2.10}$$

$$x_{ij} + x_{i+1,j} \leq 1 \qquad \forall j \in \{1, 2, \cdots, m\}, i \in \{1, 2, \cdots, m-1\} \mid d(G_i, G_{i+1}) \geq \tau \tag{4.2.11}$$

$$x_{ij} \in \{0, 1\} \qquad \forall i, j \in \{1, 2, \cdots, m\} \tag{4.2.12}$$

Selection of the Value for p

The parameter p determines how many clusters the initial time windows should be grouped into; in other words, how many phases the entire time frame is supposed to be broken down into. Usually, we are interested in a relatively small p to simplify the TDN sequence to allow us to capture the primary changes of comorbidity over time. Meanwhile, we need to avoid using a value that is too small, because an overly small p can result in very broad phases that combine windows hardly similar. The clinical advice from domain experts is essential to choose a proper value of p. Whereas, data analytic methods can also be used to support the decision on this parameter.

The Silhouette Index (*SI*) has often been used in literature to determine the value of p for p-median models [Köhn et al., 2010, Rousseeuw, 1987]. In this study, we adapted *SI* to find a proper value of p for our proposed CPMP. Let $\mathcal{C}(p) = \{C_1, C_2, \ldots C_p\}$ be a clustering of TDNs $\mathcal{G} = \{G_1, G_2, \ldots, G_m\}$; given a network $G \in \mathcal{G}$, let C_k represent the cluster that contains G and $\mathcal{G}_A$ denote the network(s) in $\mathcal{G}$ that are adjacent to G. Then, our adapted SI for G is defined as

$$SI_{\mathcal{G}}(\mathcal{C}(p), G) = \begin{cases} 0 & \text{if } |C_k| \geq \sigma|\mathcal{G}| \\ 0 & \text{if } |C_k| = 1 \text{ and } \exists \hat{G} \in \mathcal{G}_A \mid d(G, \hat{G}) < \tau \\ 1 & \text{if } |C_k| = 1 \text{ and } d(G, \hat{G}) \geq \tau, \forall \hat{G} \in \mathcal{G}_A \\ \dfrac{\Delta_{\mathcal{C}\setminus C_k}(G) - \Delta_{C_k}(G)}{\max\{\Delta_{C_k}(G), \Delta_{\mathcal{C}\setminus C_k}(G)\}} & \text{if } 2 \leq |C_k| < \sigma|\mathcal{G}| \end{cases} \tag{4.2.13}$$

The adapted SI considers four scenarios: (i) When a cluster contains too many TDNs ($\sigma|\mathcal{G}|$ or more), or (ii) a cluster contains a single TDN, but this TDN does not differ much (dissimilarity is less than τ) from an adjacent TDN, then the SI is set to be 0 to discourage the scenarios. (iii) However, when a single TDN is too dissimilar (dissimilarity is τ or larger) from adjacent TDNs to be grouped into other clusters, we let SI be 1 to allow the TDN to form a cluster by itself. (iv) When a cluster is neither too large (less than $\sigma|\mathcal{G}|$) nor too small (size is at least 2), we compute an SI that measures how a TDN is similar to its assigned cluster compared with other clusters. $\Delta_{C_k}(G) = \frac{1}{|C_k|-1} \sum_{\hat{G} \in C_k \setminus G} d(G, \hat{G})$ is the "internal distance" of G within its own cluster, defined as the average dissimilarity between G and the other networks in the cluster that G belongs to. While $\Delta_{\mathcal{C}\setminus C_k}(G) = \min\left\{ \frac{1}{|C_i|} \sum_{\hat{G} \in C_i} d(G, \hat{G}), \forall i \in \{1, 2, \cdots, p\} \,\middle|\, i \neq k \right\}$ is the "external distance" of G, and is evaluated with the smallest average dissimilarities between G and the clusters to which G does not belong. The scenarios establish "soft" bounds of 2 (lower bound) and $\sigma|\mathcal{G}|$ (upper bound) for the cluster size. "Soft" means that though discouraged, the bounds still can be exceeded if necessary.

The overall clustering quality can be then evaluated using the average SI of all TDNs in the sequence $\mathcal{G}$, i.e.

$$SI_{\mathcal{G}}(\mathcal{C}(p)) = \frac{1}{|\mathcal{G}|} \sum_{G \in \mathcal{G}} SI_{\mathcal{G}}(\mathcal{C}(p), G). \tag{4.2.14}$$

The value of $SI_{\mathcal{G}}(\mathcal{C}(p))$ falls within the range of $[-1, 1]$. The higher is the $SI_{\mathcal{G}}(\mathcal{C}(p))$, the more likely are TDNs clustered properly such that TDNs are close within each cluster, but distant across different clusters. Note that, given a TDN sequence $\mathcal{G}$, $\mathcal{C}(p)$ is determined by the solution of IP formulation (4.2.5)–(4.2.12) with the input parameter p. Hence, $SI_{\mathcal{G}}(\mathcal{C}(p))$ is essentially a function of the number of clusters $p \in \{1, 2, \cdots, m\}$. The desired value for the parameter p, p^* should be the one that maximizes this function; in other words,

$$p^* = \underset{p \in \{1,2,\ldots,m\}}{\operatorname{argmax}} \; SI_{\mathcal{G}}(\mathcal{C}(p)). \tag{4.2.15}$$

4.2.3 TDN Visualization and Disease Clustering

A major challenge of TDN visualization is that complex networks may include too many nodes and edges to be displayed in an intuitive and orderly manner. In our TDN Visualization Module, we propose and solve a *minimum atomic clique partition problem* (MACPP) to address this challenge, as elaborated in the following.

Definition 1 (Atomic Clique) *Given a collection of networks, $\mathcal{G} = \{G_1, G_2, \cdots, G_m\}$, a subset $S \subseteq \bigcup_{i=1}^{m} V(G_i)$ is called an atomic clique if S is a clique in $G_j, \forall j \in M$, but $S \cap V(G_k) = \emptyset, \forall k \notin M$, where $M = \{i \in \{1, 2, \cdots, m\} \mid S \subseteq V(G_i)\}$.*

Definition 1 requires that in any network $G_i \in \mathcal{G}$, all nodes in an atomic clique S are either forming a clique or completely absent. For example, in Figure 2.1, the atomic cliques across Window 1 and Window 2 are $\{1, 2, 3\}$, $\{4\}$, $\{5\}$. The clique $\{1, 2, 3\}$ represents the initial comorbid diseases in Window 1, while $\{4\}$ and $\{5\}$ are newly developed diseases in Window 2. They are not interconnected directly, indicating that from diseases $\{1, 2, 3\}$, patients are very likely to develop either disease $\{4\}$ or disease $\{5\}$, separately. Recall that in Section 2.2.3, we have shown that classical clique models could not necessarily capture this progression pattern. Instead, our proposed atomic clique model succeeds to address this challenge. Now, let us define MACPP which can decompose TDNs into a minimum set of atomic cliques.

Problem: Minimum atomic clique partition problem.
Input: A collection of networks, $\mathcal{G} = \{G_1, G_2, \cdots, G_m\}$.
Output: A collection of atomic cliques $\mathcal{K} = \{K_1, K_2, \ldots, K_q\}$ such that

- $K_i \cap K_j = \emptyset, \forall i, j \in \{1, 2, \ldots, q\} \mid i \neq j$
- $\bigcup_{i=1}^{q} K_i = \bigcup_{j=1}^{m} V(G_j)$
- q is minimized.

The partition nature of the problem requires that the atomic cliques are mutually exclusive and in a combination containing all nodes from the network collection. While the objective of minimizing the number of atomic cliques allows us to simplify the decomposition of the network collection as much as possible.

In this research, we developed an iterative algorithm (Algorithm 1) to find a feasible solution to MACPP. According to Definition 1, an atomic clique exists either in a single network or within an intersection of multiple networks. As a result, Algorithm 1 first finds a common node subset D across as many networks as possible through Lines 4–8. The initialization of D is performed at Line 4. Specifically, we assign the entire node set of the network G_k to D, where k is the smallest index of the networks remaining in $\mathcal{G}$. The nested while loop from Line 9 to Line 16 then seeks an atomic clique partition on all D-induced subgraphs, $G_i[D]$, $\forall i \in M$. Once we narrow down to $G_i[D]$, we can iteratively detect and remove a maximum atomic clique across all $G_i[D]$ each time by leveraging an IP formulation until a partition is formed. After an atomic clique partition is found on $G_i[D]$, the algorithm excludes D and repeats previous steps until all $G_i \in \mathcal{G}$ are empty.

The IP formulation we used to find a maximum atomic clique across $G_i[D], \forall i \in M$ is

Algorithm 1: Atomic clique partition algorithm

```
   Input: A collection of networks $\mathcal{G} = \{G_1, G_2, \cdots, G_m\}$.
   Output: An atomic clique partition $\mathcal{K}$.
1  $\mathcal{K} \longleftarrow \emptyset$
2  while $\mathcal{G} \neq \emptyset$ do
3      $M \longleftarrow \emptyset$
4      $D \longleftarrow V(G_k)$, where $k = \min\{i \mid G_i \in \mathcal{G}\}$
5      for $G_i \in \mathcal{G}$ do
6          if $D \cap V(G_i) \neq \emptyset$ then
7              $D \longleftarrow D \cap V(G_i)$
8              $M \longleftarrow M \cup \{i\}$
9      while $D \neq \emptyset$ do
10         find a subset $K \subseteq D$ such that $K$ is a clique in $G_i[D], \forall i \in M$ and $|K|$ is maximized by
             solving formulation (4.2.16)–(4.2.18)
11         $\mathcal{K} \longleftarrow \mathcal{K} \cup K$
12         for $i \in M$ do
13             $G_i \longleftarrow G_i[V(G_i) \setminus K]$
14             if $V(G_i) = \emptyset$ then
15                 $\mathcal{G} \longleftarrow \mathcal{G} \setminus G_i$
16         $D \longleftarrow D \setminus K$
17 return $\mathcal{K}$
```

presented in (4.2.16)–(4.2.18). The binary variable $x_j = 1$ if and only if $j \in D$ is selected in the solution. Constraint (4.2.17) ensures that at most one of nodes $j, k \in D$ can be included in the solution if j and k are disconnected in any single network $G_i[D]$, so the solution will be guaranteed to be an atomic clique. While the objective function aims to maximize the cardinality of the atomic clique.

$$\max \sum_{j \in D} x_j \tag{4.2.16}$$

$$x_j + x_k \leq 1 \qquad \forall \{j,k\} \in Q = \{\{j,k\} \subseteq D \mid \exists\, i \in M \text{ such that } \{j,k\} \notin E(G_i)\} \tag{4.2.17}$$

$$x_j \in \{0,1\} \qquad \forall j \in D, \tag{4.2.18}$$

Algorithm 1 is essentially a *greedy* algorithm because the IP formulation tries to find a maximum atomic clique in each iteration of the nested loop through Lines 9–16. Clearly, the algorithm returns a feasible solution to MACPP because each K found in one iteration is isolated from that found in other iterations, and $\mathcal{K}$ exhausts all nodes in $\mathcal{G}$.

4.3 Case Studies

To assess the effectiveness of our proposed system, we applied it to two case studies on analyzing and visualizing the comorbidity progressions during hospitalizations for C. Diff and stroke patients, respectively. In the case studies, our system was implemented using Python 3.7, and the IP formulations involved were solved using a state-of-the-art optimization solver—Gurobi 8.1.1 [Gurobi Optimization, LLC, 2020].

4.3.1 Data Cohorts and Data Preparation

We integrated Cerner Health Facts® EHR data warehouse as the data source into our system. Health Facts® contains clinical data extracted directly from the U.S. hospitals that operate on Cerner EHR systems. Cerner Corporation collects and integrates the data through its established operations in compliance with the Health Insurance Portability and Accountability Act (HIPAA) laws. Because the data has been completely de-identified according to HIPAA regulations, the Institutional Review Boards (IRB) at Oklahoma State University exempted the study from review.

C. Diff is a bacterial infection that is mostly hospital-acquired among senior patients [CDC, 2019], while stroke is one of the leading chronic conditions for death/disability in the U.S. [Members et al., 2016]. Our C. Diff and stroke study cohorts were extracted from Health Facts® using International Classification of Diseases 9th/10th Revision (ICD-9/10) codes (the ICD-9/10 codes are listed in the Supplementary Material). The cohorts included hospitalized encounters of female patients aged 65 or older with the onset of C. Diff/stroke between November 1999 and August 2017. Patient age, length of stay (LOS), and all diagnoses associated with the encounters were exported as well.

Our data pre-processing mainly dealt with outlying LOS, erroneous diagnoses, and diagnosis combination. In order to exclude extreme outliers in LOS, we restricted the analysis to the encounters of LOS within the range of 24 hours to 14 days, which is a common range for inpatient hospital stays. We noticed that the data included some infeasible diagnoses, such as birth/labor-related diagnoses and male conditions. Encounters with such erroneous diagnoses were excluded from the study cohorts. Furthermore, the ICD-9/10 codes used in Health Facts® can be overly specific to express disease states in the usual sense. We used the Clinical Classifications Software (CCS) [AHRQ, 2020] to aggregate ICD-9/10 codes into relatively high-level disease states. For example, CCS combines malignant neoplasms at different locations of esophagus together as the "cancer of esophagus". Our data extraction and preprocessing eventually resulted in two large datasets containing hundreds of thousands or millions of encounters and diagnosis records as shown in Table 4.1 (under the "Enct #" and "Diag #" columns).

4.3.2 TDN Construction

In Health Facts®, diagnoses were recorded in encounters, but lacking specific timestamps about at what time during the encounter a condition was diagnosed. In other words, given time points $t_1 < t_2 < \cdots < t_m$ within an encounter, we cannot tell what diagnoses occurred exactly during a time interval $[t_i,\ t_{i+1}]$. Therefore, we defined the windows based on LOS as Warner et al. did in their studies on hospital-acquired complications [Warner et al., 2013, Warner et al., 2016]. In particular, Window i includes all encounters with LOS $\in \Big[l + (i-1)\epsilon, l + i\epsilon\Big)$, where l is the smallest LOS included for analysis (24 hours in our case studies in light of the data preparation). The rationale is that when a large sample is included in a window, the statistical results based on the sample can be considered as the expected values of the attributes of a general population in the window. Then, the changes newly happened to Window $i+1$ from Window i can be well representative of the events occurring within the interval $\Big[l + i\epsilon, l + (i+1)\epsilon\Big)$ for the population. In our case studies, we specified $\epsilon = 12$ hours, which resulted in 26 windows in total, i.e., $m = 26$.

Then, we built networks over the 26 windows with SCI threshold $\theta = 0.05$. Since our interest was concentrated on the progression of C. Diff/stroke and its strongly coexisting diseases, we only considered the ego networks of C. Diff/stroke as the TDNs for analysis and visualization henceforth. The TDNs constructed based on our C. Diff and stroke cohorts are visualized in Figure 4.4 and Figure 4.5 respectively. The orders and sizes of the TDNs are listed in Table 4.1.

Table 4.1: Statistics of Encounters and TDNs in Each Window

Window	C. Diff - Senior Female Cohort				Stroke - Senior Female Cohort			
	Enct #	Diag #	$\lvert V\rvert$	$\lvert E\rvert$	Enct #	Diag #	$\lvert V\rvert$	$\lvert E\rvert$
1	158,408	1,088,614	6	15	13,070	57,703	105	602
2	173,611	1,400,515	7	21	11,641	55,473	110	727
3	200,907	1,625,593	9	34	11,721	55,131	108	723
4	196,913	1,738,845	14	90	11,201	54,259	113	776
5	242,541	2,056,285	12	65	10,251	50,313	113	821
6	173,887	1,612,831	26	321	8,697	42,569	114	843
7	164,309	1,533,572	28	372	7,392	36,497	117	889
8	128,370	1,246,318	32	482	6,348	31,215	119	936
9	116,790	1,147,284	34	547	5,259	26,660	120	1,072
10	95,243	962,605	39	692	4,680	23,286	122	1,101
11	86,439	883,730	35	584	3,850	19,480	120	1,147
12	72,727	750,208	42	811	3,507	17,752	123	1,250
13	66,558	703,248	44	931	2,868	14,422	118	1,308
14	55,194	581,826	50	1,195	2,660	13,407	120	1,371
15	47,954	519,614	54	1,401	2,061	10,627	121	1,551
16	39,010	423,053	54	1,360	1,849	9,288	117	1,516
17	33,844	380,795	61	1,738	1,501	7,780	124	1,660
18	30,021	332,668	62	1,754	1,502	7,637	125	1,655
19	25,702	295,288	62	1,806	1,157	6,028	131	1,831
20	23,662	265,881	61	1,739	1,155	6,013	133	1,864
21	20,171	233,745	68	2,155	924	4,816	117	1,682
22	18,604	213,280	66	2,039	936	4,894	125	1,859
23	15,760	187,325	70	2,275	738	3,957	133	1,893
24	15,514	177,554	74	2,484	839	4,361	122	1,771
25	13,080	155,944	75	2,595	586	3,076	123	1,707
26	13,832	160,786	67	2,139	869	4,542	121	1,770
Total	2,229,051	20,677,407	–	–	117,262	571,186	–	–

Note: TDNs are ego networks.

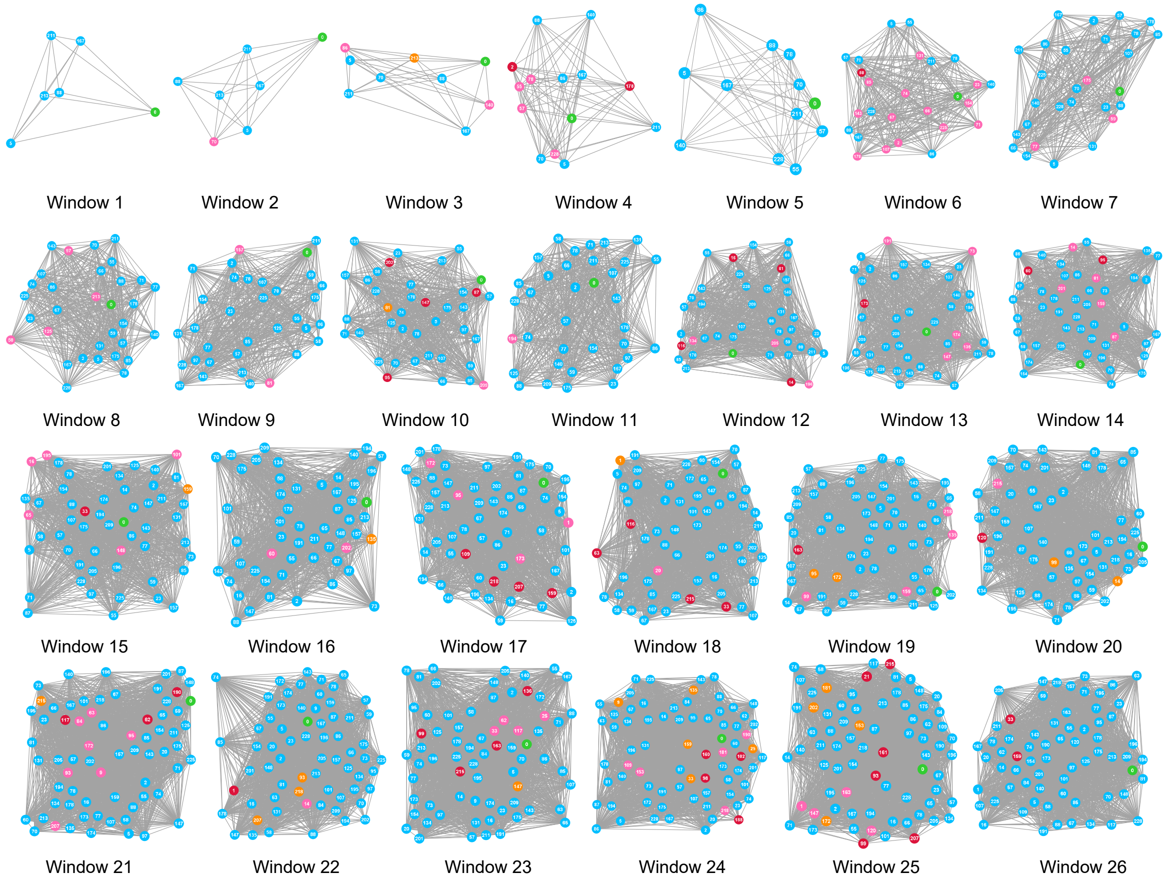
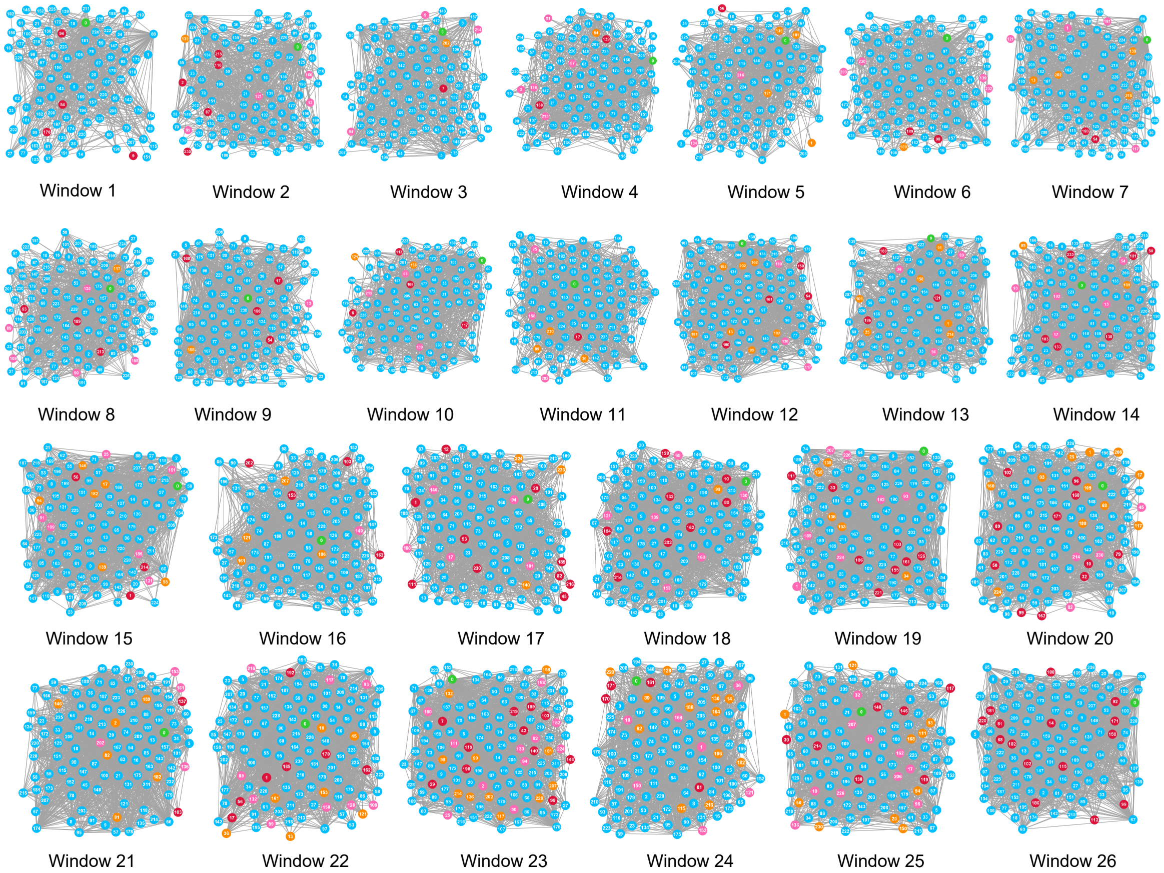


Figure 4.4: TDNs Constructed for the C. Diff Cohort (Senior Female Patients)

Note: The node color is used to indicate the existence pattern of a node in adjacent windows. C.Diff node is in green color through all windows. Given a window, a blue node indicates that the node also appears in both adjacent windows or the unique adjacent window. A red node means that the node does not appear in any adjacent window(s). Pink means that the node also appears in the next window but not in the previous window, while orange indicates that the node also occurs in the previous window but not in the next window.



Window 1 Window 2 Window 3 Window 4 Window 5 Window 6 Window 7

Window 8 Window 9 Window 10 Window 11 Window 12 Window 13 Window 14

Window 15 Window 16 Window 17 Window 18 Window 19 Window 20

Window 21 Window 22 Window 23 Window 24 Window 25 Window 26

Figure 4.5: TDNs Constructed for the Stroke Cohort (Senior Female Patients)

4.3.3 Temporal Clustering

Figure 4.6 shows the heat maps of dissimilarities among TDNs and the SI charts for p. Diagrams (A) and (C) are for C.Diff while (B) and (D) are for stroke. Let's first talk about C.Diff. The dissimilarity between each pair of the TDNs of the C. Diff cohort is calculated and plotted as a heat map shown in Figure 4.6 (A). A dark cell indicates that the two networks are similar to each other. From the heat map, we may roughly observe that (i) there exist a few dark blocks, which correspond to clusters of windows that may imply progression phases; and (ii) the phases tend to include more windows over time, indicating that comorbidity evolves more rapidly at earlier phases compared with later phases.

We now present the CPMP results on this TDN sequence to demonstrate CPMP's effectiveness to capture the observations algorithmically. In order to solve the CPMP on this TDN sequence, we firstly used the *SI* method described in Section 4.2.2 to determine a proper p^* for the TDNs. During the calculation of *SI*, we let both the parameters τ and σ be 0.5, meaning we do not intend to cluster a window with its adjacent window(s) if the dissimilarity is no less than 0.5, and we discourage a cluster that includes half or more of all windows since it might be overly broad. The *SI* result in Figure 4.6 (C) shows that $p^* = 5$, indicating that the entire window sequence should be clustered into five phases.

Given $p = p^* = 5$, the CPMP solution is: Phase 1 includes Windows $1 - 3$, Phase 2 contains Windows $4 - 5$, Phase 3 consists of Windows $6 - 11$, Phase 4 is comprised of Windows $12 - 20$, and Phase 5 includes Windows $21 - 26$. The corresponding days of the phases are shown in Table 4.2. The results are aligned with the observations we can inspect from Figure 4.6 (A), demonstrating that the proposed consecutive p-median model is capable to identify the progression patterns algorithmically.

The stroke results are presented in Figure 4.6 (B) and Figure 4.6 (D). Figure 4.6 (D)

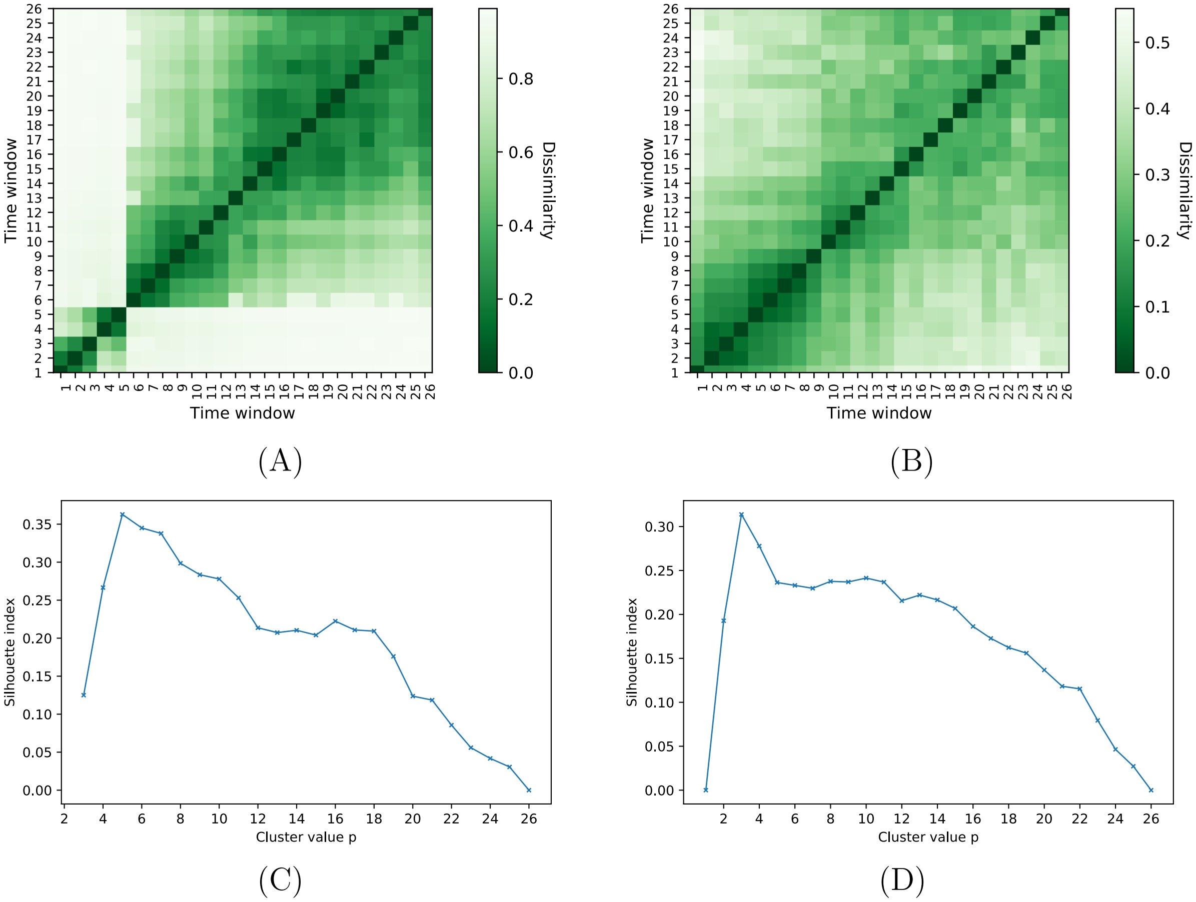


Figure 4.6: Heat Maps of Dissimilarities among TDNs and the SI Charts for p

Note: (A) and (C) are diagrams for C. Diff; (B) and (D) are diagrams for stroke.

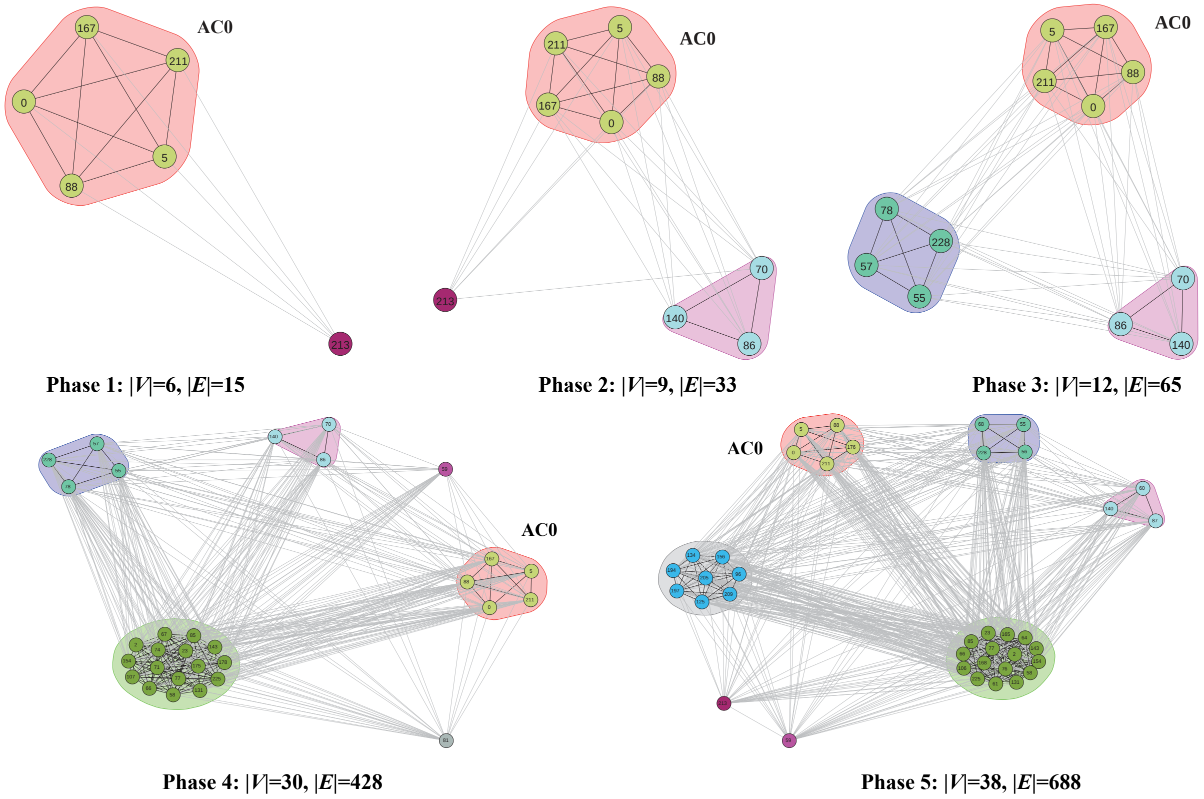
shows that $p^* = 3$, implying that hospitalized stroke patients may experience three phases: Phase 1 includes Windows 1 – 8, Phase 2 contains Windows 9 – 15, and Phase 3 consists of Windows 16 – 26, as shown in Table 4.2. Similar to the C. Diff results, the phases outlined by the proposed consecutive p-median model are also in line with what we can observe from Figure 4.6 (B).

Table 4.2: Phases and Corresponding Windows and Days

Cohorts	Time Unit	Phase 1	Phase 2	Phase 3	Phase 4	Phase 5
C. Diff	Window	1 – 3	4 – 5	6 – 11	12 – 20	21 – 26
	Day	2 – 3	3 – 4	4 – 7	7 – 11	12 – 14
Stroke	Window	1 – 8	9 – 15	16 – 26	–	–
	Day	2 – 5	6 – 9	9 – 14	–	–

4.3.4 Visualization of TDNs in Phases

By visualizing TDNs in the identified phases, we can reduce the complexity of the entire TDN sequence over time. However, the complexity inside a single TDN remains because some TDNs can include many nodes and edges. For example, the C. Diff TDN at Phase 5 includes 688 edges incident to 38 nodes. Visualizing such dense networks in a user-friendly format will significantly facilitate subsequent inspection and analysis. To that end, we firstly found an atomic clique partition using Algorithm 1. Then, for the TDN at every phase, we plotted each atomic clique together in a compact, shaded space. In addition, to keep consistency, each atomic clique was rendered in the same color across all phases.



Phase 1: $|V|$=6, $|E|$=15

Phase 2: $|V|$=9, $|E|$=33

Phase 3: $|V|$=12, $|E|$=65

Phase 4: $|V|$=30, $|E|$=428

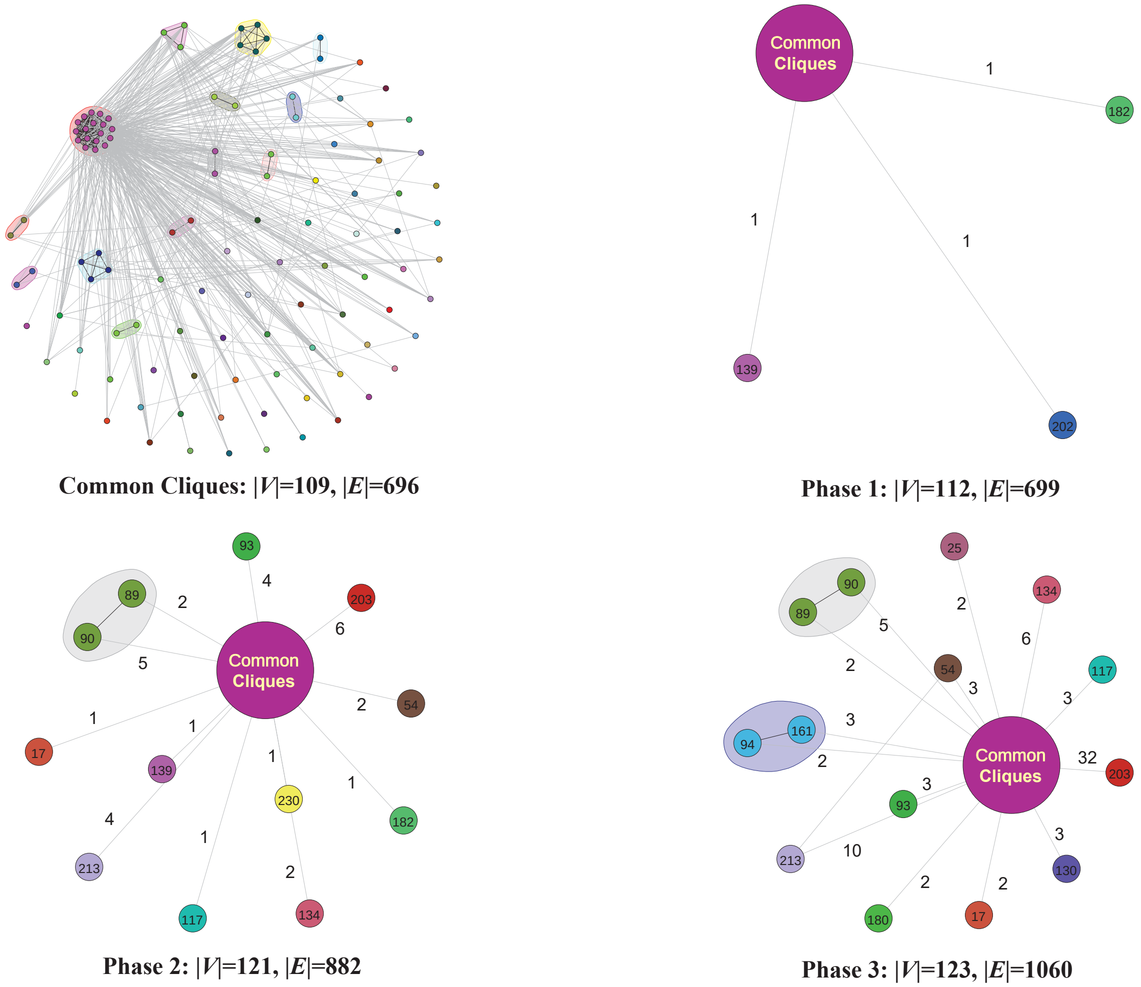
Phase 5: $|V|$=38, $|E|$=688

Figure 4.7: TDNs Constructed on the Phases of the C.Diff Cohort

The C. Diff comorbidity progression is visualized in Figure 4.7, from which we can observe that acute renal failure (node 5), fluid and electrolyte disorder (node 88), other gastrointestinal disorders (node 167), and septicemia (node 211) along with C. Diff (node 0) form an atomic clique that occurs persistently across all phases (marked as AC0 in Figure 4.7). It implies that these diseases are highly coexisting with C. Diff throughout the entire time frame. Many clinical studies [Bauer et al., 2012, Doshi et al., 2018] have reported similar findings that these diseases are highly associated with C. Diff, thus validating our VA results.

Another interesting progression pattern we can inspect from Figure 4.7 is that instead of occurring independently, the comorbidities that appeared at later phases tend to form atomic cliques as well. In other words, the onset of one of these diseases may indicate one or more other conditions in the same atomic clique. For example, urinary tract infections (UTI, node 228) appears in Phases 3 – 5, which echoes a previous study finding that UTI is associated with prolonged hospitalization of C. Diff patients [Warner et al., 2013]. Furthermore, our approach discovers that UTI occurs in an atomic clique that also includes cardiac dysrhythmias (node 55), chronic kidney disease (node 57), and disorders of lipid metabolism (node 78). It suggests that doctors should pay attention to not only UTI but also these UTI-associated diseases to prevent prolonged hospitalization.

The stroke comorbidity progression is visualized in Figure 4.8, which shows that a few diseases start to be highly coexisting with stroke after Phase 1. It implies that these disease states are highly associated with prolonged hospitalization of more than one week for stroke patients. This association of some of the diseases, such as mental health disorders (node 130) and shock (node 213), is also supported by other clinical studies [Siddiqui et al., 2018, Myint et al., 2018]. Furthermore, other two risk factors for prolonged hospitalizations—fracture of lower limb (node 89) and fracture of hip (node 90)—occur in the same atomic clique. It indicates that these two conditions are very likely to occur together, which may be resulted from post-stroke fall [Schmid et al., 2010].



Common Cliques: |*V*|=109, |*E*|=696

Phase 1: |*V*|=112, |*E*|=699

Phase 2: |*V*|=121, |*E*|=882

Phase 3: |*V*|=123, |*E*|=1060

Figure 4.8: TDNs Constructed on the Phases of the Stroke Cohort

Note: There exists a set of *common cliques* throughout all the phases. The common cliques are visualized in detail in the upper left part of the figure and simplified as a large node in the TDNs across the phases. The edge weight in the TDNs indicates how many nodes inside the set of common cliques are connected to a node outside the common cliques.

4.4 Discussion

Our proposed VA system for comorbidity progression has significant implications in both the technology advance and healthcare applications, as discussed in the following.

Technical Contributions. The highlight of this research from a technical perspective is that we investigate the temporal and disease clustering of TDNs for the *first* time. In this effort, two new problems and associated algorithms, rooted from the methodology for the single network, were extended to network sequences (i.e., TDNs) to address the challenges in implementing the temporal and disease clustering of TDNs:

- *The consecutive p-median problem* was extended from the classical p-median problem, by requiring each cluster to only include consecutive objects (TDNs in our case) to model temporal clustering. An IP formulation was developed to solve the problem, and the classical Silhouette Index was modified to determine a suitable value of p.
- *The minimum atomic clique partition problem* was extended from the minimum clique partition problem for a single network to clustering diseases across a sequence of TDNs. A greedy heuristic algorithm was developed to find a feasible solution for the problem.

Application in Healthcare. Supported by the temporal clustering module, our proposed system can automatically detect the comorbidity progression phases. Because the disease states and coexistence relationships are highly similar within each phase while remarkably distinct across different phases, the end of a phase can indicate a beginning time point of significant progression changes. Furthermore, through our visualization module, we are able to show the comorbidity coexistence relationships and progression patterns visually and concisely. It can help doctors understand when and what diseases are most likely to be comorbid with the index disease and plan prevention and treatments. For example, in our stroke case study, the VA results in Figure 4.8 show that fractures are associated with prolonged hospitalization for more than one week. Furthermore, the fractures often include both lower limb and hip fractures. By being aware of this fact, hospitals and doctors can prepare proper care resources to prevent/handle both types of fractures during patients' hospitalizations. In addition, the TDNs can be used to compare different subgroups of patients, such as matched case-control cohorts based on a certain treatment [Kim et al., 2018] to evaluate the treatment's efficacy or different gender groups [Kalgotra et al., 2017] to reveal progression disparities between genders.

Limitations. This research mainly has two limitations. First, in the literature there are many approaches for network dissimilarity measurement. The choices of the method may influence the temporal clustering results. However, a systematic review and comparison of all the methods for our problem is beyond the scope of this study. Second, our temporal and disease clustering approaches only work on undirected, unweighted networks. TDNs can be more sophisticated by carrying node attributes (like disease frequency), edge weight (like *SCI* value), and edge direction (like presence order). Performing temporal and disease clustering on such complex TDNs requires corresponding dissimilarity measurement methods and graphical cluster models. Nevertheless, many of the approaches are either still absent or require much effort for suitable adaptions. As a result, we leave these for future work.

4.5 Conclusion

Comorbidity is a prominent challenge in healthcare practice and research. We modeled comorbidity progression as a sequence of TDNs, and designed a VA system, which integrates novel temporal and disease clustering technologies to visualize progression patterns from the TDN sequence. Two case studies of applying the system to C. Diff and stroke demonstrate the effectiveness of the system. Based on the discussion in Section 4.4, we summarize two directions for our future work—*healthcare application* and *technical improvement*. From the healthcare application perspective, we plan to apply the proposed system to more diseases to mine useful insights for healthcare practice. We will also incorporate more biomarkers besides comorbidity during the applications to reveal more progression patterns. In order to improve the proposed technologies, we plan to extend our temporal and disease clustering approaches to more sophisticated TDNs that can carry node attributes and edge weights. In this study, we proposed a heuristic algorithm for MACPP, which does not necessarily find a minimized solution. Hence, we are interested in developing exact algorithms, such as IP formulations, which are able to provide optimal solutions for MACPP in our future work.

CHAPTER V

EXTRACTING PATIENT HISTORY INFORMATION FROM CLINICAL NOTES FOR MEDICAL BILLING[1]

5.1 Introduction

As we discussed in Section 1.2, medical billing is an important yet demanding task in the healthcare revenue management cycle and the extraction of patient history information is critical for proper billing. The *objective* of this study is to develop natural language processing (NLP) systems that can effectively recognize three important categories of patient history information—*chief complaint (CC)*, *history of present illness (HPI)*, and *past, family and/or social history (PFSH)* [CMS, 2020]—directly from clinical notes. An occurrence of such patient history information is called an *entity* and this clinical information extraction (CIE) study is a typical named entity recognition (NER) task.

As detailed in Table 5.1, CC is a brief statement that describes the major reason for a medical encounter, often in the patient's own words. CC can be about a symptom, problem, condition, diagnosis, or even physician recommended followup. A simple example can be "CC: Right foot pain."

HPI describes how a patient's present illness developed from the first sign/symptom or the previous encounter to the present. HPI mainly deals with eight elements, which are *quality*, *location*, *severity*, *duration*, *timing*, *context*, *modifying factors*, and *associated*

[1] This is a joint work with Tuan-Dung Le, Dr. Thanh Thieu, Dr. Andrew Gin, Dr. Phuong D.Nguyen, Dr. Tieming Liu and Dr. Zhuqi Miao.

Table 5.1: Descriptions and Examples of History Elements for E/M Services

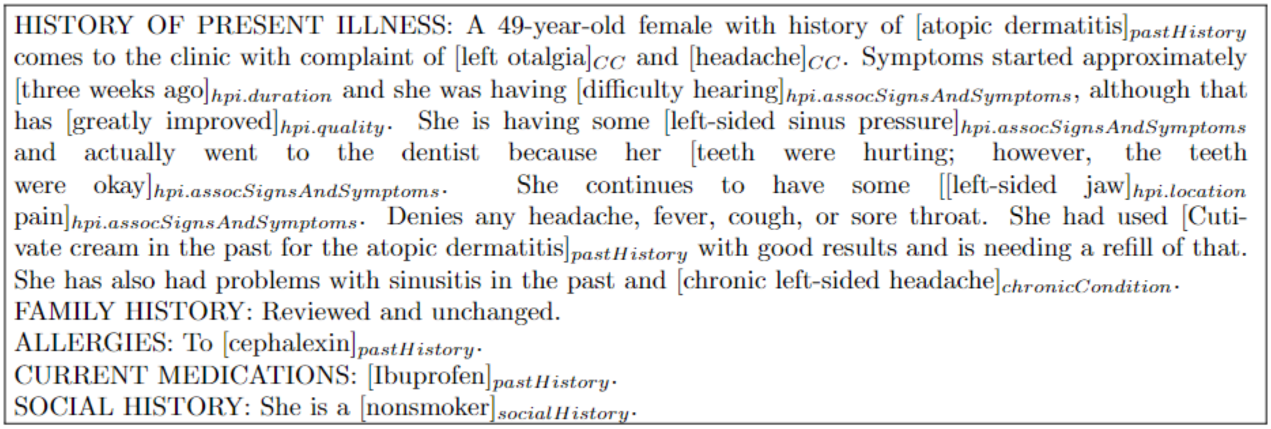
Element	**Sub-element**	**Description**	**Example**
CC		a brief statement on the reason for a medical encounter	right foot pain
HPI	Location	where the complaint is located	right foot
	Quality	the nature of the problem	aching pain
	Severity	how bad the problem is	6 on a scale of 1 to 10
	Duration	how long it has existed	it started two days ago
	Timing	any onset pattern for the complaint	constant
	Context	any specific activity associated with the main complaint	harvested corns
	Modifying Factors	what prior treatment or medication has been tried	better when heat is applied
	Associated Signs or Symptoms	what symptoms or signs that accompany the main complaint	numbness, fatigue
PFSH	Past History	the patient's past medical history	diabetes, hypertension
	Family History	the patient's family medical history	father has dementia
	Social History	the patient's social medical history	nonsmoker, drinks occasionally

signs and symptoms. Quality indicates the *nature* of the problem, symptom, or pain, often about how they feel (e.g., sharp, dull, constant, intermittent, and improved/worsening). The location, severity, and duration elements refer to *where*, *how bad* (e.g., pain scale 1–10 and mild/severe), and *how long* the problem exists. While timing, context, modifying factors, and associated signs/symptoms tell *what timing pattern* comes with (e.g., in the morning and after meals), *what activities accompany the problem*, *what actions the patient has taken to address the problem and whether the problem improves or worsens*, and *what other symptoms or signs co-occur with the problem*, respectively.

PFSH consists of reviews in three aspects regarding the patient's history before the present illness: *past medical history*, *family history*, and *social history.* Past medical history involves the patient's past medical experiences with illnesses, injuries, operations, medications, and/or allergies, among others. Family history contains a review of medical events in the patient's family, mainly about the diseases that can be inherited by or occur in the patient at risk. Social history is an age-appropriate review of the patient's past and current activities such as marital status, living status, alcohol usage, exercises, and hobbies.

Note that in the 1997 Evaluation/Management (E/M) services documentation guidelines, chronic conditions are considered auxiliary elements in HPI. Therefore, in addition to extracting the aforementioned CC, HPI, and PFSH elements, our algorithms and systems are designed to be able to recognize Chronic Conditions as well.

An excerpt from an example clinical note and the corresponding history element annotations are illustrated in Figure 5.1.



HISTORY OF PRESENT ILLNESS: A 49-year-old female with history of [atopic dermatitis]$_{pastHistory}$ comes to the clinic with complaint of [left otalgia]$_{CC}$ and [headache]$_{CC}$. Symptoms started approximately [three weeks ago]$_{hpi.duration}$ and she was having [difficulty hearing]$_{hpi.assocSignsAndSymptoms}$, although that has [greatly improved]$_{hpi.quality}$. She is having some [left-sided sinus pressure]$_{hpi.assocSignsAndSymptoms}$ and actually went to the dentist because her [teeth were hurting; however, the teeth were okay]$_{hpi.assocSignsAndSymptoms}$. She continues to have some [[left-sided jaw]$_{hpi.location}$ pain]$_{hpi.assocSignsAndSymptoms}$. Denies any headache, fever, cough, or sore throat. She had used [Cutivate cream in the past for the atopic dermatitis]$_{pastHistory}$ with good results and is needing a refill of that. She has also had problems with sinusitis in the past and [chronic left-sided headache]$_{chronicCondition}$.
FAMILY HISTORY: Reviewed and unchanged.
ALLERGIES: To [cephalexin]$_{pastHistory}$.
CURRENT MEDICATIONS: [Ibuprofen]$_{pastHistory}$.
SOCIAL HISTORY: She is a [nonsmoker]$_{socialHistory}$.

Figure 5.1: A Clinical Note Example and History Element Annotations

5.2 Methodology

As illustrated in Figure 5.2, the methods of our study involved creating rule-based algorithms and deep learning (DL) models and applying them to a set of annotated clinical notes to extract history elements therein. The CIE performances of these two approaches were then evaluated. The technical details of each process within the flowchart are elaborated in the remainder of this section.

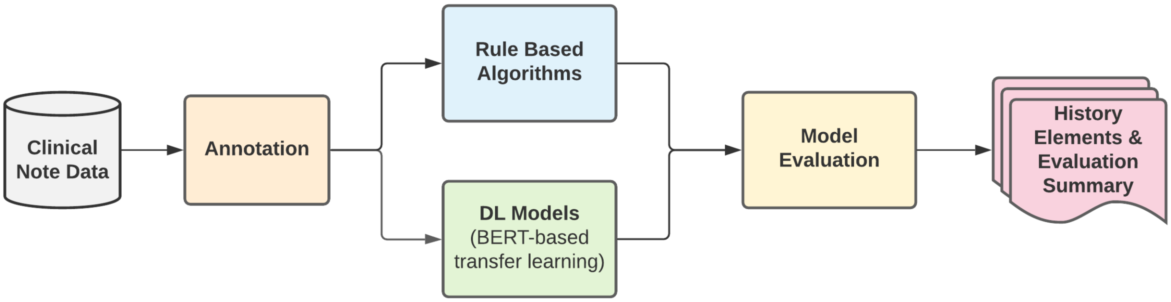

Clinical Note Data
Annotation
Rule Based Algorithms
DL Models (BERT-based transfer learning)
Model Evaluation
History Elements & Evaluation Summary

Figure 5.2: Flowchart of the Methods and Experimental Design

5.2.1 Clinical Notes

In this study, we used the Medical Transcription Sample Reports and Examples (MTSamples) as our data source. MTSamples is one of the most popular clinical note repositories among the medical and medical informatics research communities [Wang et al., 2018b, MTHelpLine, 2022]. The structure of MTSamples notes is very similar to that shown in Figure 5.1. Each note is organized into several sections with each section starting with an explicit section heading, followed by free-text narratives about patient visits. Since the notes are transcribed, the section headings are generally correct, and different sections represent different types of information (e.g., a section with a section heading “Chief Complaint” usually contains the information of the chief complaint). Note that all the MTSamples notes have been completely de-identified according to Health Insurance Portability and Accountability Act (HIPAA) regulations.

Due to the complexity of clinical notes and the annotation workload, we selected and annotated 61 clinical notes from MTSamples as the benchmark dataset for the study. The benchmark dataset included 27 consultation notes, 12 SOAP (“subjective, objective, assessment, and plan”) reports, 6 emergency room reports, 3 followup notes, 3 history and physical notes, and 10 miscellaneous notes. The annotations were first completed independently by two undergraduate students who majored in biology. They were then curated by two other collaborators, followed by verification and adjustment by an experienced physician. There were 1,648 labels annotated in total for the selected notes as described in Table 5.2. Note that there are many overlapping labels in our dataset, e.g., most Chronic Condition entities are also labeled as Past History entities at the same time.

Table 5.2: Label Counts of the Annotated Dataset

Entity type	Total	Overlap percentage (%)	Most overlap with (#overlap)
CC	138	22.64	Chronic Condition (17)
Chronic Condition	171	93.57	Past History (120)
HPI-location	69	46.38	CC (16)
HPI-quality	57	10.53	CC, HPI-associated signs/symptoms (3)
HPI-severity	27	18.52	HPI-associated signs/symptoms (4)
HPI-duration	67	4.48	HPI-context (2)
HPI-timing	37	8.11	HPI-associated signs/symptoms (3)
HPI-context	37	10.81	HPI-location (3)
HPI-modifying factors	82	2.44	Past History (2)
HPI-associated signs/symptoms	269	12.64	HPI-location (13)
Past History	520	23.46	Chronic Condition (120)
Family History	45	2.22	Chronic Condition (1)
Social History	129	0	

5.2.2 Rule-Based Algorithms

Figure 5.3 shows the design of our rule-based algorithms, which involve two steps of processes, (1) pre-classification and (2) rule-based entity recognition, as elaborated in the following.

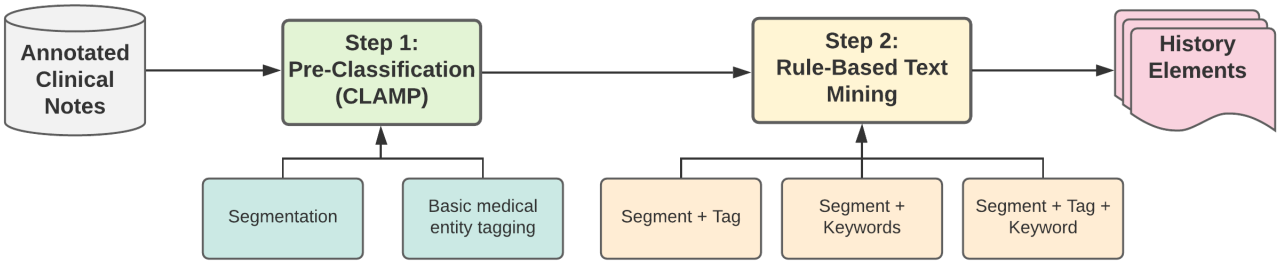


Figure 5.3: Rule-based Algorithms

Pre-Classification. Section headings can provide useful information for classifying the text included in the corresponding sections. As a result, the first process of pre-classification is to perform section segmentation on the given notes by leveraging their headings. Following the segmentation, the second process is to tag basic medical entities, including problems, tests, treatments, body parts, etc. Besides what section a word/phrase belongs to, what tag it possesses is also useful for classifying the text in the subsequent recognition step.

We used CLAMP 1.6.1 to implement the two processes. CLAMP is a popular integrated clinical NLP software and has been increasingly employed to analyze narrative patient reports. Although CLAMP possesses an embedded heading lexicon for segmentation, it was not sufficiently comprehensive to handle all the headings in our dataset. Hence, we added two types of extra keywords to CLAMP's heading lexicon to enhance the segmentation accuracy, (1) the keywords that no heading in CLAMP is exactly the same as or has a similar meaning to, such as "subjective", "diagnosis", and "service", and (2) the syntactic variations of headings already built in CLAMP, such as "course in hospital" for "hospital course".

Rule-Based Entity Recognition. This step involves three types of rules to recognize history elements from either the original text or the pre-classification outputs.

- Type-1 rules consider the combinations of *section headings* and *basic medical tags.* For instance, when a disease tag appears in the chief complaint section, it is considered a CC element. Whereas, when it appears in the past medical history section, it is considered a Past History element.

- Type-2 rules leverage the combinations of *section headings*, *keywords*, and *basic medical tags.* For example, a problem-type medical tag following the keyword "complaints of" in the section "subjective" or "history of present illness" often indicates a CC entity. Similarly, a body location tag following the keyword "issues with" or "problems with" in the two above sections also indicates a CC entity.

- Type-3 rules consider the combinations of *section headings* and *keywords*. For example, keywords such as “mild”, “moderate”, and “severe” were used to recognize HPI-Severity elements in sections “subjective” or “history of present illness”. Furthermore, in order to detect chronic conditions, we developed a library of UMLS (Unified Medical Language System) CUIs (Concept Unique Identifier) for chronic conditions defined by Chronic Conditions Data Warehouse (ccwdata.org). CUIs of entities recognized by CLAMP are matched against the library within some sections to identify chronic conditions.

5.2.3 Deep Learning Models

Since we only have a limited number of annotated notes, it is not appropriate for us to build DL models from scratch and train the models by ourselves as we often do in other studies. Instead, the DL approach employed in this study is a transfer learning scheme based on a pre-trained model named the Bidirectional Encoder Representation from Transformers (BERT). The method involves adapting three machine learning concepts—*Transformer*, *BERT*, and *Transfer Learning*—as elaborated in the following.

Transformer. Since its debut in 2017, the transformer has become an increasingly popular technique in the field of NLP [Vaswani et al., 2017]. Prior to the transformer, most DL-based NLP techniques were built using RNNs (e.g., LSTM), which process text sequentially. By contrast, transformers dispense with the sequential processing by fully leveraging the attention mechanism, which can assess every token and its context more independently at the same time. This nature allows transformers to be more parallelizable in programming, thereby making the model training of transformers on large-scale corpora possible.

BERT. The transformer’s strength in parallel computing led to the development of NLP models pre-trained on large-scale language datasets. BERT is one of such transformer-based pre-trained models. The BERT model used in our study is an open-source, clinical BERT pre-trained on MIMIC-III discharge summary notes [Alsentzer et al., 2019]. The “bidirectional” mechanism allows BERT to exploit both the left (i.e., earlier) and right (i.e., later) contexts of each token. By being trained on two unlabeled tasks — masked language modeling (a certain proportion of words are masked at random, and the model is trained to predict them from context) and next sentence prediction (given a sentence, the model is trained to predict whether another selected sentence is probably the next sentence), clinical BERT gained an “understanding” about the vocabulary, syntax, and phrasing that are used in clinical note documentation. The understanding is represented and output as a vector $C \in \mathbb{R}^{H}$, where $\mathbb{R}$ stands for the set of all real numbers. The dimension of the vector, H, also known as “hidden size”, is 768 according to the implementation [Alsentzer et al., 2019].

Transfer Learning. Given the pre-trained clinical BERT model, we used transfer learning to adapt it to our downstream problem in recognizing history elements. Transfer learning aims to transfer ML knowledge/models gained from solving one task to a different but related task [Torrey and Shavlik, 2010]. The transfer strategy we used was fine-tuning, which introduced a linear layer on top of the clinical BERT model, as shown in Figure 5.4. We used the BIO tagging schema which classifies each token in the sentence into “begin” (with a tag of $B - E$), “inside” (with a tag of $I - E$), or “outside” (with a tag of O) of an entity category E. Suppose that $x = [x^{(1)}, x^{(2)}, \cdots, x^{(n)}]$ represents a sentence which consists of a sequence of n words, $y = [y^{(1)}, y^{(2)}, \cdots, y^{(n)}]$ represents the sequence of NER tags, and the task is to predict the entity tag $y^{(i)} \in Y$ for each word $x^{(i)}$ where $y^{(i)}$ can be $O, B - cc, I - cc, B - hpi.quality, I - hpi.quality, \cdots$. The total number of tags is 27 in our setting (two possible tags for each of the 13 entity types plus an “O” tag). We used a pre-trained clinical BERT model as an encoder θ to obtain token representations and then classify them into tags by simply adding a linear layer ($w \in R^{768 \times 27}$). Applying standard fine-tuning, the model is trained to minimize the cross-entropy loss:

$$L = -\sum_{i=1}^{n} f_{i,y_i}(x^{(i)}; \theta, w) = -\sum_{i=1}^{n} f_{i,y_i}(h(i); w)$$

where $h = [h^{(1)}, h^{(2)}, \cdots, h^{(n)}]$ are token embeddings corresponding to the input x.

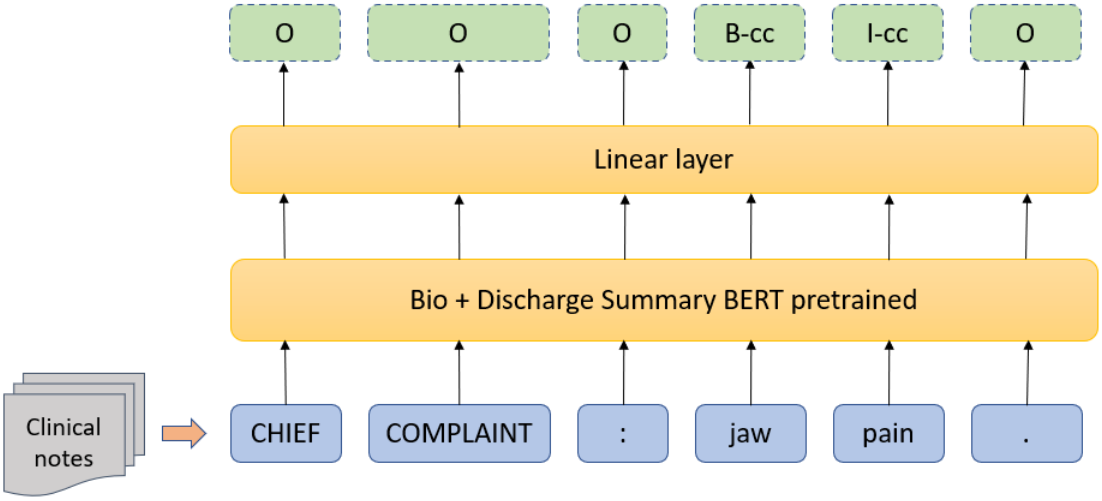


Figure 5.4: BERT Transfer Learning Architecture

5.2.4 Model Evaluation Metrics

We evaluated the two NLP models with two types of metrics: the exact-match metric and the relaxed-match metric [Li et al., 2020]. With an exact-match metric, a prediction is considered correct only when both the text span and the type of an entity are exactly the same with the gold-standard annotations. With a relaxed-match metric, if the type of an entity is correct and its text span overlaps with the ground-truth annotations, the prediction is considered partially correct. Relaxed-match metric also distinguishes different types of errors, which allows us to examine model performance in detail.

Exact-Match Metric. We used the typical exact-match metric in NLP to evaluate the performances of our rule-based and deep learning models. The metric included precision, recall, and F-1 scores. Precision is the fraction of predicted entities that are correct according to gold standard notes while recall is the fraction of the entities in gold standard notes that are successfully predicted. F-1 score is a measure that combines both precision and recall.

$$\text{Precision} = \frac{|\{\text{Predicted Labels}\} \cap \{\text{Gold-standard Labels}\}|}{|\{\text{Predicted Labels}\}|}$$

$$\text{Recall} = \frac{|\{\text{Predicted Labels}\} \cap \{\text{Gold-standard Labels}\}|}{|\{\text{Gold-standard Labels}\}|}$$

$$\text{F1 score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

The evaluations are performed with the help of the seqeval [Nakayama, 2018] script, which is a Python package for sequence labeling evaluation and tested by the widely used conlleval [Sang, 2004] script.

Relaxed-Match Metric. One widely adopted relaxed-match metric for named entity recognition (NER) is the Fifth Message Understanding Conference (MUC-5) Evaluation Metrics [Chinchor and Sundheim, 1993]. MUC-5 compares gold-standard annotations with predictions using six categories, i.e., Correct, Partially Correct, Incorrect, Spurious, Missing, and Non-committal, and defines a special formula to calculate error counts. A similar but simpler metric defines five types of errors including wrong range, wrong tag, wrong range and tag, no extraction, and no annotation [Ichihara et al., 2015]. However, these existing relaxed-match metrics are not suitable for overlapping or nested NER evaluation. For example, they all require that “Oklahoma State University” be labeled as an organization, but “Oklahoma” not be labeled as a place at the same time. In our dataset, we have many overlapping labels. Therefore, we need to develop a new relaxed-match metric.

We propose a hierarchical relaxed-match metric that can deal with overlapping or nested entity labels. For each type of history element, there are six below mutual exclusive count categories. Codes are provided in the brackets for convenience (S: text span, E: exact matching, P: partial matching, T: entity type, C: correct, I: incorrect).

- Category 1, text span exact matching and entity type correct (SE-TC)

  For an entity A of a certain type in a gold-standard note, there is an entity B in the corresponding prediction note which shares the same text span and entity type with A.

- Category 2, text span partial matching and entity type correct (SP-TC)

  For an entity A of a certain type in a gold-standard note, no Category 1 entity exists in the corresponding prediction note, but there is an entity B in the prediction note that shares a partial matching text span and the same entity type with A.

- Category 3, text span exact matching and entity type incorrect (SE-TI)

  For an entity A of a certain type in a gold-standard note, no Category 1 or 2 entity exists in the corresponding prediction note, but there is an entity B in the prediction note that shares the same text span with A and the entity type of B is incorrect.

- Category 4, text span partial matching and entity type incorrect (SP-TI)

  For an entity A of a certain type in a gold-standard note, no Category 1 or 2 or 3 entity exists in the corresponding prediction note, but there is an entity B in the corresponding prediction note which shares a partial matching text span with A and the entity type of B is incorrect.

- Category 5, Missing (MS).

  For an entity A of a certain type in a gold-standard note, the text span of A in the corresponding prediction note does not contain any entity of any type.

- Category 6, Spurious (SR).

  For an entity B of a certain type in a prediction note, the text span of B in the corresponding gold-standard note does not contain any entity of any type.

Note that the first five categories are hierarchical. For each entity of a certain type in a gold-standard note, we will go through the corresponding prediction note and count it as one of the first five categories following the order from Category 1 to Category 5 while for each entity of that type in a prediction note, we will go through the corresponding gold-standard note and check whether the Category 6 applies. During this process, multiple occurrences within the same category are only counted once for each entity in gold-standard notes.

5.3 Results

5.3.1 Exact-Match Metric Performances

Table 5.3 presents the two systems' exact-match metric performances side-by-side. The relatively low performances from both systems show that information extraction of patient history information for E/M billing is still a challenging problem. In general, CC, HPI timing, and PFSH are better identified while HPI context, HPI modifying factors, and HPI associated signs/symptoms are more difficult to recognize. This phenomenon is consistent with our assumption that some types of billing elements require more complicated inferences than others. The rule-based system performed better on multiple element types in the context of a relatively small labeled corpus. The BERT system might correctly learn a pattern, e.g., CC precision = 0.73, but did not have enough training examples to generalize, e.g., CC recall = 0.10.

Table 5.3: Exact-Match Model Performances

	Rule-based			BERT			
Element Type	Prec.	Recl.	F1	Prec.	Recl.	F1	Sprt.
CC	0.42	0.40	**0.41**	0.73	0.10	0.18	138
Chronic Condition	0.77	0.59	**0.67**	0.51	0.54	0.52	171
HPI-location	0.15	0.43	**0.23**	0.29	0.10	0.15	69
HPI-quality	0.42	0.30	**0.35**	0.17	0.19	0.18	57
HPI-severity	0.27	0.41	**0.32**	0	0	0	27
HPI-duration	0.21	0.30	0.25	0.44	0.36	**0.39**	67
HPI-timing	0.38	0.54	**0.45**	0.25	0.11	0.15	37
HPI-context	0.04	0.03	**0.03**	0.0	0.0	0.0	37
HPI-modifying factors	0.16	0.54	**0.25**	0.18	0.24	0.20	82
HPI-associated signs/symptoms	0.16	0.33	**0.22**	0.13	0.22	0.16	269
Past History	0.54	0.69	**0.61**	0.53	0.71	**0.61**	520
Family History	0.70	0.87	**0.77**	0.63	0.67	0.65	45
Social History	0.12	0.10	0.11	0.33	0.40	**0.36**	129

Note: "Prec." — Precision, "Recl." — Recall, "F1" — F1-score, "Sprt." — Support.

An interesting finding is that the rule-based system in general has a relatively higher recall but lower precision. This indicates that our rule-based system contains some low-precision rules which yield many false positive predictions. This is consistent with our conception that varying precision is a common characteristic of rule-based systems [Michelakis et al., 2009].

On the contrary, the deep learning system, in general, has lower recall rates but its precision rates are relatively higher. The deep learning system is fine-tuned on limited training samples so although it does learn some patterns, its learning capacity is compromised and thus there are many misses during the prediction.

We also compared the performances of the two systems on notes in different size groups. We divided the notes into 3-quantile groups based on their numbers of tokens, numbers of sentences, and numbers of sections, respectively. F1 scores of the two systems in these groups are reported in Tables 5.4, 5.5, and 5.6.

Table 5.4: F1 Scores for Notes with Different Numbers of Tokens

Element Type	Rule-based						BERT					
	Group 1		Group 2		Group 3		Group 1		Group 2		Group 3	
	Sp	F1	Sp	F1	Sp	F1	Sp	F1	Sp	F1	Sp	F1
CC	39	0.38	57	0.45	42	0.43	39	0.14	57	0.22	42	0.16
Chronic Condition	30	0.67	59	0.61	82	0.71	30	0.54	59	0.44	82	0.60
HPI-location	24	0.35	11	0.18	34	0.20	24	0.15	11	0.31	34	0.12
HPI-quality	13	0.43	13	0.20	31	0.36	13	0.25	13	0.26	31	0.17
HPI-severity	9	0.14	9	0.24	9	0.52	9	0.00	9	0.00	9	0.00
HPI-duration	21	0.27	16	0.20	30	0.26	21	0.31	16	0.32	30	0.48
HPI-timing	10	0.70	6	0.23	21	0.45	10	0.17	6	0.25	21	0.13
HPI-context	9	0.15	10	0.00	18	0.00	9	0.00	10	0.00	18	0.00
HPI-modifying factors	7	0.08	31	0.25	44	0.29	7	0.29	31	0.24	44	0.21
HPI-associated signs/symptoms	51	0.23	79	0.21	139	0.21	51	0.19	79	0.17	139	0.19
Past History	85	0.55	206	0.58	229	0.65	85	0.63	206	0.61	229	0.69
Family History	8	0.74	14	0.79	23	0.78	8	0.86	14	0.64	23	0.73
Social History	22	0.04	37	0.19	70	0.10	22	0.30	37	0.48	70	0.41
Average		0.396		0.429		0.430		0.352		0.411		0.417

Note: Group 1 - notes with no more than 557 tokens, Group 2 - notes with token numbers in (557, 823], Group 3 - notes with more than 823 tokens

Table 5.5: F1 Scores for Notes with Different Numbers of Sentences

Element Type	Rule-based						BERT					
	Group 1		Group 2		Group 3		Group 1		Group 2		Group 3	
	Sp	F1	Sp	F1	Sp	F1	Sp	F1	Sp	F1	Sp	F1
CC	42	0.35	46	0.47	50	0.45	42	0.09	46	0.22	50	0.21
Chronic Condition	35	0.67	62	0.67	74	0.67	35	0.60	62	0.57	74	0.49
HPI-location	20	0.20	19	0.29	20	0.20	20	0.18	19	0.15	30	0.14
HPI-quality	11	0.32	22	0.32	24	0.38	11	0.31	22	0.23	24	0.13
HPI-severity	14	0.15	6	0.24	7	0.64	14	0.00	6	0.00	7	0.00
HPI-duration	21	0.24	25	0.18	21	0.33	21	0.35	25	0.40	21	0.43
HPI-timing	9	0.37	10	0.62	18	0.39	9	0.00	10	0.37	18	0.09
HPI-context	12	0.11	11	0.00	14	0.00	12	0.00	11	0.00	14	0.00
HPI-modifying factors	20	0.17	32	0.35	30	0.19	20	0.28	32	0.22	30	0.20
HPI-associated signs/symptoms	62	0.18	87	0.23	120	0.21	62	0.18	87	0.21	120	0.17
Past History	88	0.59	165	0.59	267	0.62	88	0.61	165	0.62	267	0.68
Family History	9	0.70	13	0.79	23	0.79	9	0.80	13	0.67	23	0.73
Social History	24	0.08	36	0.09	69	0.14	24	0.21	36	0.46	69	0.45
Average		0.360		0.431		0.445		0.332		0.412		0.429

Note: Group 1 - notes with no more than 50 sentences, Group 2 - notes with sentence numbers in (50, 72], Group 3 - notes with more than 72 sentences

Table 5.6: F1 Scores for Notes with Different Numbers of Sections

Element Type	Rule-based						BERT					
	Group 1		Group 2		Group 3		Group 1		Group 2		Group 3	
	Sp	F1	Sp	F1	Sp	F1	Sp	F1	Sp	F1	Sp	F1
CC	49	0.43	46	0.48	43	0.35	49	0.15	46	0.11	43	0.27
Chronic Condition	42	0.70	58	0.56	71	0.72	42	0.53	58	0.58	71	0.52
HPI-location	23	0.19	26	0.21	20	0.28	23	0.22	26	0.10	20	0.16
HPI-quality	26	0.35	19	0.35	12	0.33	26	0.33	19	0.18	12	0.00
HPI-severity	16	0.22	6	0.53	5	0.40	16	0.00	6	0.00	5	0.00
HPI-duration	25	0.15	32	0.38	10	0.19	25	0.37	32	0.44	10	0.25
HPI-timing	15	0.41	14	0.48	8	0.47	15	0.21	14	0.09	8	0.22
HPI-context	13	0.10	10	0.00	14	0.00	13	0.00	10	0.00	14	0.00
HPI-modifying factors	42	0.27	27	0.27	13	0.16	42	0.24	27	0.28	13	0.07
HPI-associated signs/symptoms	105	0.21	78	0.16	86	0.27	105	0.20	78	0.23	86	0.13
Past History	127	0.50	156	0.62	237	0.64	127	0.54	156	0.67	237	0.69
Family History	8	0.56	21	0.93	16	0.70	8	0.46	21	0.80	16	0.73
Social History	28	0.07	42	0.19	59	0.08	28	0.15	42	0.42	59	0.52
Average		0.351		0.435		0.465		0.314		0.420		0.461

Note: Group 1 - notes with no more than 10 sections, Group 2 - notes with section numbers in (10, 12], Group 3 - notes with more than 12 sections

If we focus on the averaged F1 scores, the above three tables can be summarized in Figure 5.5. We identify two interesting findings. First, as the number of tokens, sentences, or sections in a note group increases, the averaged F1 scores of both systems in the group increase. This indicates that both systems perform better if the notes are lengthier and contain more sections. This is consistent with our assumption that short notes are generally more heterogeneous, lack of contextual information for inference, and thus more challenging for text analysis. Another finding is that the averaged F1 score gap between the two systems narrows as notes become lengthier or contain more sections. In other words, the BERT system is able to exploit contextual information and catches up quickly with the rule-based system in terms of averaged F1 score when more information exists in notes.

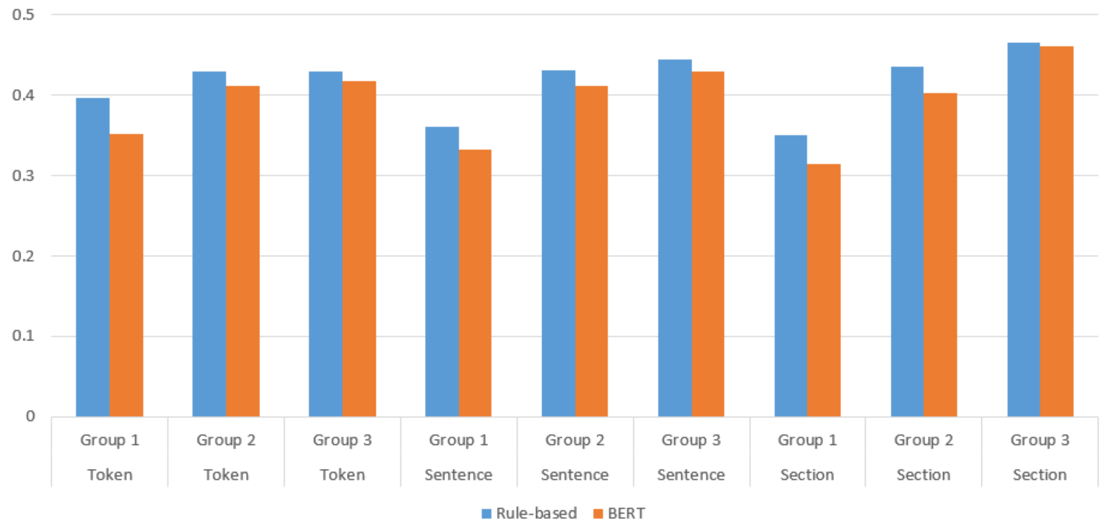


Figure 5.5: Average F1 Scores across Note Groups

5.3.2 Relaxed-Match Metric Performances

Table 5.7 shows the relaxed-match metric performances. Note that the values of the six categories are not counts but are calculated in ratios. To be specific, the first five categories are calculated by dividing each category count by the number of labels of each element type

Table 5.7: Relaxed-Match Model Performances

Element Type	Labels	Rule-based (ratios)							BERT (ratios)						
		Preds	SE-TC	SP-TC	SE-TI	SP-TI	MS	SR	Preds	SE-TC	SP-TC	SE-TI	SP-TI	MS	SR
CC	138	127	0.41	0.12	0.17	0.15	0.16	0.17	19	0.11	0.01	0.22	0.17	0.49	0.05
Chronic Condition	171	130	0.57	0.03	0.26	0.08	0.06	0.10	167	0.52	0.06	0.12	0.06	0.24	0.17
HPI-location	68	196	0.21	0.54	0.00	0.04	0.21	0.22	23	0.04	0.12	0.06	0.21	0.57	0.39
HPI-quality	56	40	0.25	0.13	0.13	0.13	0.38	0.13	46	0.20	0.11	0.07	0.13	0.50	0.28
HPI-severity	27	40	0.22	0.37	0.00	0.11	0.30	0.45	2	0.00	0.07	0.04	0.26	0.63	0.00
HPI-duration	67	95	0.28	0.34	0.00	0.06	0.31	0.43	52	0.33	0.15	0.01	0.03	0.48	0.31
HPI-timing	36	52	0.28	0.31	0.03	0.14	0.25	0.38	12	0.03	0.14	0.17	0.31	0.36	0.25
HPI-context	37	27	0.03	0.30	0.05	0.43	0.19	0.26	27	0.00	0.22	0.05	0.41	0.32	0.52
HPI-modifying factors	82	267	0.54	0.29	0.00	0.04	0.13	0.35	86	0.24	0.11	0.15	0.10	0.40	0.40
HPI-associated signs/symptoms	269	522	0.30	0.29	0.04	0.11	0.26	0.38	334	0.20	0.19	0.01	0.07	0.53	0.44
Past History	520	640	0.68	0.10	0.12	0.04	0.06	0.34	610	0.70	0.07	0.02	0.02	0.19	0.24
Family History	45	54	0.87	0.04	0.02	0.07	0.00	0.22	38	0.67	0.09	0.11	0.00	0.13	0.11
Social History	129	103	0.10	0.64	0.01	0.05	0.19	0.09	120	0.39	0.38	0.02	0.04	0.18	0.18

Note: "Labels" — the number of gold-standard labels, "Preds" — the number of predicted labels, cell values of the first five categories — counts divided by the number of gold-standard labels, cell values of the sixth category — counts divided by the number of predicted labels

in gold-standard notes while the sixth category is calculated by dividing the count of the sixth category by the number of predictions of that element type in prediction notes.

An interesting finding is about Category 2 (SP-TC). In relaxed-match metrics, Category 2 predictions are generally considered as partially correct. If we add Category 1 and Category 2 together (Figure 5.6), the percentage of "correct" entities of the two systems will be further higher, highlighting the potential of our NLP systems in computer-assisted medical billing.

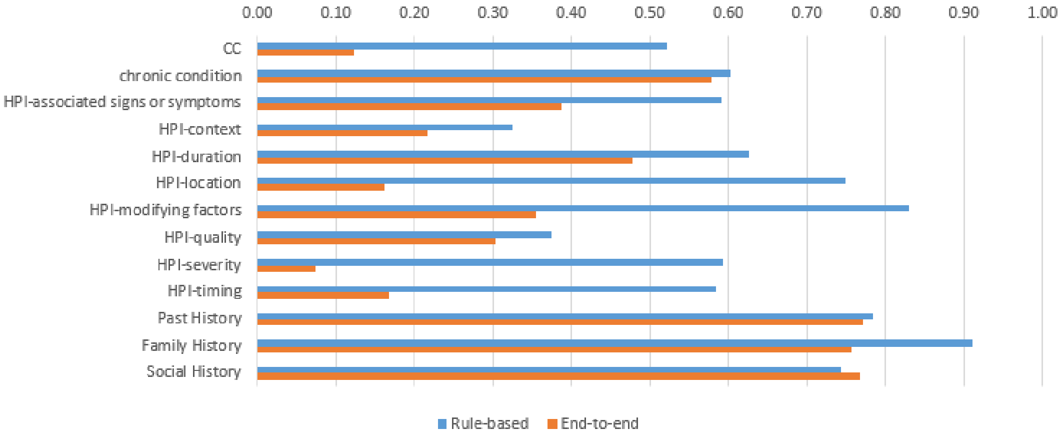


Figure 5.6: Percentages of Correct Entities in Gold-standard Notes

5.3.3 Select Example Analysis

Here we select some example outputs of the two systems and examine them intuitively.

In Example 1 below (G: gold-standard, R: rule-based model, B: BERT model), for the rule-based model prediction, the CC entity type has three Category 1 counts, the Chronic Condition type has one Category 1 count (*[diabetes mellitus]*) and one Category 3 count (*[hypercholesterolemia]*); for the BERT model prediction, the CC entity type has three Category 1 counts, the Chronic Condition type has two Category 3 counts (*[diabetes mellitus]* and *[hypercholesterolemia]*), the Past History type has one Category 6 count (*[Followup]*).

Example 1:
(G) CHIEF COMPLAINT: Followup on *[diabetes mellitus]*$_{CC,chronicCondition}$, *[hypercholesterolemia]*$_{CC,chronicCondition}$, and *[sinusitis]*$_{CC}$.
(R) CHIEF COMPLAINT: Followup on *[diabetes mellitus]*$_{CC,chronicCondition}$, *[hypercholesterolemia]*$_{CC}$, and *[sinusitis]*$_{CC}$.
(B) CHIEF COMPLAINT: *[Followup]*$_{pastHistory}$ on *[diabetes mellitus]*$_{CC}$, *[hypercholesterolemia]*$_{CC}$, and *[sinusitis]*$_{CC}$.

Similarly, in Example 2, for the rule-based model prediction, the HPI-Associated Signs and Symptoms entity type has one Category 2 count (either *[sinus]* or *[congestion]* or *[drainage]*), and the HPI-Duration type has one Category 1 count (*[several days]*); for the end-to-end BERT model prediction, the HPI-Associated Signs and Symptoms entity type has one Category 2 count (either *[sinus congestion]* or *[drainage]*), and the HPI-Duration type has one Category 2 count (*[last several days]*).

Example 2:
(G) She does complain of some *[sinus congestion and drainage]*$_{hpi.assocSignsAndSymptoms}$ for the last *[several days]*$_{hpi.duration}$.
(R) She does complain of some *[sinus]*$_{hpi.assocSignsAndSymptoms,hpi.location}$ *[congestion]*$_{hpi.assocSignsAndSymptoms}$ and *[drainage]*$_{hpi.assocSignsAndSymptoms}$ for the last *[several days]*$_{hpi.duration}$.
(B) She does complain of some *[sinus congestion]*$_{hpi.assocSignsAndSymptoms}$ and *[drainage]*$_{hpi.assocSignsAndSymptoms}$ for the *[last several days]*$_{hpi.duration}$.

In Example 3, most history information is contained in only one section, making extraction even more challenging. For example, BERT did a worse job in identifying *CC* entities although a strong hint ("complains of") exists.

Example 3:

(G) SUBJECTIVE: The patient is in with several medical problems. She complains of *[numbness]*$_{CC}$, *[tingling]*$_{CC}$, and a *[pain]*$_{CC}$ in the *[toes primarily of her right foot]*$_{hpi.location}$ described as a *[moderate]*$_{hpi.severity}$ pain. She initially describes it as a *[sharp quality pain]*$_{hpi.quality}$, but is unable to characterize it more fully. She has had it for about *[a year]*$_{hpi.duration}$, but seems to be *[worsening]*$_{hpi.quality}$. She has *[little bit]*$_{hpi.quality}$ of *[paraesthesias]*$_{hpi.assocSignsAndSymptoms}$ in the *[left toe]*$_{hpi.location}$ as well and seem to involve *[all the toes of the right foot]*$_{hpi.location}$. They are *[not worse with walking]*$_{hpi.modifyingFactors}$.

(R) SUBJECTIVE: The patient is in with *[[several medical]*$_{hpi.modifyingFactors}$ *problems]*$_{hpi.assocSignsAndSymptoms}$. She complains of *[numbness]*$_{CC}$, *[tingling]*$_{CC}$, and a *[pain in the [toes]*$_{hpi.location}$*]*$_{CC}$ primarily of her *[right foot]*$_{hpi.location}$ described as a *[[moderate]*$_{hpi.severity}$ *pain]*$_{CC}$. She initially describes it as a *[sharp quality pain]*$_{hpi.assocSignsAndSymptoms}$, but is *[unable]*$_{hpi.assocSignsAndSymptoms}$ to characterize it more fully. She has had it for about a year, but seems to be *[worsening]*$_{hpi.quality,hpi.assocSignsAndSymptoms}$. She has little bit of *[paraesthesias]*$_{hpi.assocSignsAndSymptoms}$ in the *[left toe]*$_{hpi.location}$ as well and seem to involve all the toes of the right foot. They are not *[[worse]*$_{hpi.quality}$ *with walking]*$_{hpi.modifyingFactors}$.

(B) SUBJECTIVE: The patient is in with several medical problems. She complains of *[numbness]*$_{hpi.assocSignsAndSymptoms}$, *[tingling]*$_{hpi.assocSignsAndSymptoms}$, and a *[pain]*$_{hpi.assocSignsAndSymptoms}$ in the *[toes primarily of her right foot]*$_{hpi.location}$ described as a *[moderate]*$_{hpi.severity}$ pain. She initially describes it as a *[sharp quality pain]*$_{hpi.quality}$, but is unable to characterize it more fully. She has had it for about *[a year]*$_{hpi.duration}$, but seems to be *[worsening]*$_{hpi.quality}$. She has *[little]*$_{hpi.assocSignsAndSymptoms}$ bit of *[paraesthesias]*$_{hpi.assocSignsAndSymptoms}$ in the left toe as well and seem to involve all the toes of the right foot. They are *[not worse]*$_{hpi.quality}$ *[with walking]*$_{hpi.assocSignsAndSymptoms}$.

In Example 4, most history information is also contained in only one section but the BERT system successfully identified *CC* entities. The hint "complaint of" is similar to the hint in Example 3. This indicates the difficulty to interpret deep learning models.

Example 4:

(G) HISTORY OF PRESENT ILLNESS: A 49-year-old female with history of *[atopic dermatitis]*$_{pastHistory}$ comes to the clinic with complaint of *[left otalgia]*$_{CC}$ and *[headache]*$_{CC}$. Symptoms started approximately *[three weeks ago]*$_{hpi.duration}$ and she was having *[difficulty hearing]*$_{hpi.assocSignsAndSymptoms}$, although that has *[greatly improved]*$_{hpi.quality}$. She is having some *[left-sided sinus pressure]*$_{hpi.assocSignsAndSymptoms}$ and actually went to the dentist because her *[teeth were hurting; however, the teeth were okay]*$_{hpi.assocSignsAndSymptoms}$. She continues to have some *[[left-sided jaw]*$_{hpi.location}$ *pain]*$_{hpi.assocSignsAndSymptoms}$.

(R) HISTORY OF PRESENT *[ILLNESS]*$_{pastHistory}$: A 49-year-old female with history of *[atopic dermatitis]*$_{pastHistory,hpi.assocSignsAndSymptoms}$ comes to the clinic with *[complaint]*$_{pastHistory}$ of left *[otalgia]*$_{CC}$ and *[headache]*$_{CC}$. Symptoms started approximately three *[weeks ago]*$_{hpi.duration}$ and she was having *[difficulty hearing]*$_{hpi.assocSignsAndSymptoms}$, although that has greatly *[improved]*$_{hpi.quality}$. She is having some *[[left-sided sinus]*$_{hpi.location}$ *pressure]*$_{hpi.assocSignsAndSymptoms}$ and actually went to the dentist because her teeth were hurting; however, the teeth were okay. She continues to have some *[[left-sided jaw]*$_{hpi.location}$ *pain]*$_{hpi.assocSignsAndSymptoms}$.

(B) HISTORY OF PRESENT ILLNESS: A 49-year-old female with history of *[atopic dermatitis]*$_{pastHistory,chronicCondition}$ comes to the clinic with complaint of *[[left]*$_{hpi.location}$ *otalgia]*$_{CC}$ and *[headache]*$_{CC}$. Symptoms started approximately *[three weeks ago]*$_{hpi.duration}$ and she was having *[difficulty hearing]*$_{hpi.assocSignsAndSymptoms}$, although that has *[greatly improved]*$_{hpi.quality}$. She is having some *[left-sided sinus pressure]*$_{hpi.assocSignsAndSymptoms}$ and actually went to the dentist because her *[teeth were hurting]*$_{hpi.assocSignsAndSymptoms}$; however, the teeth *[were okay]*$_{hpi.assocSignsAndSymptoms}$. She continues to have some *[[left-sided jaw pain]*$_{hpi.assocSignsAndSymptoms}$.

5.4 Conclusion

Medical billing is a major challenge in the healthcare revenue management cycle in the United States. This study is the first in the academic community on extracting patient history information directly from clinical notes to facilitate E/M billing. It proposes a framework and develops two prototype systems – a rule-based and a deep-learning-based.

The two prototype systems developed in this study meet our expectations in their capacities to extract essential patient history information. On average, extraction performances are better for such elements as PFSH, CC, Chronic Condition, and HPI-Duration. Performances are less satisfactory for elements including HPI-Context, HPI-Quality, and HPI-Timing. This is generally consistent with our assumption that some elements need more semantic reasoning which unfortunately is still technically challenging at this moment.

Another interesting finding is that performances of both systems improve as note size increases which indicates rich notes should be used as input data for medical billing. In addition, the deep learning BERT model may have more potential in terms of performance improvements and portability if we train it on a larger dataset.

In short, the proposed framework and the two prototype systems exhibit promising values. We summarize the contributions and limitations of this study as follows.

Methodological Contributions

- This study proposes the first framework in the academic community to use publicly or academically available resources to extract patient history information to facilitate E/M medical billing. A rule-based and a deep learning systems are constructed and tested on an annotated dataset.
- It also proposes comprehensive metrics to evaluate NER performances including the exact-match metric and the relaxed-match metric which is a novel hierarchical metric suitable for quantifying the NER outcomes for notes with text span overlapping entities.

Application Contributions

- This study demonstrates the application potential and feasibility to reduce medical billing costs by developing clinical information extraction systems.
- It also introduces clinical NLP resources currently available, discusses the challenges needed to be addressed, and provides some technical solutions (e.g., libraries and knowledge extraction rules).

Limitations

- The annotated dataset for this study is small which limits the knowledge discovery capacities of both rule-based and deep learning systems. A larger dataset may significantly improve model performances.
- The rule-based system has many assumptions and may not generalize well to other datasets. For example, one assumption is that notes have explicit section headings while many raw clinical notes do not have such headings. Another assumption is that information is documented in appropriate sections which may not be true.
- The deep learning model is built on a pre-trained BERT model using the transfer learning approach, which does not guarantee to have similar performances with newly trained models. A comparison between transfer learning and newly trained BERT models may be necessary.

Future work includes annotating a larger dataset for model training and testing, comparing more cutting-edge deep learning algorithms, comparing transfer learning and newly trained deep learning models, and validating the systems in real-world clinical settings.

CHAPTER VI

CONCLUSION

In this dissertation, we focus on building machine learning models and systems to analyze structured EHR data or unstructured clinical notes with a goal to improve health care quality, reduce healthcare costs, and reduce healthcare disparities.

The three research topics include analyzing longitudinal structured EHR data for diabetic retinopathy prediction, constructing and clustering temporal disease networks to better visualize comorbidity progression, and designing systems to extract patient history information to facilitate E/M medical billing.

For the first topic, we studied temporal analysis methods for diabetic retinopathy prediction, constructed both deep learning temporal models and non-temporal random forests models, and evaluated the models on a large-scale dataset. The dataset is extracted from one of the largest real-world EHR databases in America, containing patient demographics, lab tests, and comorbidity variables. Experimental results show that deep learning temporal models outperformed non-temporal random forest models in terms of AUPRC and Recall.

At the methodological level, to the best of our knowledge, this is the first study that implements deep learning temporal models to analyze longitudinal EHR data for DR prediction. The deep learning architectures also have a multi-branching output mechanism to address the imbalanced dataset issue. At the application level, the study shows that developing a temporal DR prediction model using widely available longitudinal EHR data may be a better alternative to assist current DR screening, which potentially can help curb the

DR prevalence.

Future work for this study includes examining more temporal (e.g., the knowledge-based temporal abstraction approach) and baseline models, incorporating more data representation techniques, and performing a more detailed study on the association between model performance and the number of multi-branching outputs.

For the second topic, we designed a method to construct disease networks, proposed a consecutive p-median clustering method to group temporal disease networks into phases, and simplified the visualization using a disease clustering method. Two case studies on C.Diff and stroke demonstrated that the methods are effective.

At the methodological level, contributions of this study include the consecutive p-median clustering and the disease atomic clustering methods. At the application level, the proposed framework can be applied in real-world clinical settings to visualize comorbidity progression to improve clinical decision-making.

Future work includes examining more network dissimilarity measurements (which may influence the temporal clustering results) and testing the methods on more complex disease networks (e.g., directed or weighted networks which carry more clinical information than the undirected and unweighted networks in our current study).

For the third topic, we proposed a framework to extract patient history information directly from clinical notes to facilitate E/M medical billing. Two approaches and their corresponding prototype systems, one rule-based and one deep-learning-based, were developed. The two prototype systems show that extraction of patient history information is still a technically challenging task, but computer-assisted medical billing has promising potentials.

At the methodological level, to the best of our knowledge, it is the first framework proposed for E/M medical billing in the academic community. The rule-based and the BERT transfer learning architectures represent the two major technique solutions. A comprehensive evaluation metric including a novel hierarchical relaxed-match metric is proposed to quantify named entity recognition outcomes on notes with text span overlapping entities. At the application level, the framework and technical solutions proposed in this study can be applied in real-world billing practice to relieve the billing burden.

Future work of this study includes annotating a larger dataset for model training and testing, comparing more cutting-edge deep learning algorithms, and improving and validating the systems in real-world clinical settings.

REFERENCES

[Abràmoff et al., 2018] Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N., and Folk, J. C. (2018). Pivotal trial of an autonomous ai-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ digital medicine*, 1(1):1–8.

[AHRQ, 2020] AHRQ (2020). HCUP Tools and Software. Healthcare Cost and Utilization Project (HCUP). `https://www.hcup-us.ahrq.gov/tools_software.jsp`.

[Allen, 1983] Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

[Alsentzer et al., 2019] Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., and McDermott, M. (2019). Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA.

[AMA, 2019] AMA (2019). CPT Evaluation and Management (E/M) Office or Other Outpatient and Prolonged Services Code and Guideline Changes. `https://www.ama-assn.org/system/files/2019-06/cpt-office-prolonged-svs-code-changes.pdf`.

[Arndt et al., 2017] Arndt, B. G., Beasley, J. W., Watkinson, M. D., Temte, J. L., Tuan, W.-J., Sinsky, C. A., and Gilchrist, V. J. (2017). Tethered to the ehr: Primary care physician workload assessment using ehr event log data and time-motion observations. *The Annals of Family Medicine*, 15(5):419–426.

[Aronson and Lang, 2010] Aronson, A. R. and Lang, F.-M. (2010). An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

[Assale et al., 2019] Assale, M., Dui, L. G., Cina, A., Seveso, A., and Cabitza, F. (2019). The revival of the notes field: Leveraging the unstructured content in electronic health records. *Frontiers in Medicine*, 6(66).

[Bai et al., 2018] Bai, S., Kolter, J. Z., and Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

[Barabási, 2007] Barabási, A.-L. (2007). Network medicine — from obesity to the "diseasome". *The New England Journal of Medicine*, 357(4):404–407.

[Barabási et al., 2011] Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12:56–68.

[Bardhan and Thouin, 2013] Bardhan, I. R. and Thouin, M. F. (2013). Health information technology and its impact on the quality and cost of healthcare delivery. *Decision Support Systems*, 55(2):438–449.

[Barjis et al., 2013] Barjis, J., Kolfschoten, G., and Maritz, J. (2013). A sustainable and affordable support system for rural healthcare delivery. *Decision Support Systems*, 56:223–233.

[Bauer et al., 2012] Bauer, M. P., Hensgens, M. P., Miller, M. A., Gerding, D. N., Wilcox, M. H., Dale, A. P., Fawley, W. N., Kuijper, E. J., and Gorbach, S. L. (2012). Renal failure and leukocytosis are predictors of a complicated course of clostridium difficile infection if measured on day of diagnosis. *Clinical Infectious Diseases*, 55(suppl_2):S149–S153.

[Berlingerio et al., 2013] Berlingerio, M., Koutra, D., Eliassi-Rad, T., and Faloutsos, C. (2013). Network similarity via multiple social theories. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 1439–1440.

[Bhaskaranand et al., 2019] Bhaskaranand, M., Ramachandra, C., Bhat, S., Cuadros, J., Nittala, M. G., Sadda, S. R., and Solanki, K. (2019). The value of automated diabetic retinopathy screening with the eyeart system: a study of more than 100,000 consecutive encounters from people with diabetes. *Diabetes technology & therapeutics*, 21(11):635–643.

[Bodenreider, 2004] Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

[Bradley et al., 2006] Bradley, A., Duin, R., Paclik, P., and Landgrebe, T. (2006). Precision-recall operating characteristic (p-roc) curves in imprecise environments. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 4, pages 123–127. IEEE.

[Bright et al., 2012] Bright, T. J., Wong, A., Dhurjati, R., Bristow, E., Bastian, L., Coeytaux, R. R., Samsa, G., Hasselblad, V., Williams, J. W., Musty, M. D., Wing, L., Kendrick, A., Sanders, G., and Lobach, D. (2012). Effect of clinical decision-support systems: a systematic review. *Annals of Internal Medicine*, 157(1):29–43.

[Brown and Kachura, 2019] Brown, A. and Kachura, J. (2019). Natural language processing of radiology reports in patients with hepatocellular carcinoma to predict radiology resource utilization. *Journal of the American College of Radiology*, 16(6):840–844.

[Brunson and Laubenbacher, 2018] Brunson, J. C. and Laubenbacher, R. C. (2018). Applications of network analysis to routinely collected health care data: a systematic review. *Journal of the American Medical Informatics Association*, 25(2):210–221.

[Burger et al., 2016] Burger, G., Abu-Hanna, A., de Keizer, N., and Cornet, R. (2016). Natural language processing in pathology: a scoping review. *Journal of clinical pathology*, 69(11):949–955.

[Caban and Gotz, 2015] Caban, J. J. and Gotz, D. (2015). Visual analytics in healthcare – opportunities and research challenges. *Journal of the American Medical Informatics Association*, 22(2):260–262.

[Candrilli et al., 2007] Candrilli, S. D., Davis, K. L., Kan, H. J., Lucero, M. A., and Rousculp, M. D. (2007). Prevalence and the associated burden of illness of symptoms of diabetic peripheral neuropathy and diabetic retinopathy. *Journal of Diabetes and its Complications*, 21(5):306–314.

[Capobianco and Lio, 2013] Capobianco, E. and Lio, P. (2013). Comorbidity: a multidimensional approach. *Trends in Molecular Medicine*, 19(9):515–521.

[Catling and Wolff, 2020] Catling, F. J. and Wolff, A. H. (2020). Temporal convolutional networks allow early prediction of events in critical care. *Journal of the American Medical Informatics Association*, 27(3):355–365.

[CDC, 2019] CDC (2019). Antibiotic resistance threats in the United States. `https://www.cdc.gov/drugresistance/pdf/threats-report/2019-ar-threats-report-508.pdf`.

[CDC, 2021] CDC (2021). Diabetes and vision loss. `https://www.cdc.gov/diabetes/managing/diabetes-vision-loss.html`.

[Champagnie, 2019] Champagnie, S. J. (2019). Medicare Loses Billions to Billing Errors. *SSRN Electronic Journal*.

[Chapman et al., 2001] Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., and Buchanan, B. G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310.

[Chawla, 2009] Chawla, N. V. (2009). Data mining for imbalanced datasets: An overview. *Data mining and knowledge discovery handbook*, pages 875–886.

[Chen et al., 2009] Chen, L., Blumm, N., Christakis, N., Barabási, A., and Deisboeck, T. (2009). Cancer metastasis networks and the prediction of progression patterns. *British Journal of Cancer*, 101(5):749–758.

[Chen et al., 2015] Chen, Y., Zhang, X., Zhang, G., and Xu, R. (2015). Comparative analysis of a novel disease phenotype network based on clinical manifestations. *Journal of Biomedical Informatics*, 53:113–120.

[Chinchor and Sundheim, 1993] Chinchor, N. and Sundheim, B. M. (1993). Muc-5 evaluation metrics. In *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993.*

[Chiticariu et al., 2013] Chiticariu, L., Li, Y., and Reiss, F. (2013). Rule-based information extraction is dead! long live rule-based information extraction systems! In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 827–832.

[Chmiel et al., 2014] Chmiel, A., Klimek, P., and Thurner, S. (2014). Spreading of diseases through comorbidity networks across life and gender. *New Journal of Physics*, 16(11):115013.

[Cho et al., 2014] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078.*

[Choi et al., 2016] Choi, E., Bahadori, M. T., Kulas, J. A., Schuetz, A., Stewart, W. F., and Sun, J. (2016). Retain: an interpretable predictive model for healthcare using reverse time attention mechanism. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 3512–3520.

[CMS, 1997] CMS (1997). 1997 documentation guidelines for evaluation and management services. https://www.cms.gov/outreach-and-education/medicare-learning-network-mln/mlnedwebguide/downloads/97docguidelines.pdf.

[CMS, 2020] CMS (2020). Evaluation and management services guide. https://www.cms.gov/outreach-and-education/medicare-learning-network-mln/mlnproducts/downloads/eval-mgmt-serv-guide-icn006764.pdf.

[Collier and Takeuchi, 2004] Collier, N. and Takeuchi, K. (2004). Comparison of character-level and part of speech features for name recognition in biomedical texts. *Journal of Biomedical Informatics*, 37(6):423–435.

[Combi and Shahar, 1997] Combi, C. and Shahar, Y. (1997). Temporal reasoning and temporal data maintenance in medicine: issues and challenges. *Computers in biology and medicine*, 27(5):353–368.

[Cramer et al., 2010] Cramer, A., Waldorp, L. J., van der Maas, H. L. J., and Borsboom, D. (2010). Comorbidity: A network perspective. *Behavioral and Brain Sciences*, 33:137–193.

[Cruickshanks et al., 1993] Cruickshanks, K. J., Ritter, L. L., Klein, R., and Moss, S. E. (1993). The association of microalbuminuria with diabetic retinopathy: the wisconsin epidemiologic study of diabetic retinopathy. *Ophthalmology*, 100(6):862–867.

[Davazdahemami and Delen, 2018] Davazdahemami, B. and Delen, D. (2018). A chronological pharmacovigilance network analytics approach for predicting adverse drug events. *Journal of the American Medical Informatics Association*, 25(10):1311–1321.

[De Groot et al., 2003] De Groot, V., Beckerman, H., Lankhorst, G. J., and Bouter, L. M. (2003). How to measure comorbidity: a critical review of available methods. *Journal of Clinical Epidemiology*, 56(3):221–229.

[Denecke, 2014] Denecke, K. (2014). Extracting medical concepts from medical social media with clinical nlp tools: a qualitative study. In *Proceedings of the fourth workshop on building and evaluation resources for health and biomedical text processing*.

[Denny et al., 2010] Denny, J. C., Ritchie, M. D., Basford, M. A., Pulley, J. M., Bastarache, L., Brown-Gentry, K., Wang, D., Masys, D. R., Roden, D. M., and Crawford, D. C. (2010). Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics*, 26(9):1205–1210.

[Divo et al., 2015] Divo, M. J., Casanova, C., Marin, J. M., Pinto-Plata, V. M., de Torres, J. P., Zulueta, J. J., Cabrera, C., Zagaceta, J., Sanchez-Salcedo, P., Berto, J., Davila, R. B., Alcaide, A. B., Cote, C., and Celli, B. R. (2015). Copd comorbidities network. *European Respiratory Journal*, 46:640–650.

[Divo et al., 2014] Divo, M. J., Martinez, C. H., and Mannino, D. M. (2014). Ageing and the epidemiology of multimorbidity. *European Respiratory Journal*, 44(4):1055–1068.

[Doshi et al., 2018] Doshi, R., Desai, J., Shah, Y., Decter, D., and Doshi, S. (2018). Incidence, features, in-hospital outcomes and predictors of in-hospital mortality associated with toxic megacolon hospitalizations in the United States. *Internal and Emergency Medicine*, 13(6):881–887.

[Du et al., 2019] Du, Z., Li, W., Huang, D., and Wang, Y. (2019). Encoding visual behaviors with attentive temporal convolution for depression prediction. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–7. IEEE.

[Feinstein, 1970] Feinstein, A. R. (1970). The pre-therapeutic classification of co-morbidity in chronic disease. *Journal of Chronic Diseases*, 23(7):455–468.

[Fichman et al., 2011] Fichman, R. G., Kohli, R., and Krishnan, R. (2011). Editorial overview—the role of information systems in healthcare: current research and future trends. *Information Systems Research*, 22(3):419–428.

[Ford et al., 2016] Ford, E., Carroll, J. A., Smith, H. E., Scott, D., and Cassell, J. A. (2016). Extracting information from the text of electronic medical records to improve case detection: a systematic review. *Journal of the American Medical Informatics Association*, 23(5):1007–1015.

[Fotouhi et al., 2018] Fotouhi, B., Momeni, N., Riolo, M. A., and Buckeridge, D. L. (2018). Statistical methods for constructing disease comorbidity networks from longitudinal inpatient data. *Applied Network Science*, 3:46.

[Friedman et al., 2004] Friedman, C., Shagina, L., Lussier, Y., and Hripcsak, G. (2004). Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association*, 11(5):392–402.

[Garla and Brandt, 2012] Garla, V. N. and Brandt, C. (2012). Ontology-guided feature engineering for clinical text classification. *Journal of biomedical informatics*, 45(5):992–998.

[Gevarter, 1984] Gevarter, W. B. (1984). Expert systems: Artificial intelligence applied. *Telematics and Informatics*, 1(3):239–251.

[Ghosh et al., 2018] Ghosh, S., Das, N., Gonçalves, T., Quaresma, P., and Kundu, M. (2018). The journey of graph kernels through two decades. *Computer Science Review*, 27:88–111.

[Gibson, 2015] Gibson, D. M. (2015). The geographic distribution of eye care providers in the united states: implications for a national strategy to improve vision health. *Preventive medicine*, 73:30–36.

[Gijsen et al., 2001] Gijsen, R., Hoeymans, N., Schellevis, F. G., Ruwaard, D., Satariano, W. A., and van den Bos, G. A. M. (2001). Causes and consequences of comorbidity: A review. *Journal of Clinical Epidemiology*, 54(7):661–674.

[Goldberg, 2016] Goldberg, Y. (2016). A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420.

[Gottlieb et al., 2018] Gottlieb, J. D., Shapiro, A. H., and Dunn, A. (2018). The complexity of billing and paying for physician care. *Health Affairs*, 37(4):619–626.

[Grout et al., 2018] Grout, R. W., Cheng, E. R., Carroll, A. E., Bauer, N. S., and Downs, S. M. (2018). A six-year repeated evaluation of computerized clinical decision support system user acceptability. *International Journal of Medical Informatics*, 112:74–81.

[Gulshan et al., 2016] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410.

[Guo et al., 2019] Guo, M., Yu, Y., Wen, T., Zhang, X., Liu, B., Zhang, J., Zhang, R., Zhang, Y., and Zhou, X. (2019). Analysis of disease comorbidity patterns in a large-scale china population. *BMC Medical Genomics*, 12(Suppl 12):177.

[Gupta et al., 2020] Gupta, A., Liu, T., and Crick, C. (2020). Utilizing time series data embedded in electronic health records to develop continuous mortality risk prediction models using hidden markov models: A sepsis case study. *Statistical methods in medical research*, 29(11):3409–3423.

[Gupta and Sharda, 2013] Gupta, A. and Sharda, R. (2013). Improving the science of healthcare delivery and informatics using modeling approaches. *Decision Support Systems*, 2(55):423–427.

[Gurobi Optimization, LLC, 2020] Gurobi Optimization, LLC (2020). Gurobi optimizer reference manual. http://www.gurobi.com.

[Hartman et al., 2022] Hartman, M., Martin, A. B., Washington, B., Catlin, A., Team, N. H. E. A., et al. (2022). National health care spending in 2020: Growth driven by federal spending in response to the covid-19 pandemic: National health expenditures study examines us health care spending in 2020. *Health Affairs*, pages 10–1377.

[HealthAPT, 2022] HealthAPT (2022). Condition categories. https://www2.ccwdata.org/web/guest/condition-categories.

[Henry et al., 2016] Henry, J., Pylypchuk, Y., Searcy, T., and Patel, V. (2016). Adoption of electronic health record systems among us non-federal acute care hospitals: 2008–2015. *ONC Data Brief*, 35:1–9.

[Hidalgo et al., 2009] Hidalgo, C. A., Blumm, N., Barabási, A.-L., and Christakis, N. A. (2009). A dynamic network approach for the study of human phenotypes. *PLoS Computational Biology*, 5(4):1–11.

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

[Hossain et al., 2020] Hossain, M. E., Uddin, S., Khan, A., and Moni, M. A. (2020). A framework to understand the progression of cardiovascular disease for type 2 diabetes mellitus patients using a network approach. *International Journal of Environmental Research and Public Health*, 17(2):596.

[Huang et al., 2011] Huang, M., Névéol, A., and Lu, Z. (2011). Recommending mesh terms for annotating biomedical articles. *Journal of the American Medical Informatics Association*, 18(5):660–667.

[Huerta et al., 2013] Huerta, T. R., Thompson, M. A., Ford, E. W., and Ford, W. F. (2013). Electronic health record implementation and hospitals' total factor productivity. *Decision Support Systems*, 55(2):450–458.

[Ichihara et al., 2015] Ichihara, M., Komiya, K., Iwakura, T., and Yamazaki, M. (2015). Error analysis of named entity recognition in bccwj. *Recall*, 61:2641.

[Irace et al., 2011] Irace, C., Scarinci, F., Scorcia, V., Bruzzichessi, D., Fiorentino, R., Randazzo, G., Scorcia, G., and Gnasso, A. (2011). Association among low whole blood viscosity, haematocrit, haemoglobin and diabetic retinopathy in subjects with type 2 diabetes. *British Journal of Ophthalmology*, 95(1):94–98.

[Jensen et al., 2014] Jensen, A. B., Moseley, P. L., Oprea, T. I., Ellesøe, S. G., Eriksson, R., Schmock, H., Jensen, P. B., Jensen, L. J., and Brunak, S. (2014). Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature Communications*, 5(4022).

[Jeong et al., 2017] Jeong, E., Ko, K., Oh, S., and Han, H. W. (2017). Network-based analysis of diagnosis progression patterns using claims data. *Scientific Reports*, 7(1):1–12.

[Jiang et al., 2011] Jiang, M., Chen, Y., Liu, M., Rosenbloom, S. T., Mani, S., Denny, J. C., and Xu, H. (2011). A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *Journal of the American Medical Informatics Association*, 18(5):601–606.

[Jiwani et al., 2014] Jiwani, A., Himmelstein, D., Woolhandler, S., and Kahn, J. G. (2014). Billing and insurance-related administrative costs in united states' health care: synthesis of micro-costing evidence. *BMC Health Services Research*, 14(1):1–9.

[Johnson et al., 2014] Johnson, M. P., Zheng, K., and Padman, R. (2014). Modeling the longitudinality of user acceptance of technology with an evidence-adaptive clinical decision support system. *Decision Support Systems*, 57:444–453.

[Jurafsky and Martin, 2022] Jurafsky, D. and Martin, J. H. (2022). *Speech and Language Processing (3rd ed. draft)*. https://web.stanford.edu/∼jurafsky/slp3.

[Kalgotra et al., 2017] Kalgotra, P., Sharda, R., and Croff, J. M. (2017). Examining health disparities by gender: a multimorbidity network analysis of electronic medical record. *International Journal of Medical Informatics*, 108:22–28.

[Kamsu-Foguem et al., 2012] Kamsu-Foguem, B., Tchuenté-Foguem, G., Allart, L., Zennir, Y., Vilhelm, C., Mehdaoui, H., Zitouni, D., Hubert, H., Lemdani, M., and Ravaux, P. (2012). User-centered visual analysis using a hybrid reasoning architecture for intensive care units. *Decision Support Systems*, 54(1):496–509.

[Karami and Safdari, 2016] Karami, M. and Safdari, R. (2016). From information management to information visualization: development of radiology dashboards. *Applied Clinical Informatics*, 7(2):308.

[Kawamoto et al., 2005] Kawamoto, K., Houlihan, C. A., Balas, E. A., and Lobach, D. F. (2005). Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ*, 330(7494):765.

[Kelley et al., 2005] Kelley, E., Moy, E., Stryer, D., Burstin, H., and Clancy, C. (2005). The national healthcare quality and disparities reports: an overview. *Medical care*, pages I3–I8.

[Kim et al., 2018] Kim, M., Banerjee, S., Zhao, Y., Wang, F., Zhang, Y., Zhu, Y., DeFerio, J., Evans, L., Park, S. M., and Pathak, J. (2018). Association networks in a matched case-control design – Co-occurrence patterns of preexisting chronic medical conditions in patients with major depression versus their matched controls. *Journal of Biomedical Informatics*, 87:88–95.

[Klastorin, 1985] Klastorin, T. D. (1985). The p-median problem for cluster analysis: A comparative test using the mixture model approach. *Management Science*, 31(1):84–95.

[Klein et al., 1984] Klein, R., Klein, B. E., Moss, S. E., Davis, M. D., and DeMets, D. L. (1984). The wisconsin epidemiologic study of diabetic retinopathy: Ii. prevalence and risk of diabetic retinopathy when age at diagnosis is less than 30 years. *Archives of ophthalmology*, 102(4):520–526.

[Köhn et al., 2010] Köhn, H., Steinley, D., and Brusco, M. J. (2010). The p-median model as a tool for clustering psychological data. *Psychological Methods*, 15(1):87.

[Kok et al., 2020] Kok, C., Jahmunah, V., Oh, S. L., Zhou, X., Gururajan, R., Tao, X., Cheong, K. H., Gururajan, R., Molinari, F., and Acharya, U. R. (2020). Automated prediction of sepsis using temporal convolutional network. *Computers in Biology and Medicine*, 127:103957.

[Kolari et al., 2006] Kolari, P., Java, A., Finin, T., Oates, T., Joshi, A., et al. (2006). Detecting spam blogs: A machine learning approach. In *Proceedings of the national conference on artificial intelligence*, volume 21, pages 1351–1356. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

[Kong, 2019] Kong, H.-J. (2019). Managing unstructured big data in healthcare system. *Healthcare Informatics Research*, 25(1):1–2.

[Kraff, 2020] Kraff, C. (2020). Why do you need to get an eye exam? `https://kraffeye.com/blog/do-you-need-to-get-an-eye-exam`. Last accessed 2/18/2022.

[Krishnamurthy et al., 2018] Krishnamurthy, M., Marcinek, P., Malik, K. M., and Afzal, M. (2018). Representing social network patient data as evidence-based knowledge to support decision making in disease progression for comorbidities. *IEEE Access*, 6:12951–12965.

[Lea et al., 2016] Lea, C., Vidal, R., Reiter, A., and Hager, G. D. (2016). Temporal convolutional networks: A unified approach to action segmentation. In *European Conference on Computer Vision*, pages 47–54. Springer.

[Lee et al., 2021] Lee, K., Ray, J., and Safta, C. (2021). The predictive skill of convolutional neural networks models for disease forecasting. *Plos one*, 16(7):e0254319.

[Li, 2018] Li, H. (2018). Deep learning for natural language processing: advantages and challenges. *National Science Review*, 5:24–26.

[Li et al., 2020] Li, J., Sun, A., Han, J., and Li, C. (2020). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

[Li et al., 2017] Li, Y., Vo, A., Randhawa, M., and Fick, G. (2017). Designing utilization-based spatial healthcare accessibility decision support systems: A case of a regional health plan. *Decision Support Systems*, 99:51–63.

[Lin et al., 2013] Lin, Y.-K., Chen, H., and Brown, R. A. (2013). Medtime: A temporal information extraction system for clinical narratives. *Journal of biomedical informatics*, 46:S20–S28.

[Mane et al., 2012] Mane, K. K., Bizon, C., Schmitt, C., Owen, P., Burchett, B., Pietrobon, R., and Gersing, K. (2012). Visualdecisionlinc: A visual analytics approach for comparative effectiveness-based clinical decision support in psychiatry. *Journal of Biomedical Informatics*, 45(1):101–106.

[Martel et al., 2016] Martel, M. M., Levinson, C. A., Langer, J. K., and Nigg, J. T. (2016). A network analysis of developmental change in adhd symptom structure from preschool to adulthood. *Clinical Psychological Science*, 4(6):988–1001.

[McElroy et al., 2018] McElroy, E., Fearon, P., Belsky, J., Fonagy, P., and Patalay, P. (2018). Networks of depression and anxiety symptoms across development. *Journal of the American Academy of Child & Adolescent Psychiatry*, 57(12):964–973.

[Members et al., 2016] Members, W. G., Mozaffarian, D., Benjamin, E. J., Go, A. S., Arnett, D. K., Blaha, M. J., Cushman, M., Das, S. R., de Ferranti, S., Després, J., et al. (2016). Heart disease and stroke statistics-2016 update: a report from the american heart association. *Circulation*, 133(4):e38–e360.

[Meystre et al., 2008] Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., and Hurdle, J. F. (2008). Extracting information from textual documents in the electronic health record: A review of recent research. *Yearbook of medical informatics*, 47(Suppl 1):128–144.

[Michelakis et al., 2009] Michelakis, E., Krishnamurthy, R., Haas, P. J., and Vaithyanathan, S. (2009). Uncertainty management in rule-based information extraction systems. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 101–114.

[Mohamed et al., 2007] Mohamed, Q., Gillies, M. C., and Wong, T. Y. (2007). Management of diabetic retinopathy: a systematic review. *Jama*, 298(8):902–916.

[Moor et al., 2019] Moor, M., Horn, M., Rieck, B., Roqueiro, D., and Borgwardt, K. (2019). Early recognition of sepsis with gaussian process temporal convolutional networks and dynamic time warping. In *Machine Learning for Healthcare Conference*, pages 2–26. PMLR.

[Moores, 2012] Moores, T. T. (2012). Towards an integrated model of it acceptance in healthcare. *Decision Support Systems*, 53(3):507–516.

[Moskovitch and Shahar, 2015] Moskovitch, R. and Shahar, Y. (2015). Classification of multivariate time series via temporal abstraction and time intervals mining. *Knowledge and Information Systems*, 45(1):35–74.

[Moskovitch et al., 2019] Moskovitch, R., Shahar, Y., Wang, F., and Hripcsak, G. (2019). Temporal biomedical data analytics. *Journal of biomedical informatics*, 90:103092–103092.

[MTHelpLine, 2022] MTHelpLine (2022). Transcribed medical transcription sample reports and examples. `https://mtsamples.com/`.

[Myint et al., 2018] Myint, P. K., Sheng, S., Xian, Y., Matsouaka, R. A., Reeves, M. J., Saver, J. L., Bhatt, D. L., Fonarow, G. C., Schwamm, L. H., and Smith, E. E. (2018). Shock index predicts patient-related clinical outcomes in stroke. *Journal of the American Heart Association*, 7(18):e007581.

[Nadj et al., 2020] Nadj, M., Maedche, A., and Schieder, C. (2020). The effect of interactive analytical dashboard features on situation awareness and task performance. *Decision Support Systems*, 135:113322.

[Nadkarni et al., 2011] Nadkarni, P. M., Ohno-Machado, L., and Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551.

[Nakayama, 2018] Nakayama, H. (2018). seqeval: A python framework for sequence labeling evaluation. `https://github.com/chakki-works/seqeval`.

[Nam et al., 2019] Nam, Y., Lee, D.-g., Bang, S., Kim, J. H., Kim, J.-H., and Shin, H. (2019). The translational network for metabolic disease – from protein interaction to disease co-occurrence. *BMC Bioinformatics*, 20(576).

[Nelson et al., 2019] Nelson, O., Sturgis, B., Gilbert, K., Henry, E., Clegg, K., Tan, J. M., Wasey, J. O., Simpao, A. F., and Gálvez, J. A. (2019). A visual analytics dashboard to summarize serial anesthesia records in pediatric radiation treatment. *Applied Clinical Informatics*, 10(4):563.

[Ogunyemi and Kermah, 2015] Ogunyemi, O. and Kermah, D. (2015). Machine learning approaches for detecting diabetic retinopathy from clinical and public health records. In *AMIA Annual Symposium Proceedings*, volume 2015, page 983. American Medical Informatics Association.

[Olsen et al., 2000] Olsen, B. S., Sjølie, A.-K., Hougaard, P., Johannesen, J., Borch-Johnsen, K., Marinelli, K., Thorsteinsson, B., Pramming, S., Mortensen, H. B., of Diabetes, T. D. S. G., et al. (2000). A 6-year nationwide cohort study of glycaemic control in young people with type 1 diabetes: risk markers for the development of retinopathy, nephropathy and neuropathy. *Journal of diabetes and its complications*, 14(6):295–300.

[Osborne et al., 2016] Osborne, J. D., Wyatt, M., Westfall, A. O., Willig, J., Bethard, S., and Gordon, G. (2016). Efficient identification of nationally mandated reportable cancer cases using natural language processing and machine learning. *Journal of the American Medical Informatics Association*, 23(6):1077–1084.

[Papakyriakopoulos et al., 2020] Papakyriakopoulos, O., Hegelich, S., Serrano, J. C. M., and Marco, F. (2020). Bias in word embeddings. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 446–457.

[Patel et al., 2008] Patel, D., Hsu, W., and Lee, M. L. (2008). Mining relationships among interval-based events for classification. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 393–404.

[Peleg et al., 2009] Peleg, M., Asbeh, N., Kuflik, T., and Schertz, M. (2009). Onto-clust—a methodology for combining clustering analysis and ontological methods for identifying groups of comorbidities for developmental disorders. *Journal of Biomedical Informatics*, 42(1):165–175.

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

[Piri et al., 2017] Piri, S., Delen, D., Liu, T., and Zolbanin, H. M. (2017). A data analytics approach to building a clinical decision support system for diabetic retinopathy: developing and deploying a model ensemble. *Decision Support Systems*, 101:12–27.

[Pons et al., 2016] Pons, E., Braun, L. M., Hunink, M. M., and Kors, J. A. (2016). Natural language processing in radiology: a systematic review. *Radiology*, 279(2):329–343.

[Pržulj et al., 2004] Pržulj, N., Corneil, D. G., and Jurisica, I. (2004). Modeling interactome: scale-free or geometric? *Bioinformatics*, 20(18):3508–3515.

[Ramos et al., 2003] Ramos, J. et al. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer.

[Rawat and Wang, 2017] Rawat, W. and Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9):2352–2449.

[Rind et al., 2013] Rind, A., Wang, T. D., Aigner, W., Miksch, S., Wongsuphasawat, K., Plaisant, C., and Shneiderman, B. (2013). Interactive information visualization to explore and query electronic health records. *Foundations and Trends in Human-Computer Interaction*, 5(3):207–298.

[Rousseeuw, 1987] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

[Rumelhart et al., 1986] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

[Sang, 2004] Sang, E. T. K. (2004). conlleval: evaluate result of processing conll-2000 shared task. `https://www.clips.uantwerpen.be/conll2000/chunking/conlleval.txt`.

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

[Schmid et al., 2010] Schmid, A. A., Wells, C. K., Concato, J., Dallas, M. I., Lo, A. C., Nadeau, S. E., Williams, L. S., Peixoto, A. J., Gorman, M., Boice, J. L., et al. (2010). Prevalence, predictors, and outcomes of poststroke falls in acute hospital setting. *Journal of Rehabilitation Research and Development*, 47(6):553–562.

[Schäfer et al., 2014] Schäfer, I., Kaduszkiewicz, H., Wagner, H., Schön, G., Scherer, M., and Bussche, H. v. d. (2014). Reducing complexity: a visualisation of multimorbidity by combining disease clusters and triads. *BMC Public Health*, 14(1285).

[Semeraro et al., 2011] Semeraro, F., Parrinello, G., Cancarini, A., Pasquini, L., Zarra, E., Cimino, A., Cancarini, G., Valentini, U., and Costagliola, C. (2011). Predicting the risk of diabetic retinopathy in type 2 diabetic patients. *Journal of Diabetes and its Complications*, 25(5):292–297.

[Sethi, 2009] Sethi, S. P. (2009). *Healthcare Industry in the United States*, pages 9–16. Palgrave Macmillan, New York.

[Shahar, 1997] Shahar, Y. (1997). A framework for knowledge-based temporal abstraction. *Artificial intelligence*, 90(1-2):79–133.

[Shu et al., 2019] Shu, Z., Liu, W., Wu, H., Xiao, M., Wu, D., Cao, T., Ren, M., Tao, J., Zhang, C., He, T., Li, X., Zhang, R., and Zhou, X. (2019). Symptom-based network classification identifies distinct clinical subgroups of liver diseases with common molecular pathways. *Computer Methods and Programs in Biomedicine*, 174:41–50.

[Siddiqui et al., 2018] Siddiqui, N., Dwyer, M., Stankovich, J., Peterson, G., Greenfield, D., Si, L., and Kinsman, L. (2018). Hospital length of stay variation and comorbidity of mental illness: a retrospective study of five common chronic medical conditions. *BMC Health Services Research*, 18(1):1–10.

[Sidorov et al., 2014] Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., and Chanona-Hernández, L. (2014). Syntactic n-grams as machine learning features for natural language processing. *Expert Systems with Applications*, 41(3):853–860.

[Simpao et al., 2015a] Simpao, A. F., Ahumada, L., and Rehman, M. (2015a). Big data and visual analytics in anaesthesia and health care. *British Journal of Anaesthesia*, 115(3):350–356.

[Simpao et al., 2015b] Simpao, A. F., Ahumada, L. M., Desai, B. R., Bonafide, C. P., Gálvez, J. A., Rehman, M. A., Jawad, A. F., Palma, K. L., and Shelov, E. D. (2015b). Optimization of drug–drug interaction alert rules in a pediatric hospital's electronic health record system using a visual analytics dashboard. *Journal of the American Medical Informatics Association*, 22(2):361–369.

[Simpao et al., 2014] Simpao, A. F., Ahumada, L. M., Gálvez, J. A., and Rehman, M. A. (2014). A review of analytics and clinical informatics in health care. *Journal of Medical Systems*, 38(4):45.

[Simpao et al., 2018] Simpao, A. F., Ahumada, L. M., Martinez, B. L., Cardenas, A. M., Metjian, T. A., Sullivan, K. V., Gálvez, J. A., Desai, B. R., Rehman, M. A., and Gerber, J. S. (2018). Design and implementation of a visual analytics electronic antibiogram within an electronic health record system at a tertiary pediatric hospital. *Applied Clinical Informatics*, 9(1):37.

[Singh et al., 2015] Singh, A., Nadkarni, G., Gottesman, O., Ellis, S. B., Bottinger, E. P., and Guttag, J. V. (2015). Incorporating temporal ehr data in predictive models for risk stratification of renal function deterioration. *Journal of biomedical informatics*, 53:220–228.

[Sinsky et al., 2016] Sinsky, C., Colligan, L., Li, L., Prgomet, M., Reynolds, S., Goeders, L., Westbrook, J., Tutty, M., and Blike, G. (2016). Allocation of physician time in ambulatory practice: A time and motion study in 4 specialties. *Annals of Internal Medicine*, 165(11):753–760.

[Sohn et al., 2014] Sohn, S., Clark, C., Halgrim, S. R., Murphy, S. P., Chute, C. G., and Liu, H. (2014). Medxn: an open source medication extraction and normalization tool for clinical text. *Journal of the American Medical Informatics Association*, 21(5):858–865.

[Sokolova et al., 2017] Sokolova, E., Oerlemans, A. M., Rommelse, N. N., Groot, P., Hartman, C. A., Glennon, J. C., Claassen, T., Heskes, T., and Buitelaar, J. K. (2017). A causal and mediation analysis of the comorbidity between attention deficit hyperactivity disorder (adhd) and autism spectrum disorder (asd). *Journal of Autism and Developmental Disorders*, 47(6):1595–1604.

[Soleymani et al., 2020] Soleymani, R., Granger, E., and Fumera, G. (2020). F-measure curves: A tool to visualize classifier performance under imbalance. *Pattern Recognition*, 100:107146.

[Sorbello et al., 2017] Sorbello, A., Ripple, A., Tonning, J., Munoz, M., Hasan, R., Ly, T., Francis, H., and Bodenreider, O. (2017). Harnessing scientific literature reports for pharmacovigilance: prototype software analytical tool development and usability testing. *Applied Clinical Informatics*, 8(1):291.

[Soysal et al., 2018] Soysal, E., Wang, J., Jiang, M., Wu, Y., Pakhomov, S., Liu, H., and Xu, H. (2018). Clamp - a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association*, 25(3):331–336.

[Stitt et al., 2016] Stitt, A. W., Curtis, T. M., Chen, M., Medina, R. J., McKay, G. J., Jenkins, A., Gardiner, T. A., Lyons, T. J., Hammes, H.-P., Simo, R., et al. (2016). The progress in understanding and treatment of diabetic retinopathy. *Progress in retinal and eye research*, 51:156–186.

[Suo et al., 2017] Suo, Q., Ma, F., Canino, G., Gao, J., Zhang, A., Veltri, P., and Agostino, G. (2017). A multi-task framework for monitoring health conditions via attention-based recurrent neural networks. In *AMIA annual symposium proceedings*, volume 2017, page 1665. American Medical Informatics Association.

[Tai-Seale et al., 2017] Tai-Seale, M., Olson, C. W., Li, J., Chan, A. S., Morikawa, C., Durbin, M., Wang, W., and Luft, H. S. (2017). Electronic health record logs indicate that physicians split time evenly between seeing patients and desktop medicine. *Health Affairs*, 36(4):655–662.

[Tantardini et al., 2019] Tantardini, M., Ieva, F., Tajoli, L., and Piccardi, C. (2019). Comparing methods for comparing networks. *Scientific Reports*, 9(17557).

[Thompson, 1968] Thompson, K. (1968). Programming techniques: Regular expression search algorithm. *Communications of the ACM*, 11(6):419–422.

[Ting et al., 2017] Ting, D. S. W., Cheung, C. Y.-L., Lim, G., Tan, G. S. W., Quang, N. D., Gan, A., Hamzah, H., Garcia-Franco, R., San Yeo, I. Y., Lee, S. Y., et al. (2017). Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *Jama*, 318(22):2211–2223.

[Ting et al., 2016] Ting, D. S. W., Cheung, G. C. M., and Wong, T. Y. (2016). Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review. *Clinical & experimental ophthalmology*, 44(4):260–277.

[Topuz et al., 2018] Topuz, K., Zengul, F. D., Dag, A., Almehmi, A., and Yildirim, M. B. (2018). Predicting graft survival among kidney transplant recipients: A bayesian decision support model. *Decision Support Systems*, 106:97–109.

[Torrey and Shavlik, 2010] Torrey, L. and Shavlik, J. (2010). Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global.

[Tscholl et al., 2018] Tscholl, D. W., Handschin, L., Neubauer, P., Weiss, M., Seifert, B., Spahn, D. R., and Noethiger, C. B. (2018). Using an animated patient avatar to improve perception of vital sign information by anaesthesia professionals. *British Journal of Anaesthesia*, 121(3):662–671.

[Tutubalina et al., 2018] Tutubalina, E., Miftahutdinov, Z., Nikolenko, S., and Malykh, V. (2018). Medical concept normalization in social media posts with recurrent neural networks. *Journal of biomedical informatics*, 84:93–102.

[Van Hecke et al., 2005] Van Hecke, M. V., Dekker, J. M., Stehouwer, C. D., Polak, B. C., Fuller, J. H., Sjolie, A. K., Kofinis, A., Rottiers, R., Porta, M., and Chaturvedi, N. (2005). Diabetic retinopathy is associated with mortality and cardiovascular disease incidence: the eurodiab prospective complications study. *Diabetes care*, 28(6):1383–1389.

[Van Leiden et al., 2002] Van Leiden, H. A., Dekker, J. M., Moll, A. C., Nijpels, G., Heine, R. J., Bouter, L. M., Stehouwer, C. D., and Polak, B. C. (2002). Blood pressure, lipids, and obesity are associated with retinopathy: the hoorn study. *Diabetes care*, 25(8):1320–1325.

[Van Valkenhoef et al., 2013] Van Valkenhoef, G., Tervonen, T., Zwinkels, T., De Brock, B., and Hillege, H. (2013). Addis: a decision support system for evidence-based medicine. *Decision Support Systems*, 55(2):459–475.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

[Vishwanathan et al., 2010] Vishwanathan, S. V. N., Schraudolph, N. N., Kondor, R., and Borgwardt, K. M. (2010). Graph kernels. *The Journal of Machine Learning Research*, 11:1201–1242.

[Wang et al., 2021] Wang, R., Miao, Z., Liu, T., Liu, M., Grdinovac, K., Song, X., Liang, Y., Delen, D., and Paiva, W. (2021). Derivation and validation of essential predictors and risk index for early detection of diabetic retinopathy using electronic health records. *Journal of Clinical Medicine*, 10(7):1473.

[Wang et al., 2020] Wang, T., Qiu, R. G., Yu, M., and Zhang, R. (2020). Directed disease networks to facilitate multiple-disease risk assessment modeling. *Decision Support Systems*, 129:113171.

[Wang et al., 2018a] Wang, Y., Liu, S., Afzal, N., Rastegar-Mojarad, M., Wang, L., Shen, F., Kingsbury, P., and Liu, H. (2018a). A comparison of word embeddings for the biomedical natural language processing. *Journal of biomedical informatics*, 87:12–20.

[Wang et al., 2018b] Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S., et al. (2018b). Clinical information extraction applications: a literature review. *Journal of biomedical informatics*, 77:34–49.

[Wang and Yao, 2022] Wang, Z. and Yao, B. (2022). Multi-branching temporal convolutional network for sepsis prediction. *IEEE Journal of Biomedical and Health Informatics*, 26(2):876–887.

[Warner et al., 2015] Warner, J. L., Denny, J. C., Kreda, D. A., and Alterovitz, G. (2015). Seeing the forest through the trees:uncovering phenomic complexity through interactive network visualization. *Journal of the American Medical Informatics Association*, 22:324–329.

[Warner et al., 2016] Warner, J. L., Zhang, P., Liu, J., and Alterovitz, G. (2016). Classification of hospital acquired complications using temporal clinical information from a large electronic health record. *Journal of Biomedical Informatics*, 59:209–217.

[Warner et al., 2013] Warner, J. L., Zollanvari, A., Ding, Q., Zhang, P., Snyder, G. M., and Alterovitz, G. (2013). Temporal phenome analysis of a large electronic health record cohort enables identification of hospital-acquired complications. *Journal of the American Medical Informatics Association*, 22:e281–e287.

[West, 1996] West, D. B. (1996). *Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River, NJ.

[Williamson et al., 2000] Williamson, D. F., Thompson, T. J., Thun, M., Flanders, D., Pamuk, E., and Byers, T. (2000). Intentional weight loss and mortality among overweight individuals with diabetes. *Diabetes care*, 23(10):1499–1504.

[Wills and Meyer, 2020] Wills, P. and Meyer, F. G. (2020). Metrics for graph comparison: A practitioner's guide. *PLOS ONE*, 15(2):1–54.

[Wishart et al., 2008] Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., and Hassanali, M. (2008). Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research*, 36(suppl_1):D901–D906.

[Woolhandler and Himmelstein, 2014] Woolhandler, S. and Himmelstein, D. U. (2014). Administrative work consumes one-sixth of us physicians' working hours and lowers their career satisfaction. *International Journal of Health Services*, 44(4):635–642.

[Wright et al., 2013] Wright, A., McCoy, A. B., Henkin, S., Kale, A., and Sittig, D. F. (2013). Use of a support vector machine for categorizing free-text notes: assessment of accuracy across two institutions. *Journal of the American Medical Informatics Association*, 20(5):887–890.

[Wu et al., 2020] Wu, S., Roberts, K., Datta, S., Du, J., Ji, Z., Si, Y., Soni, S., Wang, Q., Wei, Q., Xiang, Y., et al. (2020). Deep learning in clinical natural language processing: a methodical review. *Journal of the American Medical Informatics Association*, 27(3):457–470.

[Xie et al., 2022] Xie, F., Yuan, H., Ning, Y., Ong, M. E. H., Feng, M., Hsu, W., Chakraborty, B., and Liu, N. (2022). Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies. *Journal of biomedical informatics*, 126(103980).

[Xie and Wang, 2020] Xie, J. and Wang, Q. (2020). Benchmarking machine learning algorithms on blood glucose prediction for type i diabetes in comparison with classical time-series models. *IEEE Transactions on Biomedical Engineering*, 67(11):3101–3124.

[Xu et al., 2010] Xu, H., Stenner, S. P., Doan, S., Johnson, K. B., Waitman, L. R., and Denny, J. C. (2010). Medex: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association*, 17(1):19–24.

[Xu et al., 2012] Xu, Y., Hong, K., Tsujii, J., and Chang, E. I.-C. (2012). Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries. *Journal of the American Medical Informatics Association*, 19(5):824–832.

[Yau et al., 2012] Yau, J. W., Rogers, S. L., Kawasaki, R., Lamoureux, E. L., Kowalski, J. W., Bek, T., Chen, S.-J., Dekker, J. M., Fletcher, A., Grauslund, J., et al. (2012). Global prevalence and major risk factors of diabetic retinopathy. *Diabetes care*, 35(3):556–564.

[Yet et al., 2013] Yet, B., Bastani, K., Raharjo, H., Lifvergren, S., Marsh, W., and Bergman, B. (2013). Decision support system for warfarin therapy management using bayesian networks. *Decision Support Systems*, 55(2):488–498.

[Zhuang et al., 2013] Zhuang, Z. Y., Wilkin, C. L., and Ceglowski, A. (2013). A framework for an intelligent decision support system: A case in pathology test ordering. *Decision Support Systems*, 55(2):476–487.

[Zolbanin et al., 2015] Zolbanin, H. M., Delen, D., and Zadeh, A. H. (2015). Predicting overall survivability in comorbidity of cancers: A data mining approach. *Decision Support Systems*, 74:150–161.

VITA

Suhao Chen

Candidate for the Degree of

Doctor of Philosophy

Dissertation: MACHINE LEARNING OF STRUCTURED AND UNSTRUCTURED HEALTHCARE DATA

Major Field: Industrial Engineering and Management

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Industrial Engineering and Management at Oklahoma State University, Stillwater, Oklahoma in July, 2022.

Completed the requirements for the Master of Management in Corporate Management at Shanghai Jiao Tong University, Shanghai, China in 2010.

Completed the requirements for the Bachelor of Management in Information Management and Systems at Nanjing University, Nanjing, China in 2007.

Experience:

Ping-an Asset Management Co., Ltd., Shanghai, China, 2014-2018
Changsheng Fund Management Co., Ltd., Shanghai, China, 2012-2014
Penghua Fund Management Co., Ltd., Shanghai, China, 2010-2012