APPLICATIONS OF SEQUENTIAL MINING AND DATA MODELING FOR PERSONALIZED MEDICINE

A Thesis
Presented to
The Academic Faculty

by

Kunal Malhotra

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Computer Science

Georgia Institute of Technology
December 2016

APPLICATIONS OF SEQUENTIAL MINING AND DATA MODELING FOR PERSONALIZED MEDICINE

Approved by:

Professor Shamkant B. Navathe
Advisor and, Committee Chair
School of Computer Science
Georgia Institute of Technology

Professor Jimeng Sun, Co-Advisor
School of Computer Science and Engineering
Georgia Institute of Technology

Professor Leo Mark
School of Computer Science
Georgia Institute of Technology

Professor Polo Chau
School of Computer Science and Engineering
Georgia Institute of Technology

Professor Arvind Ramanathan
Computational Science and Engineering Division, Health Data Sciences Institute
Oak Ridge National Laboratory

Date Approved: 7 November 2016

To Mom & Dad for being there selflessly

ACKNOWLEDGEMENTS

I would like to express gratitude and warm appreciation to my advisor Professor Sham Navathe for playing a pivotal role in extending me an offer for the prestigious Ph.D. program in Computer Science at Georgia Tech. I truly feel honored to have been mentored by one of the pioneers in database research during my time in graduate school. Professor Navathe gave me complete freedom to pursue the areas of research that interested me and connected me with top notch researchers in my area to collaborate with. I have always been at awe of his creative and critical thinking and I believe that he has played a significant role in transforming me from a student to a scientist. I would also like to thank Dr.Jimeng Sun for accepting me in his lab in 2014 and assigning me some of the most exciting projects in healthcare with some of the well known industry partners and providing me with all the resources I needed to work in a comfortable environment. He is one of the best scientists I have worked with and I couldn't have asked for a better opportunity to hone my skill set and grow as a researcher. Dr.Jimeng placed his trust in me by making me lead some of the critical projects in his lab and I truly honor such a responsibility placed upon me. I would also like to thank Dr.Arvind Ramanathan for providing me with a research fellow position at Oak Ridge National Lab which is one of the most esteemed national labs in the United States. It has been an amazing experience to work with a great intellect like him and brainstorming with him on projects has really helped me whet my creative thinking ability. I am extremely grateful to Dr.Leo Mark and Dr.Polo Chau for mentoring me through out my Ph.D and advising me on various aspects of my research. At Georgia Tech I have worked with some of the smartest students like Sungate An, Josh Kulas, Robert Chen, Kimis Perros, Yuyu Zhang, Hang Su, Taha, Siddharth, Rocky, Balaji, Shibani Medhekar and Malvika Paul. It has been a great experience collaborating with them at several levels and being mentored by them at the same time. I would also like extend my heartfelt thanks to my friends Karttikay, Vineith, Swati, Sneha, Chinmay, Ankit, Roshna, Rubina, Snigdha and Anchal who have made my stay in Atlanta a cherishable one. A special word of thanks to Jess who has been with me through thick and thin since the time I came to the US and has played an instrumental role in helping me cross some of the most difficult hurdles in graduate school.

I am fortunate to have a fiance like Pallavi back in India who has been my pillar of strength and guided me through all my problems during my tenure in the Ph.D program. She has been extremely supportive and patient and can't thank her enough for keeping me going. I would also like to thank my sister Charu, Sachin and Saesha for always being there and motivating me at all times. Finally and most importantly I would like to thank my parents for being there with me selflessly all these years and having me in their prayers every moment of their lives, always inspiring me to keep my head high even when the chips are down.

Contents

List of Tables

List of Figures

SUMMARY

Healthcare data modeling and analytics as an area of study has gathered momentum amongst clinicians and computer scientists especially after the increased adoption of Electronic Health Records (EHRs) by various provider facilities. Clinicians are getting valuable insights on the efficacy of treatments based on historical medical data of patients. All patients do not respond similarly to existing treatment regimens and hence it is advantageous to stratify patients so that customized treatment plans can be designed, which in turn reduces their financial burden as well. Personalized medicine is a relatively new area in the field of medicine which involves identifying patient profiles and recommending appropriate medical interventions to them based on clinical, genomic and other social factors. Healthcare providers across the United States tend to have a difference of opinion in characterizing and treating patients which makes it inevitable to find out significant sequences of treatment events that show strong correlation with favorable outcomes to be recommended to clinicians. The existing standard coding structures used by EHRs today are extremely difficult to interpret while analyzing data which triggers the need for dependencies on medical ontologies to improve profiling of patients.

Given the above context, this dissertation begins with examining sequential mining approaches to study treatment patterns for a variety of diseases ranging from the rarest of rare cancers such as Glioblastoma (GBM) to some of the more prevalent disease worldwide including heart disease and epilepsy. We propose a non-conventional graph based approach to mine sequential patterns from medical data and come up with clinically relevant constraints to be applied on the graph. We mined the patterns in the sequences and leverage them as features in predictive analytics to build models to solve clinical problems such as survival prediction of GBM patients. We have also conducted a nation-wide analysis of treatment patterns for Autism, Heart disease and Breast cancer revealing the variations in actual practices followed by clinicians. We apply sequential pattern mining to develop epilepsy treatment pathways and report the minor variations which exist across different age groups and types of epilepsy. Anti-epileptic drug resistant patients are identified with the help of a predictive model as part of the analysis.

A known challenge in developing predictive models in healthcare involves utilizing the diagnosis information in a clinically significant manner. Medical ontologies store diagnosis codes in a hierarchical tree representing parent-child relationships which is not leveraged in the process of feature construction for predictive modeling. We exploit this hierarchical information to develop an approach to select the appropriate granularity level for each diagnosis code and develop the most effective feature set for a predictive analytics problem. The thesis also focuses on efficient management of healthcare data. The current EHR systems have been designed to be generic with respect to diseases and have the capability to store common data elements (CDEs). In spite of existence of such CDEs clinical researchers prefer customization of such systems which in turn requires schema modifications at the back-end. We present an automated approach to dynamically modify database schemas without compromising consistency.

Chapter I

INTRODUCTION

1.1 Healthcare Industry in the United States of America

The healthcare system in the U.S. is very different from other advanced industrialized countries. Although a recent legislation was enacted which mandates heathcare coverage for almost every individual, it still does not have a universal healthcare coverage or a uniform healthcare system. In 2010, 86 % of the U.S. population are covered by some type of health insurance out of which 66% are covered by a private health insurance plan. Among the insured population, approximately 36.5% of the population received coverage through the U.S government via Medicare, Medicaid and/or Veterans administration or other military care [43]. Comparisons with other countries helps in tracking the performance of the U.S. health care system and identify the areas of weakness and scope of accelerated improvement . Such analysis is performed based on data obtained from the Organization for Economic Cooperation and Development (OECD), in addition to other sources to evaluate certain health outcomes in addition to the total spending and utilization of the U.S. health care system relative to the other high income countries which include Australia, Canada, Denmark, France, Germany, Japan, Netherlands, New Zealand, Norway, Sweden, Switzerland, the United Kingdom [141, 42].

Expenditure in healthcare by U.S is a lot higher than the other aforementioned countries. In spite of not having a publicly financed universal health system, the U.S spends more public money on healthcare than all but two of the other countries. Ironically americans have fewer hospital admissions and visits as compared to

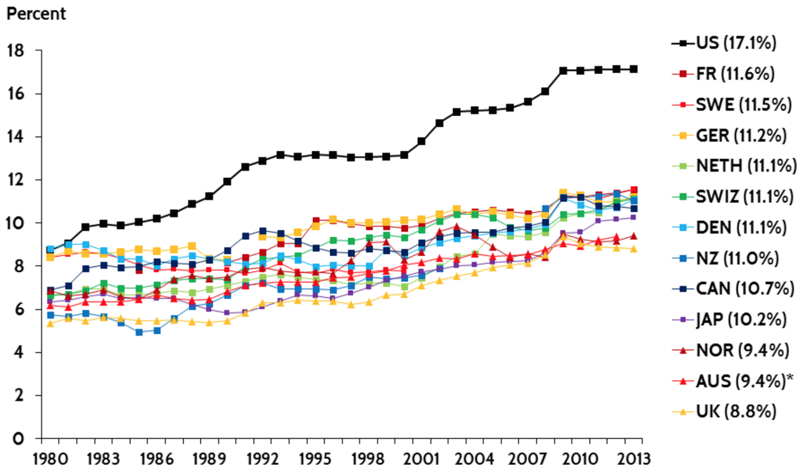


Figure 1: Health Care Spendings as a Percentage of GDP (1980 - 2013). Source: OECD Health Data 2015 [142]

the other countries but still end up paying exorbitant amounts of money towards healthcare. This is attributed to the greater use of expensive technologies such as Magnetic Resonance Imaging (MRI), etc. This finding is corroborated by analyzing the cross-national pricing data which shows notably higher prices for healthcare in the U.S [28]. Figure 1 clearly shows that U.S spent 17.1% of its gross domestic product (GDP) on healthcare in 2013 which was approximately 50% more than France which immediately follows U.S and almost double of what U.K spent (8.8%).

Although the healthcare crisis is severely affected by the increasing cost of healthcare, another concerning factor adding to this problem is the existence of preventable chronic illnesses. Smoking and obesity have stifled the healthcare system in the US by resulting in a wave of different chronic illnesses such as heart diseases, hypertension, diabetes, etc. which eventually results in heavy expenditure in healthcare and unhealthy communities. According to a recent report by CDC, nearly 20% of the Americans smoke which leads to approximately 443,000 deaths annually [61]. 67% of the Americans are obese or overweight, which is a major reason for many serious diseases such as type 2 diabetes, stroke , heart diseases, cancer, etc [54].

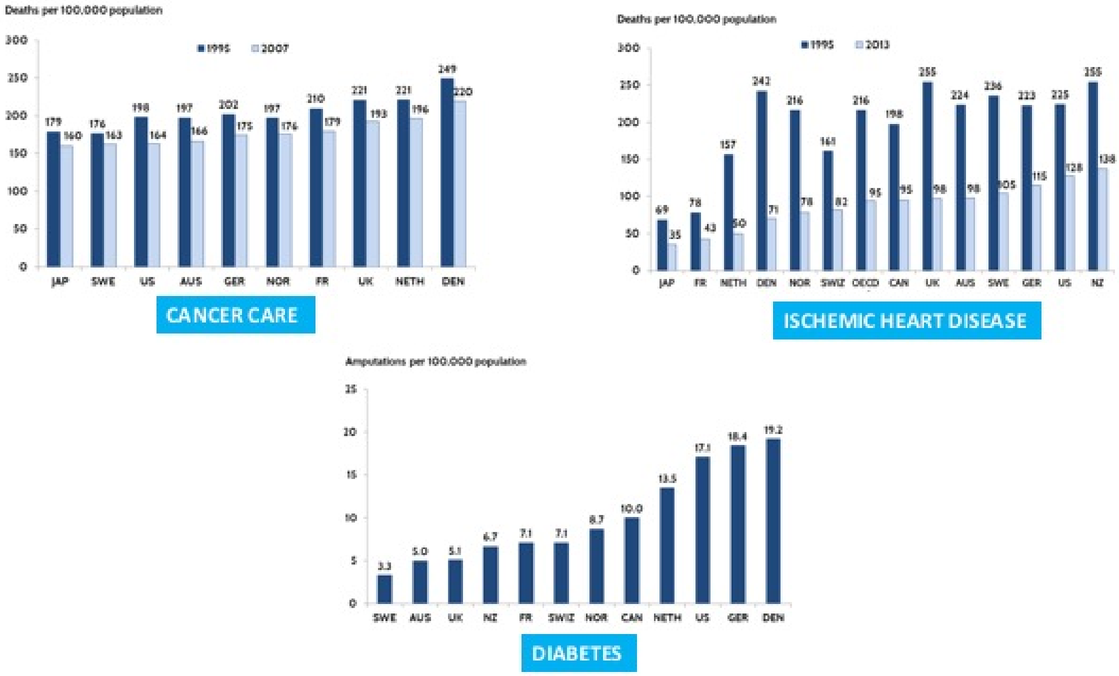


Figure 2: Mortality as a Result of Ischemic Heart Disease (1995 - 2013). Source: OECD Health Data 2015 [142]

With skyrocketing investments made by U.S. in the healthcare sector, there still exist various key healthcare outcomes such as management and prevalence of chronic diseases and life expectancy, which have shown very poor results. Even though the mortality rates from cancer have fallen considerably relative to the other countries, the mortality from ischemic heart disease and amputations as a result of diabetes are still very high [142]. Figure 2 shows a comparison of outcome control achieved by various countries with respect to cancer, heart disease and diabetes and it is evident that while other countries have been successful in achieving favorable outcomes, U.S is still struggling to do so.

Healthcare in the United States and other parts of the world is faced with primarily three challenges with respect to data management: data collection, data sharing and most importantly data analytics [129]. With the immense growth of clinical data being accumulated in electronic health record (EHR) systems attributed to the incentives of EHR adoption in the U.S. provided by Health Information Technology for Economic and Clinical Health (HITECH) Act [26], clinicians and computer scientists are encouraged to use this data to improve quality of care provided and clinical research. The recent landmark ruling of the United States supreme court to largely uphold the Patient Protection and Affordable Care Act (ACO) of 2010 was a major step taken in the direction of redesigning and improving the state of the healthcare system in the country [117]. It also helps in the development of healthcare practices and innovation, improves coordination of care and promotes the use of HITECH by emphasizing the development of a seamless system of care in addition to reducing medical errors which cause more than 100,000 deaths in a year [12]. The clinical data obtained from EHR systems can range from structured forms such as image scans, lab results, diagnosis and pharmacy claims to completely unstructured ones such as textual notes from clinicians, reports, medical literature, etc. A lot of effort is being made by institutions such as Kaiser-Permanente which already has more than 30 petabytes of data to prepare for a data intensive future [75]. The American Society of Clinical Oncology (ASCO) is in the process of developing a Cancer Learning Intelligence Network for Quality (CancerLinQ) which is supposed to provide clinicians with facilities such as clinical decision support, generating insights from data and visualization [137]. The market of global healthcare analytics is expected to reach the $18.7 billion mark in 2020 from $5.8 billion dollars in 2015 and is segmented based on types, applications, components, delivery modes, end users, and regions. Based on the type of analytics, there exists descriptive or retrospective analytics, predictive analytics and prescriptive analytics. The segment of healthcare providers is expected to grow at the highest compound annual growth rate (CAGR). Directives from the federal government to implement EHRs and shift to ICD-10 code sets along with higher focus on value-based medicine and quality of care has played a significant role in the growth of this segment. Region-wise segmentation of the healthcare analytics market involves North America, Europe, Asia and the Rest of the World (RoW). North America accounts for the largest share of this market and with its current CAGR being the highest it is expected to continue dominating the same. The reasons for its dominance could be attributed to the strong federal mandates encouraging adoption of healthcare information technology and significant venture capital investments for analytics [124].

The ongoing research in the area of healthcare analytics is still in its early stages. Data from EHRs are being leveraged to develop models to study important clinical outcomes. One common area of focus has been the use of data analytics to identify the patients who are at risk of being readmitted to a hospital setting after discharge within 30 days. The motivation to solve this problem is attributed to the strict penalization of hospitals by the US Centers for Medicare and Medicaid Services (CMS) Readmissions Reduction Program for excessive number of readmissions. This has revolutionized the ongoing research in healthcare and a lot of effort is being made both in industrial and academic settings to develop models for predicting hospital readmissions [8, 47, 63]. In addition to this, researchers are also interested in solving other important problems such as identification of children with asthma [5], detection of postoperative complications [52] and potential delays in cancer diagnosis [106] .

1.2 Personalized Medicine and Healthcare Analytics

Personalized medicine has been identified as a very recent and rapidly advancing field of healthcare which involves developing treatment regimens customized to a patient's clinical, genetic, genomic and environmental information. These factors play a major role in developing a unique profile for a patient and thus influencing the onset of a particular disease, the course of the same, the response of the patients to drugs or other interventions, etc. Identification of such factors for every patient is of prime importance in order to make individualized and accurate predictions about a person's susceptibility of developing a disease, course of disease and response to treatment. The field of personalized medicine offers both the patients and clinicians the ability to make informed decisions, higher probability of desired outcomes as a result of targeted therapies, early and accurate intervention to help the patient recover sooner and last but not the least, reduce the cost of healthcare [2].

As mentioned before, the striking increase of electronic health data being stored in EMRs recording information about the condition of patients, diagnostic tests, labs, image scans, genomics, proteomics, etc. has made it increasingly important to analyze this wealth of information to derive non-intuitive insights which can eventually improve quality of care and reduce cost of the same. Healthcare analytics covers a wide spectrum of analytical techniques and can be driven by both knowledge and data. Knowledge driven approaches rely on knowledge repositories such as scientific and medical literature, published clinical trials, clinical guidelines, journals, textbooks, etc [24, 90, 76]. A lot of research done in the area of natural language processing and deep learning has translated into products such as Watson Discovery Advisor [3] which makes the machine learn domain knowledge from unstructured knowledge bases to augment decision making and to help clinicians in evaluating alternate diagnoses and treatments. Data driven analytics, on the other hand is driven solely based on underlying observational data collected at the time of giving care. Published guidelines usually focus on a single disease for an average patient or may be targeted towards a geographically biased patient population and thus may not be the best evidence to learn from when developing tools to assist clinicians in giving care. Mining of the real world observational data collected at the individual patient level from different sources can help in a complete understanding of the general healthcare delivery system and make recommendations at a personalized level [76].

1.3 Need for Dynamic Modeling of Data

The organization and delivery of healthcare services is a tedious and an information intensive effort. Some of the major concerns and challenges faced by the healthcare industry include providing better patient care, make use of the best practices, come up with innovative techniques and means of comprehensive healthcare and last but not the least efficient modeling of data. [154]. In the United States the electronic health record systems are expected to gain popularity and their adoption in a healthcare setting is inevitable spurred by financial incentives and penalties mandated in the American Recovery and Reinvestment Act [80]. Even though Electronic Medical Record (EMR) systems have been developed for maintaining consistency in capturing data, healthcare data is anything but consistent. Clinicians and providers have been trained over years to document clinical facts and findings in the most convenient way possible with little regard for standardization and potential analysis which can be performed on this data to improve decision making. Common Data Elements (CDEs) facilitate information to be collected and stored in consistent formats to ensure data coming from multiple sources are standardized and follow the same vocabulary. In spite of these efforts to standardize data, there exist difference of opinions amongst providers and clinical experts based on how they define and characterize a patient suffering from a particular clinical condition. For example, a group of clinicians may define a cohort of asthmatic patients differently than another group. In instances where there exists a consensus, with new clinical findings and medical discoveries clinicians tend to acquire new knowledge and may eventually differ marginally from each other when characterizing a clinical condition. In the recent times, NoSQL databases have gained immense popularity and encompass a wide variety of database technologies which were developed in response to the demands of the modern day applications. When compared to a relational database system, NoSQL data models have a few key characteristics such as higher scalability, superior performance and most importantly flexible data model which in turn can solve the problem of the dynamic nature of healthcare data. But the main drawback of NoSQL technology is that it cannot guarantee strong consistency and does not support secondary indexing, both of which are the strongest qualities of a relational database system. Many of the EHR systems are developed on top of relational databases and they are not amenable to switch easily to NoSQL data model. This makes it inevitable to develop approaches to handle the ever changing healthcare data model in a relational database system.

1.4 Thesis Statement

Improving quality of care provided to patients in a healthcare setting presents many challenges which can be overcome by formulating sequential pattern mining approaches and extensions for treatment pathways and developing medical ontology guided clinically interpretable predictive models.

1.5 Goals and Contributions of the thesis

In this section we present the contribution of this dissertation aimed towards solving most of the problems mentioned before.

1. **Sequential Pattern Mining Extensions - Graph Formulation and Constraints**

   (a) **Non conventional graph approach to combine existing sequential algorithms - SPADE and GSP to come up with extensions to sequential mining.**

   We transform the medical data into a graph consisting of nodes and edges.

The nodes consist of patient nodes and event nodes connected to the patient nodes by edges signifying the events which a particular patient undergoes. These events could be prescription of medications, labs, procedures, etc. The other type of edge is between these events signifying the sequence in which these events occur for a particular patient. Since multiple patients can undergo events in the same sequence, there could potentially be more than one edge between the same set of events distinguishing it from a conventional graph.

(b) **Formulation of Temporal Constraints in the context of medical data - Exact order constraint, Temporal overlap constraint.**

We introduce two clinically relevant constraints which improve the relevance of the the sequential patterns mined. The 1st constraint is the Exact Order constraint which restricts an event node in the sequence to occur immediately after the previous node. This constraint is very useful when studying medical data since certain events such as medications or procedures have a different effect when they are prescribed in an immediate sequence as compared to being prescribed in the same sequence but allowing intermediate events to occur as well. The other constraint is the node overlap constraint which involves transforming the graph structure by combining the nodes of the graph into one to represent events which overlap for a time greater than a threshold.

(c) **Single Node and Combination Node Approach in the graph based method to sequential pattern mining.**

We introduce two approaches to form sequential patterns in the graph called the Single Node Approach and the Combination Node Approach. In both these approaches we have directed edges between a source node and a target node. The difference between the approaches is highlighted when there is a partial overlap between events i.e multiple events occur together for certain period of time but also occur individually or there is a complete overlap i.e multiple events co-occur for a certain period.

   (d) **Use of sequential patterns for feature construction to enhance prediction power of models.**

   We use the sequences of events mined and apply them in the context of predictive data modeling in healthcare. We build predictive models for certain case studies using features from the clinical and often the genomic domain. In addition to these we use the sequential patterns generated as features in the model to enhance the prediction of the models.

2. **Application of different modes of sequential mining to a variety of healthcare contexts.**

   (a) Sequential pattern mining for predictive analytics: This work involves extracting treatment patterns from Glioblastoma patient data and associating these patterns with the class of patients with predicted longevity of over one year.

   (b) Sequential pattern mining for comparative analysis of nation wide treatment treatment trajectories for common disorders.

3. **Treatment pathway analysis for epilepsy & early identification of drug resistant patients**

   (a) An extensive analysis is performed to develop treatment pathways for epilepsy patients in order to understand how clinicians currently treat epilepsy patients. The analysis is done across different age groups and the two major epilepsy types: Idiopathic Generalized Epilepsy (IGE) and Symptomatic Localization Related Epilepsy (SLRE).

(b) We develop a predictive model to detect drug resistant or refractory epilepsy patient population at a very early stage in their treatment for clinicians to perform an intervention with new drugs which have shown promising results in the clinical trials for such patients. The model leverages patients characteristics related to demographics, comorbidities, treatment, encounters and financial status.

4. **Interpretable clinical predictive models via ontology guided feature construction**

   The primary goal of this work is to leverage tree based ontologies which store information about the medical diagnosis codes in a manner so as to optimally select the set of diagnosis codes for feature construction in a predictive model. The approach is inspired from the popular concepts of entropy reduction and information gain which are primarily used by tree based classifiers such as decision trees to perform the split when classifying data.

5. **Automated schema evolution approach via meta-database management**

   The last piece of work involves developing an approach to handle the dynamic nature of healthcare data in a relational database. The work is motivated by the fact that there always exists difference of opinion amongst clinicians about the data elements required to characterize certain diseases. We propose a form based dynamic database system which has the capability to automatically modify the database at the back-end based on the changes made by the user on front-end with minimal intervention of a database administrator. The system can create new schemas based on predefined forms, update and customize schemas based on changing user requirements, etc.

Chapter II

SEQUENTIAL PATTERN MINING EXTENSIONS - GRAPH FORMULATION AND CONSTRAINTS

2.1 Introduction to Sequential Mining

Sequential pattern mining refers to the task of finding frequently occurring events always appearing in a particular sequence as patterns [68]. These patterns are extracted from an input transactional database in which each transaction consists of items annotated with a timestamp which in turn are a set of literals following a particular order. The problem is to find all the sequential patterns of items with a predetermined user specified minimum support where support of a pattern refers to the percentage of transaction containing the pattern [143]. An example of sequential pattern in the healthcare domain would be ‘patients showing symptoms of fever and vomiting in the first visit to the clinic are diagnosed with Malaria within 3 days’. In this example the symptoms of ‘fever’ , ‘vomiting’ and the phenomena of diagnosis of ‘Malaria’ are all items which are part of data collected for patients and each patient visit can be considered a transaction. The retail industry also benefits significantly by applying various sequential pattern mining techniques to make decisions on product placements on shelves in stores [7]. The education sector makes use of sequential pattern mining to extract patterns in source codes to detect cases of plagiarism. This approach also has a lot of significance in web usage mining. Website designers are interested in understanding the user behavior in using websites. The sequence in which certain pages are accessed more frequently than the others can generate a lot of insight about the material on the web, that the users are interested in viewing. Page traversal patterns can help in improving the connectivity of certain pages.

In the healthcare domain it can help advance medical research by extracting non intuitive clinically relevant information in the form of patterns hidden in medical events. For example clinical researchers studying onset of diseases can leverage sequential pattern mining to identify symptoms and diseases that precede or follow certain specific diseases which in turn can assist in improving decision making when planning treatment plans for patients [143, 160].

Agarwal and Srikant in 1995 introduced the problem of sequential mining while studying customer purchase sequences. It states that given a user defined minimum support threshold and a set of sequences in which a sequence consists of an event list and an event consists of a set of items, we need to find all frequent subsequences with frequency greater than or equal to the user defined threshold [6]. Table 1 shows an example of a sequence database consisting of multiple transactions where each transaction represents a patient and the ordered event list of each patient is a set of symptoms presented by him. Assuming the minimum support threshold set is 50% i.e. a subsequence should occur in at least 50% of the transactions for being categorized as a frequent subsequence, it is observed that the subsequence <fever flu >is one of the subsequences which satisfies the minimum support criteria. Traditionally sequential pattern mining was considered to be based on an apriori approach before the advent of other approaches and thus is often called an extension of association rule mining with an additional constraint of timestamps associated with events [68]. Data sources exist in various forms such as transactional , streams, time series, etc. which has led the algorithm development in this area more focussed on the development and improvement for a specific domain. The data used for sequence mining is not limited to the temporal or longitudinal format. For example genome data consists of nucleotides and bioinformaticians are interested in identifying motif patterns. Web logs, alarm data in telecommunications networks and population health data also

Table 1: Sequence database of symptoms presented by patients

Patient_ID	**Sequence of symptoms**
Patient$_1$	<headache fever flu >
Patient$_2$	<fever headache flu doctor_visit >
Patient$_3$	<doctor_visit fever doctor_visit headache >
Patient$_4$	<doctor_visit fever headache flu >

contain data points in a similar fashion and can be viewed as a series of ordered or semi-ordered events, or episodes, occurring at specific times or in a specific order [153] . Thus the problem becomes a search for collections of events that occur frequently perhaps according to some pattern.

2.2 Related Work

There has been a lot of research which has been done in the area of sequential pattern mining to date. Most of the recent work done in this area has focussed on algorithm modifications in specific domains such as telecommunications [111, 155], spatial and geographic domains [23], etc. The algorithms proposed in this area can be categorized into 3 broad classes - Apriori based, horizontal or vertical database format and pattern growth based. The GSP (Generalized Sequence Pattern) mining algorithm [6] which is an Apriori based algorithm extracts frequent sets of items using a downward-closure property of patterns. The approach involves filtering items which are more frequent than a user defined minimum threshold followed by generating all possible candidate sequences by combining items after every iteration. Finally the patterns which satisfy the threshold criteria are retained. Sequential Pattern Discovery using Equivalent Classes (SPADE) is another approach which has been designed to work with a vertical data format and annotates each item with identifiers representing their time of occurrence in a particular transaction [158]. Multiple items are joined on this ID_list to generate sequences. Bellazi et al. [38] have worked on generating temporal association rules using an Apriori approach to help improve care delivery for specific pathologies. These rules consist of antecedents and consequents signifying that if the antecedent occurs then the consequent would also occur with a certain probability. Another algorithm, which is based on temporal association rules is KarmaLego [104]. This is a fast time-interval mining method, which exploits the transitivity inherent in temporal relations.

Frequent pattern growth is another popular technique and unlike the Apriori technique does not have a candidate generation step and is used by a lot of algorithms to mine sequences such as PrefixSpan , SPAM, etc.[68]. The search strategy of SPAM involves a depth-first traversal which generates a bitmap representation of the transaction database followed by pruning based on minimum support threshold [14] . PrefixSpan on the other hand develops prefix and suffix patterns and concatenates them generate sequential patterns [114]. Our approach is inspired by Apriori based methods and reads the data as a graph of events to mine only those sequences which exist in the graph instead of analyzing all possible combination of events. To appropriately apply treatment pattern mining in healthcare context, we introduce several important and clinically relevant constraints for mining sequences such as 'Exact-Order' and 'Temporal Overlap'.

2.3 Non Conventional Graph Approach to Sequential Mining

Graphs have increasingly been used for modeling structures ranging from simple ones like chemical compounds to the more complex ones such as biological networks and protein structures. Almost all the graph based algorithms that have been developed to model data and extract information from it use the conventional graph structure of nodes with single edges between every pair of nodes. In the healthcare domain especially when modeling treatment for a particular disease, there are limited treatment choices for a clinician to choose from when making a decision on the treatment plan for a patient. Using the conventional approach, a graph model would involve generating a graph for every patient consisting of medications prescribed to that patient. This leads to duplication of medication nodes since multiple patients could be treated with the same medication.

We develop a non conventional graph approach to mine sequential patterns from treatment data. It is non conventional due to the fact that nodes are not duplicated under any circumstance, instead multiple edges are created between the same pair of nodes to signify multiple associations.

Since a graphs have the ability to project rich networks and relationships between events in a network, the data at hand is modeled as a graph consisting of nodes, categorized as ‘patient node’ and ‘event type node’. The edges are categorized as ‘undergoes edge’ and ‘sequence edge’. We use a graph based approach to enhance the existing algorithms such as GSP by generating candidates which actually exist in the data with certain frequency as opposed to generating all possible combinations of sequences followed by pruning based on a user defined threshold. Figure 3 shows an illustrative representation of the data as a graph consisting of two patients. The patient nodes can only have patient specific properties such as ‘patient id’, ‘age at diagnosis’, etc. Events for which patterns need to be extracted such as prescribed drugs, labs , BMI (Body Mass Index) elevations and drops, etc. would be represented as event type nodes. The undirected ‘undergoes edge’ between a patient node and an event type node signifies that the patient underwent a particular event with properties corresponding to the process of the event being carried out for that patient. The ‘sequence edge’ is a directed edge between event type nodes signifying the sequence in which these events occurred for every patient. These edges may have properties which would be defined based on the nature of the events being studied. For example, if the events under consideration are medications, then properties such as medication gap and number of days of overlap could be recorded as properties of the sequence edge between medications since they are representative of the relationship between the medications are not patient specific properties. In the figure, the undergoes edge between 'Patient_1' and 'Medical Event B' is representative of the fact that 'Medical Event B' was performed for 'Patient_1'. The start and end dates of this medical event are stored as properties of the 'undergoes edge'. Similarly the same patient undergoes 'Medical Event C' after 'Medical Event B' with a gap of 3 months denoted by the sequence edge between the two events and its gap property. The existing sequential pattern mining approaches have been developed with the aim of reporting the information about the sequence of events which are frequent in a dataset. In addition to that they consider all events occurring together as a single event which may be inaccurate when dealing with medical events especially drug prescriptions which may overlap with each other for variable periods of time. Our graph based approach not only extracts patterns based on time of occurrence of an event in a sequence but also has the capability of factoring in the scenarios where events may overlap for a variable periods of time. We tailor the traditional methods SPADE [158] and GSP [6] to develop a graph based approach with the introduction of constraints customized to the medical field namely the 'Consecutive Occurrence' and 'Temporal Event Overlap'.

2.4 Candidate Generation

In this section we explain how we have generated candidates for mining sequences from the graph. The pseudo-code in Algorithm 1 provides the details of the proposed approach. We define a concept of 'N-path event set', consisting of a sequence of 'N+1' events (treatment instances) joined by 'N' sequence edges. For example, Medical Event_C →Medical Event_B →Medical Event_A is a 2-path event set from the graph model shown in figure 3, consisting of three events forming a sequence of consecutive events. The approach involves extraction of sequences of edge length '1' and prune the ones below the minimum threshold provided. Additional sequences are built up

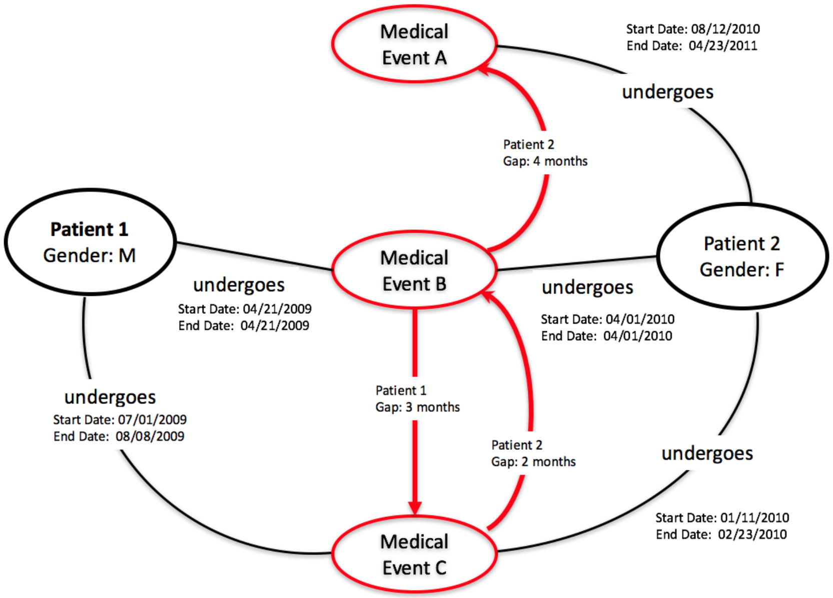


Figure 3: Data represented as a graph

by increasing the edge length in an incremental manner. At each iteration the last N-1 subsequences are compared with the first N-1 subsequences and joined if there is a match and the support criteria are met. For example, to generate 2-path sequences we compare the first and the last subsequences of length '1' to look for potential joins. Figure 4 illustrates the candidate generation step with which each node representing an event type node. C →D and B →D could be potentially joined with D →A to form 2 path sequences such as C →D →A and B →D →A . Similarly to generate the 3-path sequences from 2-path sequences we compare the first and the last subsequences of length 2 and so on. At every step we check the support criteria and prune out sequences which do not meet this criteria. We continue this process till we cannot join any more sequences to form longer sequences while the support for the sequences is above the threshold. Currently with this approach there is potential for generating combinations which may not be following a sequence. For example, a sequence such as D →F can be joined with F →A to form D →F →A but it could be possible that F →A occurred before D →F. To avoid these spurious combinations we introduce event identifiers for every event type node which identifies the time of occurrence (or

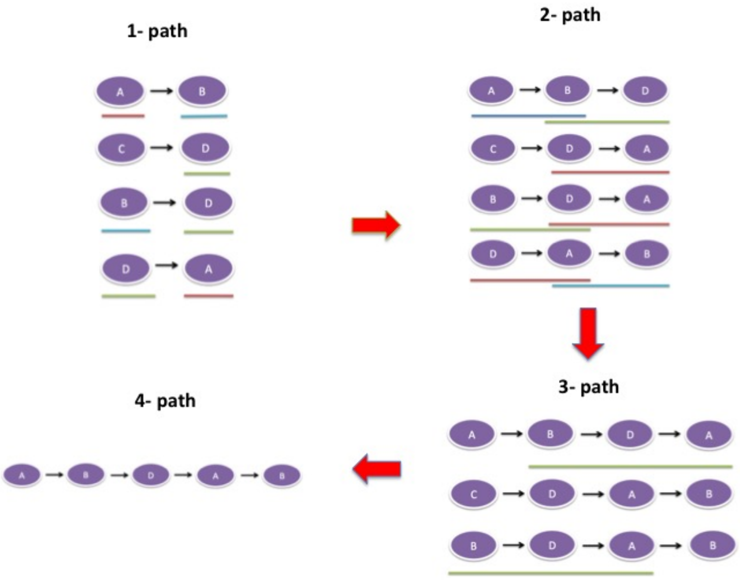


Figure 4: Candidate Generation

position in sequence) of these nodes. The join of two sequences would only take place if the first and the last (N-1)th subsequence of two sequences and the event identifiers match. For example, D.1 →F.2 can be joined with F.2 →A.3. With this condition we introduce a constraint in our approach called the 'Consecutive Occurrence' constraint.

Algorithm 1 MineSequentialPatterns

```
 1: procedure MINETREATMENTPATTERNS
 2:     $N \leftarrow$ Length of $path$
 3:     $minSup \leftarrow$ minimum support
 4:     $N = 1$
 5: REPEAT Steps 6 to 13 UNTILL size of N+1$^{\text{th}}$ path pool = 0;
 6:     $S \leftarrow$ Set of $N$-path sequences of treatment events with support $\geq$
        $minSup$
 7:     for all sequence s $\in S$ do
 8:         for all sequence s' $\in S$ - s do
 9:             $A \leftarrow$ first $N$ treatment events of s'
10:             $B \leftarrow$ last $N$ treatment events of s
11:             if A == B && frequency of $N$+$1^{th}$ sequence >$minSup$ then
12:                 add $N$+$1^{th}$ sequence to $N$+$1$ path pool
13:     $N$++;
```

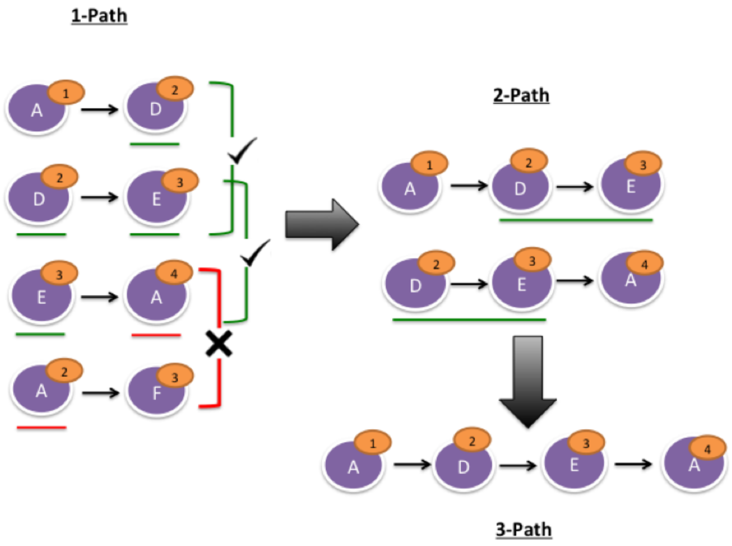


Figure 5: Candidate Generation with Consecutive Occurence Constraint

2.4.1 Consecutive Occurence Constraint

This constraint is introduced to avoid analyzing the set of events which are separated by gaps consisting of intermediate events. For example, in the event set <fever headache flu>, 'fever' and 'flu' has an intermediate event 'headache' which signifies that a patient develops fever which is followed by a headache which in turn is followed by flu. In the process of counting the support for a sequence <fever flu >, this particular patient would not be considered and if there is no patient which has the sequence <fever flu >then this sequence would not be extracted. Figure 5 illustrates the candidate generation step, each node representing an event type. We observe that the combination of sequences E.3 →A.4 and A.2 →F.3 is denied due to mismatch of the event identifiers for the event node A in spite of having the same head and tail. The candidate generation process continues till edge length can no longer be increased due to lack of potential combinations or the support threshold is not met.

2.5 Graph Modifications - Single Node Approach, Combination Node Approach

Certain events co-occur frequently or have a certain period of time when they co-occur. In the medical domain there are events which have the same effect if they occur independently or in concurrence with another event. Certain lab tests ordered for a patient could be done on the same day but clinically it makes no difference if they are performed on the same day or different days. But certain events such as medications may have different effect on the outcome if given individually as opposed to being given with another medication. The overlap of events in time could be either partial or complete. The former is the case when they overlap in time for at least one unit of time and occur independently before or after the overlap whereas the latter occurs when the events begin and end concurrently. Figure 6 shows an illustration of a timeline of a patient who is on multiple medications at different points in time. Figure 6(a) shows the graph representation of the medication prescription data using the medications as individual nodes. We call this approach the ‘Individual Node Approach’. This approach deals with completely overlapping prescriptions such as ‘Lamotrigine’ and ’Gabapentin’ as individual nodes and creates a sequence edge to the next medication prescribed which is ‘Levetiracetam’. For partially overlapping prescriptions such as ‘Phenytoin’ and ‘Topiramate’, the sequence edge is directed from the prescription which begins first to the prescription which begins later. In scenarios where the prescriptions begin on the same date but end on different dates, the edge is directed from the one that ends first to the one that ends later. This leads us to think about a constraint which deals with incorporating overlapping events in the sequences mined.

2.5.1 Temporal Event Overalp Constraint

The temporal overlap constraint, refers to a situation when multiple events can overlap partially or completely. To represent such situations we formulate a 'Combination Node' approach shown in figure 6(b) according to which new nodes representative of co-occurrent prescriptions are introduced whether the overlap is partial or complete. For example, if we have a sequence of events such as E1 →E2 →E3 →E4 and there exists a partial overlap between events E2 and E3, a new node would be created to represent this overlap <E2_E3 >and the graph would transform to E1 →E2 →**E2_E3** →E3 →E4. In case of total overlap the graph transformation would be E1 →**E2_E3** →E4. The timeline shown in the figure shows prescription of medications for a patient. The orange nodes represent individual drugs which already exist in the graph and the yellow nodes are introduced to represent overlapping prescriptions. The 'Combination Node' approach is useful for datasets in which combination of certain events are very frequent. For example, for a particular disease if certain set of drugs are frequently prescribed together, then the set as a whole may become a candidate for certain sequences. But for cases where such combinations are very rare and a variety of combinations exist but not with enough support, then the individual node approach is more useful in using the events in the data as individual nodes to generate sequential patterns above the minimum threshold criteria.

2.6 Summary of Extensions to Sequential Pattern Mining

In this chapter, we introduce a non conventional graph based approach to mine sequential patterns from data. The data is represented as a graph with patient and event nodes along with multiple edges connecting the event nodes based on the number of patients undergoing a certain sequence. The approach has been designed and tailored to incorporate the various scenarios which have been observed in analyzing medical data. The scenarios include discarding patterns in which the events have

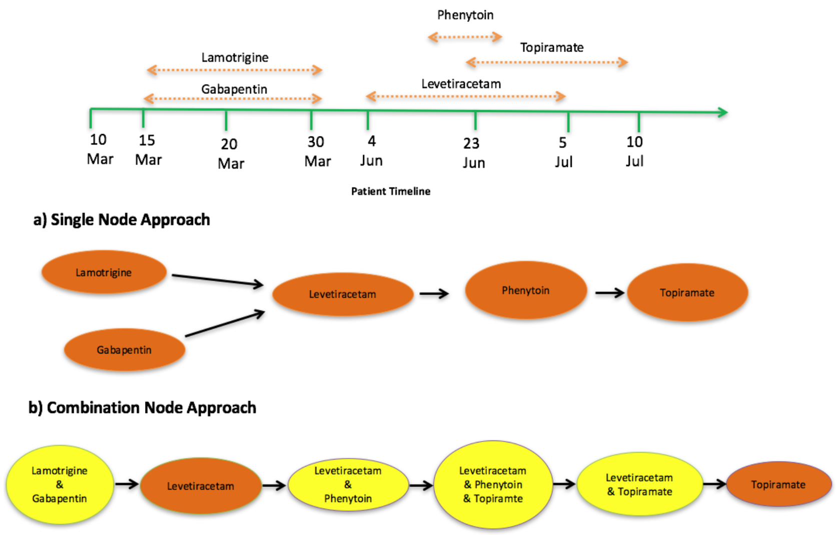


Figure 6: Approaches for Treatment Plan Generation (a) Single Node Approach (b) Combination Node Approach

other intermediate events occurring between them and taking into consideration the possibility of certain events overlapping for certain periods of time such as drug prescriptions which can have a clinically significant outcome. To cater to these cases, we add constraints such as the 'Consecutive Occurrence' and the 'Temporal Event Overlap' to make sure the events in a pattern immediately follow one another and overlapping events are appropriately incorporated in a sequential pattern. We also formulate two approaches namely the 'Individual Node' and the 'Combination Node' approach to generate candidate events which overlap partially or completely and may be used in different circumstances based on the nature of the dataset. We apply the constraint based sequential mining algorithm in the next chapter in the context of predicting survival for Glioblastoma patients and in the chapter after that to perform a comparative analysis of treatment prescribed across United States for three popular disease conditions. Parts of this chapter have been published in [97].

Chapter III

SEQUENTIAL PATTERN MINING FOR PREDICTIVE ANALYTICS - SURVIVAL PREDICTION FOR GLIOBLASTOMA PATIENTS

3.1 Introduction

Approximately 14 million cases of cancers have been reported worldwide along with 8.2 million cancer related deaths in 2012 [147]. Every cancer type has multiple modalities such as chemotherapy, radiation therapy, treatment regimens, etc. and an early diagnosis of cancer could lead to formulation of potentially effective and optimal treatment plans. Even though the primary goal of clinicians and oncologists is to arrest the progression of cancer as early as possible and prolong the survival period of patients, the patient's quality of life is also an important factor to consider when developing treatment plans [110].

Glioblastoma (GBM) is the most lethal and biologically the most aggressive brain cancer with patients having a median survival of 12-15 months [152]. A small percentage of patients survive for longer period of times. Understanding what factors prolong survival and promote treatment responses can be of value to treating physicians.The Cancer Genome Atlas (TCGA) [108], a project of the National Institutes of Health (NIH), led to work done on classifying Glioblastoma into four distinct molecular subtypes which may lead to different treatment regimens [127]. Patients with certain molecular subtypes may have greater overall survival than other patient subtypes and analyzing gene expression levels, copy number variation (CNV), and mutations may give us information about their correlation with survival periods.

The current standard of care for new GBM patients involves surgical resection followed by radiation therapy and chemotherapy with the oral alkylating agent Temodar [112]. Krex et al. [86] and Walid et al. [150] have analyzed newly diagnosed GBM patients undergoing therapy and discovered certain clinical and molecular features which play a significant role in prolonging the survival period. High dimensional gene expression profiling studies in GBM patients have identified gene signatures associated with epidermal growth factor receptor (EGFR) overexpression and survival [56, 92, 103, 105, 109, 116, 125, 134, 148]. Genomic abnormalities associated with TP53 and RB1 mutations have been identified in TCGA along with GBM-associated mutations in genes such as PIK3R1, NF1, ERBB2. CNV and mutation data on TP53, RB and receptor tyrosine kinase pathways revealed that the majority of GBM tumors have abnormalities in all these pathways suggesting that this is a core requirement for GBM pathogenesis [149]. However, no study has systematically examined those genomic factors together with clinical and treatment information for predicting GBM survival outcome, which is a focus of this study.

Predictive survival models have been developed in the past utilizing imaging and clinical features of patients [99] and there also exists ongoing clinical trials on certain drugs to test their effect on survival but to our knowledge there is a lack of comprehensive data-driven work in this space which studies the impact of clinical features, genomic features along with patterns in treatment together on the survival of Glioblastoma patients.

The high mortality rate of GBM patients, where long-term survival is a rare phenomenon, has drawn significant attention to improving treatment of these tumors. After first line standard of care treatment, there are different treatment combinations chosen by oncologists. The sequence in which the next set of drugs or therapy is prescribed adds to the level of complexity since drugs given in a particular sequence may have a better therapeutic effect than the same drugs given in some other order. Furthermore, other drugs such as steroids and antiepileptics are administered in conjunction while treating GBM, which adds another layer of complexity. We believe analyzing the treatment plans of patients from the TCGA may provide insight to certain drug sequences, which may be associated with greater overall patient survival. Based on our knowledge, there is no existing literature that analyzes medication patterns that may influence survival for new GBM patients One of the many challenges in the field of medicine is to make the best decisions about optimal treatment plans for patients. Medical practitioners often have differing opinions about the best treatment among multiple available options. While standard protocols are in place for the first and second lines of treatment for most diseases, a lot of variation exists in the treatment plans subsequently chosen. As a representative disease we study Glioblastoma Multiforme (GBM) which is a rare form of brain tumor. The goal of our study is to predict patients surviving for greater than the median survival period for GBM and discover in addition to clinical and genomic factors, certain treatment patterns which influence longevity. Our study makes the following contributions:

1. We represent GBM patient data in a non-conventional graph to capture the event sequences hidden in the data.

2. We extend existing sequential pattern mining algorithms by incorporating the ‘Consecutive Occurence’ and ‘Temporal Event Overlap’ constraints explained in chapter 2 to mine significant treatment patterns from the available data.

3. We follow a data-driven approach to build and evaluate a predictive model for treatment effectiveness of GBM patients by treating temporal treatment patterns as features in addition to the existing clinical and genomic features.

3.2 Background

It is important to integrate the clinical data in the EHR with the genomic data of patients with GBM since they have a poor prognosis and have a median survival of one year.

Sources of Data: TCGA program and cBioPortal

TCGA began as a three-year pilot in 2006 with an investment of $50 million each from the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI). The initiative has a vision that an atlas of changes could be created for specific cancer types. It also showed that results could be pooled together from different research and technology teams working on related projects and be made publicly accessible for researchers around the world to validate important discoveries [108].

The cBioPortal [33] is developed and maintained by the Computational Biology Center at Memorial Sloan Kettering Cancer Center and provides visualization, analysis and downloading of large scale cancer genomics data sets. The clinical and treatment related data about patients was obtained from TCGA and the data pertaining to mRNA expression, CNV and methylation status of those patients was obtained from cBioPortal. The mRNA expression levels and CNV data was collected for a specific set of genes which have been observed to play a role in classifying GBM patients into 4 genomic subtypes namely classical, mesenchymal, proneural and neural [149]. The methylation status of the promoter region of MGMT gene was also used for our analysis since it has been observed to have an influence on the survival period of the GBM patients [70].

3.3 Data

We have constructed a rich dataset of newly diagnosed GBM patients by integrating two different databases called the The Cancer Genome Atlas Portal [108] and the cBioPortal [33, 58]. TCGA consists of clinical and treatment data pooled together from different research and technology teams, which is publicly accessible around the world to validate important discoveries. The genomic data for the same patients was obtained from cBioPortal, a web resource for multidimensional cancer genomics data maintained by the Memorial Sloan Kettering Cancer Center.

3.3.1 Features

For our study, we analyzed data from 309 newly diagnosed GBM patients spanning over a period of 2 years. The data was categorized into 'Clinical', 'Genomic' and 'Treatment' domains. The clinical domain includes demographic information about the patient along with some basic clinical features such as Karnofsky performance score (KPS), histopathology, prior glioma history, and whether the patient is alive or deceased. Under the genomic domain, the mRNA expression levels and CNV data was collected for a specific set of genes which play a role in classifying GBM patients into 4 genomic subtypes,namely,'Classical','Mesenchymal','Proneural', and 'Neural' [149]. The log2 copy number values were collected from Affymetric SNP6 for each gene and for mRNA expression, Z-scores were used from Agilent microarray. The methylation status of the promoter region of MGMT gene was also used for our analysis [70]. The treatment domain consists of treatment plans for each patient. We use the method of sequential mining to mine significant patterns in their treatment plans and use them as features in the dataset in addition to clinical and genomic features. Table 2 summarizes the dimensions of the dataset categorized by the domain.

Table 2: Dataset summary

Feature Statistics	**Number of Features**
Clinical Domain	11
Genomic Domain	33
Treatment Domain	49
	Total: 93
Data Statistics	
Number of Patients	309
Patients surviving for more than a year	140
Patients surviving for less than a year	169
Race	White (243), Black (42), Asian (24)
Gender	Males (229) , Females (80)

3.3.2 Target Variable

The goal of this study is to apply our extended modeling protocol to effectively predict patients used for model validation who survived for greater than 12 months. The pool of patients used for the study consists of 2 classes of patients. The ones which died within 1 year of diagnosis are assigned a target variable of '0' whereas the ones which survive more than 1 year after diagnosis are assigned a value of '1' . The latter category of patients consist of both living patients and deceased patients who at least lived for more than a year after diagnosis.

3.4 Methodology

This section gives an overview of the predictive modeling pipeline developed to predict long term surviving patients.

3.4.1 Predictive Analytics Pipeline

The pipeline consists of 4 modules, namely, 'Data Standardization and Cleaning', 'Sequential Pattern Mining', 'Feature Construction' and 'Prediction and Evaluation'. As shown in figure 7, the raw data is fed into the 'Data Standardization and Cleaning' module to filter out noisy data. The 'Sequential Pattern Mining' module extracts

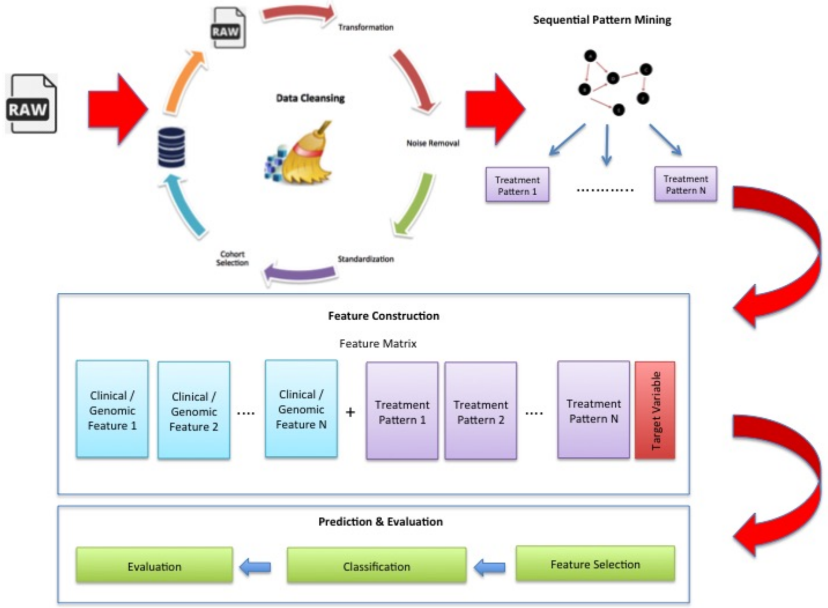


Figure 7: Predictive Modeling Pipeline

significant medication patterns from the treatment data. The clinical and genomic features are combined with the treatment patterns to form a binary feature matrix in the 'Feature Construction' module, each row corresponding a single patient. It also assigns a target variable for every patient. The 'Prediction and Evaluation' module selects predictive features and performs classification to predict the long term surviving patients.

3.4.2 Data Standardization and Cleaning

Data standardization is one of the most important and time consuming steps when building predictive models. Every hospital contributing data to TCGA uses a different format to store data and in some cases a different nomenclature is used for some data elements. For instance for drug names we observed instances of both generic names and trade names. The Anatomic Therapeutic Chemical Classification (ATC) System which is the one of the most commonly used taxonomy for drugs was initially considered to map the drug names to ATC defined codes. We observed that ATC sometimes has multiple codes for a single drug since it is dependent on therapeutic use of the drug and some drugs have multiple therapeutic uses. For example, Prednisone has two ATC codes associated with it namely A07EA03 and H02AB07 and Sirolimus also has two ATC codes L04AA10 and S01XA23. Another approach to standardize the drugs involved generalizing the drug names into broader categories but that would have resulted in reducing the inter drug variability as the distinct number of drugs in our dataset was small (approximately 100). Due to all the above consideration we chose to convert all the drug names to generic names manually. For certain fields such as 'additional chemotherapy' which had binary values '1' signifying the fact that additional chemotherapy was done for a particular patient and '0' signifying otherwise, we found keywords such as 'Completed' and 'Not Applicable'. Based on consultation with the oncologists we decided to replace the value 'Completed' with binary '1' since it means that additional chemotherapy was completed for that patient. The value 'Not Applicable' in this case was replaced with binary '0' since it signifies that additional chemotherapy was never done for this patient and thus was not applicable. Missing data is another common issue. For instance, 10% of data records had missing values for either start or end dates of specific drugs which were imputed based on the mean duration of that drug for other patients since the variance in the duration was small. The data standardization module identifies these different data formats, missing values, and creates a standardized clean data set for further analysis. This module has been customized to clean data coming from TCGA and would require changes when dealing with other datasets.

3.4.3 Sequential Pattern Mining

The GBM data was represented as a graph to capture the medication sequences hidden in the data set. The events in this case study are drugs and radiation therapies which have prescription edges with the patient nodes and sequence edges between them signifying the sequence of prescription. The 'Exact-Order' and 'Temporal Overlap' constraints explained before were applied for this case study since from a clinical perspective medication sequences which have an immediate order of prescriptions have more relevance than sequences in which there exists intermediate unfrequent set of medications which may or may not affect the outcome. We generate a set of frequent sequences which have support greater than a pre specified threshold. These sequences are used as features in the predictive model to study their effects on patient survival.

3.4.4 Feature and Cohort Construction

We construct a feature matrix with a feature vector per patient and a target variable that represents the targeted outcome of treatment. The clinical and the genomic features used for the study are extracted at the time of diagnosis for every patient. The treatment patterns are extracted from therapies prescribed within 6 months from diagnosis. Since the data consists of both numeric and categorical data types we convert the dataset into a binary feature matrix. Each binary feature indicates whether the corresponding clinical, genomic or treatment patterns are present (value 1) or not (value 0). The target variable in our study is constructed based on the patient's survival period. Deceased patients who survived for more than a year are assigned a target variable of 1 and those who survived for less than a year are assigned 0. For living patients, if their last follow up date was after one year of diagnosis, they were assigned a target variable of 1 otherwise that patient was not considered for this study since there is no positive conclusion about survival period.

3.4.5 Prediction and Evaluation

Our goal is to use all the relevant features to predict if patients would survive for longer than a year. 10-fold cross validation is used to partition the data into a training set and a test set multiple times and evaluate the classifier. To avoid overfitting, we make sure sequential patterns are re-evaluated for each training set and then test against the blind test set. No information from test set is used in extracting sequential patterns. In this case, for every iteration of the cross validation the sequential treatment patterns are extracted from the treatment plans of patients in the training set and used as features along with other clinical and genomic features to be tested on the test set. Forward feature selection method is used to prune out irrelevant features and keep the top 10 relevant predictive features to be used by the predictive model. Since the outcome variable has been engineered to be a binomial variable we trained a Logistic Regression based classifier.

3.5 Results

In this section, we present the quantitative results of the predictive models as well as qualitative results for the selected features.

3.5.1 Quantitative Results

In Table 3, we report the performance of various models using both 'single node' and 'combination node' approaches with different mixes of features. Comparison of the performance of these 2 approaches would not be fair since the combination node approach is formulated to represent overlapping medication prescriptions for better representation of features. The C-statistic and accuracy from logistic regression have been reported. Amongst the single domain models the best performance is obtained when only the genomic features are considered. Inclusion of more features increases the prediction accuracy as well as the c-statistic (see table 3). Among the multiple domain models, the best performance is achieved when features from all three domains are analyzed together. Table 4 shows the predictive clinical and genomic features along with influence they have in prolonging overall survival beyond 1 year. The predictive treatment patterns shown in table 5 contain treatment events, which consist of the drug/radiation type with the event identifier in curly brackets categorized by the approach used to form sequences. The bracketed number in the treatment patterns indicates the order number in the event sequence in which the drugs were prescribed. For example, Temozolomide{2}→Lomustine{3} indicates that Temozolomide prescribed as the second drug followed by Lomustine as the third drug in a treatment plan is statistically significant and is predictive of survival.

Table 3: Performance of various models in predicting patients surviving for >1 year using Logistic Regression (LR).

	Single Node Approach		**Combination Node Approach**	
Individual Domain Models	**C-Statistic**	**Accuracy (%)**	**C-Statistic**	**Accuracy (%)**
	LR	**LR**	**LR**	**LR**
Genomic	0.76	78.1	0.76	78.1
Clinical	0.71	72.2	0.71	72.2
Treatment	0.69	71.2	0.60	63.3
Multiple Domain Models				
Clinical + Genomic + Treatment	0.85	86.4	0.85	86.2
Treatment + Genomic	0.84	84.8	0.78	81.0
Clinical + Genomic	0.83	84.5	0.83	84.5
Clinical+ Treatment	0.78	78.6	0.75	74.5

Table 4: Predictive clinical and genomic features from the model: Clinical + Genomic + Treatment

Predictive Features	**Percentage of times selected (%)**	**Influence on Survival >1 year**	**P-value**
Genomic			
Unmethylated MGMT promoter region	40	Negative	0.05
High expression of TP53 gene	40	Negative	0.03
High expression of GABRA1 gene	40	Positive	<0.0001
Clinical			
Patient's age at diagnosis between 25 & 50 years	40	Positive	0.018
Karnofsky performance score >70	40	Positive	0.02
Prescription of Neoadjuvant therapy	30	Positive	0.002

3.5.2 Qualitative Results

Besides accurate prediction results, the predictive features are also clinically meaningful. Methylation of the MGMT gene has been reported to be crucial for some

Table 5: Predictive treatment patterns

Predictive Treatment Patterns	**Percentage of times selected (%)**	**Influence on survival >1 year**	**P-value**
Single Node Approach			
Radiation Therapy{2}→Treatment Termination	50	Negative	0.0061
Lomustine{2}→Treatment Termination	40	Negative	0.05
Procarbazine{2}→Treatment Termination	30	Positive	0.05
Temozolomide{2}→Lomustine{3}	40	Positive	0.04
Combination Node Approach			
Temozolomide{1} →(Temozolomide +Radiation){2}	30	Negative	0.05
(Temozolomide + Radiation){1}→Temodar {2} →Lomustine{3}	30	Positive	0.04

of the standard of care chemotherapeutics such as Temozolomoide (Temodar) to be effective which in turn prolongs survival [70, 122] . GBM patients having an unmethylated promoter region of the MGMT gene are less likely to survive for more than a year. In addition, a higher expression of the TP53 gene is associated with shorter survival periods, while higher expression of the GABRA1 gene, also called the gamma-aminobutyric acid (GABA) A receptor, alpha 1, is associated with longer survival periods. In the clinical domain, younger patients, in the age group of 25-50 years, have a higher chance of surviving longer. Another factor is the Karnofsky performance score which is a score ranging from 0 to 100 , assigned by clinicians to GBM patients based on their functional status prior to treatment (refer to the appendix for score descriptions)[53]. Patients with a higher score are healthier than the ones with a lower score. Patients having a score greater than 70 were observed to have survived for longer than a year. Another predictive clinical factor is neo-adjuvant treatment which is given as the first step to shrink the tumor before the main treatment is begun. Patients receiving neo-adjuvant treatment were found to survive for longer periods. (Neo adjuvant drugs include PolyLCLC, Mivobulin isethionate, Oxaliplatin, O6-Benzylguanine and Carmustine). Most importantly, our study also discovered treatment patterns, which have had both positive and negative effects on the survival period. The standard first line of treatment consists of surgery followed by fractionated External Beam Radiation Therapy (EBRT) with concurrent and adjuvant Temozolomide therapy. This combination is associated with the best survival in GBM patients and is the standard of care. Fractionated radiation is given solely for some patients if they cannot tolerate chemotherapy. We have also found that treatment consisting of EBRT or the chemotherapeutic Lomustine, as the second event in the treatment timeline present individually or in combination with another drug reduces the likelihood of longer survival. This can be explained by patients having unresectable tumors. As a result, prescribing EBRT may not be effective and does not lead to greater overall survival. Lomustine prescribed as the second drug in the treatment is also unusual since most clinicians prescribe the standard of care Temozolomide treatment and Lomustine is not prescribed early. Two treatment patterns using single node approach were found to have a positive influence on the survival period, one consisting of Procarbazine prescribed second in the treatment plan in combination with other drugs or by itself followed by termination of treatment and the second one consisting of Temozolomide prescribed second in the treatment plan immediately followed by Lomustine.

Using the combination node approach, we found that if Temozolomide is prescribed individually as the first event followed by Temozolomide with concurrent EBRT then there is a negative effect on survival. We believe this could be due to the explanation given before about patients not having a resectable tumor or it is also possible that if radiation therapy is not coupled with Temodar as the first event, which is the standard of care, then the treatment does not turn out to be effective. The other predictive treatment pattern, which we have found to have a positive effect on survival, is Temozolomide with concurrent radiation therapy followed by Temozolomide prescribed individually which is in turn followed by a prescription of Lomustine as the third event.

3.6 Conclusion

In this case study, we discuss a pipeline performing data standardization, mining sequential treatment patterns, and constructing features of predicting GBM patients surviving for longer periods (greater than 12 months). A novel constraint based sequential mining approach is applied to capture clinically meaningful patterns by adding exact-order and overlap constraints. Accurate prediction (0.85 c-statistic) can be obtained with logistic regression model using combination of clinical, genomic and treatment pattern features. Many predictive features can also offer interesting clinical insights. This study is a preliminary step in providing extensive treatment guidance to oncologists and neurosurgeons about the efficacy of certain sequence of drugs and therapies as part of a treatment plan. Currently the study is focused and driven by care provided in the area of cancer treatment. In the future we would like to explore the possibility of extending the current approach for chronic conditions such as diabetes and hope to find interesting patterns in patient trajectories since the volume of data would be large as opposed to acute conditions like Glioblastoma. Currently, the treatment patterns consist of the drug names and their event of prescription. We also plan to add more constraints in the model such as a 'gap' constraint, which would limit the temporal gap between events for inclusion in a sequence. We believe this would help in filtering out clinically insignificant treatment patterns. We have discussed these results with our medical collaborator for this work Dr. Hadjipanayis and he considers these significant in the light of existing clinical treatment literature about which we have limited clinical expertise. This work has been published in [97].

Chapter IV

SEQUENTIAL PATTERN MINING FOR COMPARATIVE ANALYSIS OF NATION WIDE TREATMENT TRAJECTORIES

4.1 Introduction

Electronic medical records (EMR) are central to the success of modern healthcare: they offer detailed information into a patient's medical history along with the clinical practices followed from diagnoses, to treatments and recovery [71]. EMRs currently record ICD9 (International Classification of Diseases and Related Health Problems, v9) and CPT (Current Procedural Terminology) codes for describing a clinical condition and the clinical care provided, respectively. Due to their rich information content, EMRs have become indispensable tools for aiding physicians in delivering data-driven clinical decisions and care for their patients [77, 60, 82, 73, 131, 93]. Accessing EMRs for quantitatively assessing public health and epidemiological research, however, can be challenging [100, 17, 102, 27]. One major concern is the data privacy issues that are inherent to EMRs [98, 20, 16]. A second concern is that even though several EMR vendors support similar information regarding clinical practice (coded via ICD9 and CPT procedure codes), this information is often not throughput. Thus, EMRs face immense challenges in homogenizing information across hospitals and individual medical practices [98, 29, 59]. This makes data sharing, hosting, analysis and visualization extremely challenging. To overcome these challenges, we use electronic healthcare reimbursement claims (EHRC) for public health and epidemiological research. EHRCs are regularly collected as part of financial transactions processed by individual medical practices or retail pharmacies to treat, or to dispense prescription drugs to patients. These datasets include ICD9 and CPT codes and therefore capture timely information regarding any medical condition specific to a patient. EHRC data also record patient specific information (e.g., age, sex, state-specific location) and physician information. Therefore, we hypothesize that claims data can provide insights into patient-centric clinical trajectories that can provide information about how patients are treated for specific medical conditions. To our knowledge, this is the first attempt in using claims data to reconstruct patient-specific clinical trajectories. Our *contributions* can be summarized as follows:

- We leverage sequential mining algorithms and adapt them to work with the claims data to construct patient specific clinical trajectories.
- We apply our sequential mining tool to three medical conditions: (a) autism spectrum disorder (ASD), (b) heart disease (HD) and (c) breast cancer (BC). By tracking the sequence of clinical procedures across a large cohort of patients, we uncover the heterogeneity in clinical procedures followed and the costs incurred across the entire US.

4.2 RelatedWork

Several studies have focused on using temporal mining techniques to analyze EMR data [19, 161, 126, 35, 22, 113]. These studies have used either administrative codes or narratives from EMRs to identify temporal patterns from sets of clinical episodes. Other studies have used temporal mining approaches to monitor adverse effects for specific vaccinations [145] or to mine associations between diagnostic codes [69]. A few applications have focused on visual analytics for clinical trajectories constructed from EMRs [65, 64]. Malhotra et al. [97] used temporal mining to predict survivability of glioblastoma patients from the cancer genome atlas network. We use the same temporal mining framework as described in [97], however, we adapt it to work with the claims dataset from IMS Health for the year 2009-2010 (described further in the

Data section). In our previous work, we have described how claims data can provide novel insights into the spread of infectious diseases such as the flu [120, 121, 119] . However, the application of temporal mining to construct patient specific clinical trajectories from claims data is novel. Further, we have used the sequential mining approach to analyze a massive volume of claims data (with over 1 billion total claims) and across three different disease conditions.

4.3 Data

Table 6: Summary of EHRC data analyzed

	ASD	**HD**	**BC**
Patients	89,624	1,620,605	1,045,986
CPT	7,343	15,983	5,402
Claims	1,228,239	1,367,369	25,403,204

We acquired one year worth of EHRC data spanning a period of one year (from Apr 1, 2009 to Mar 31, 2010) from IMS Health after removal of sensitive identifiable information from the IMS Health data warehouse. The processed EHRC data includes only the ambulatory care reimbursement claims data, which provides access to a patient's claim record consisting of a unique patient identifier (not tied to the original identifiable record), claim identifier, the CPT code, ICD9 code, physician identifier (who treated the patient) and date of service provided to the patient. In addition to the claims data, we also used two allied datasets that provided demographic information for physicians and patients. For this study, we did not include additional fields that were part of the claim, since we were only interested in constructing the temporal history of CPT procedures for the patients.

Table 6 provides a summary of the claims data, and in particular for the claims related to the three medical conditions examined in this study. The rationale for choosing these three conditions was based on recent statistics: (a) ASD represents one of the most complex medical conditions to diagnose and treat, and an estimated 1 in 68 children suffer from it [44]; (b) HD represents one of the most prevalent conditions within the US population, with an estimated 25% (in 2008) of all deaths attributed to HD [107] ; and (c) BC represents one of the most commonly occurring cancers and in particular has very established clinical procedures [66]. The claims represent about an estimated 47% (about 1 billion) of all electronic healthcare reimbursement claims processed throughout the country and therefore represents a massive volume of data to be analyzed.

4.4 Results

1. **Claims data are correlated with disease occurrence from CDC-NHDS** Prior to analyzing the data and constructing patient specific profiles of treatments being provided, we evaluated whether the claims data are accurate in capturing disease incidence rates across the country. In our previous studies [120, 121], we showed that the the incidence rates for influenza across the entire nation as well as individual Human and Health Services (HHS) regions I-X are correlated with incidence rates from influenza-like-illnesses network (ILINet) published by the CDC. In this study, we show that the claims data is correlated with publicly available datasets such as the National Hospital Discharge Survey (NHDS) from the Centers of Disease Control (CDC) for three other widespread medical conditions namely, autism spectrum disorder (ASD), heart disease (HD) and breast cancer (BC).

   To compare the claims with NHDS data, we extracted the average weekly counts of claims for the ICD9 codes corresponding to ASD, HD and BC and normalized them by the total population. While the claims data provide access to individual zip codes of physicians who treated individuals, NHDS data provides access to total numbers of individuals (with a particular condition) at the state level. Here, we chose to compare the NHDS and claims data at the national level (although the correlations between the two datasets across individual HHS regions are also similar).

As shown in Figure 8A, for ASD, there is very little correlation between claims data and the NHDS, indicating the lack of any agreement between the two datasets. This is not surprising, given the fact that NHDS examines only hospital discharge rates (and is limited only about 239 non-institutional hospitals for the year 2009 and excludes any Federal, military or Veteran Administration hospitals) and ASD usually do not require hospitalization of patients (unless under extreme cases). On the other hand, the claims data reveal a significantly large number of ASD diagnoses in 2009-2010. The correlations for HD and BC claims are quite high when compared with the corresponding NHDS data (average correlation of 0.8 with p-value of 1E-6), and interestingly the weekly trends reflect remarkably well across the two datasets. Notably, both HD and BC require significant hospital care, and therefore, we observe a strong correspondence between the two datasets.

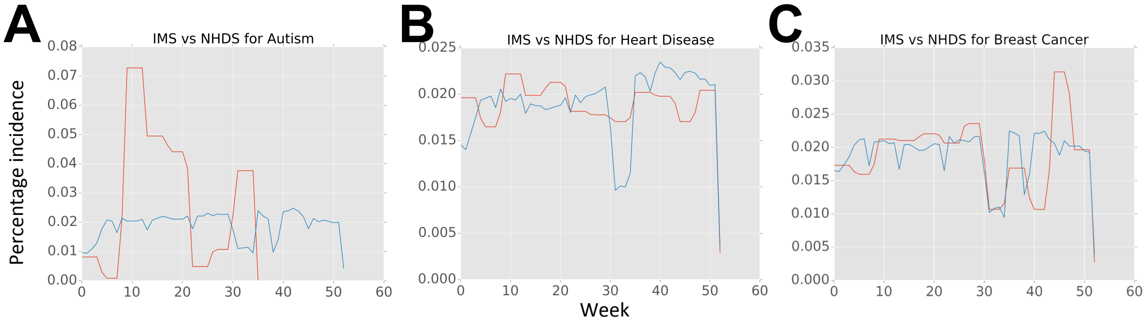


Figure 8: Comparison of IMS claims (red line) versus CDC-NHDS data (blue line) in weekly trends of (A) ASD, (B) HD and (C) BC incidence across the entire US. Note that the counts are normalized by estimates of total population across the US (in 2009-2010), similar to NHDS [1].

2. **Findings regarding clinical procedures followed in multiple regions of the country**

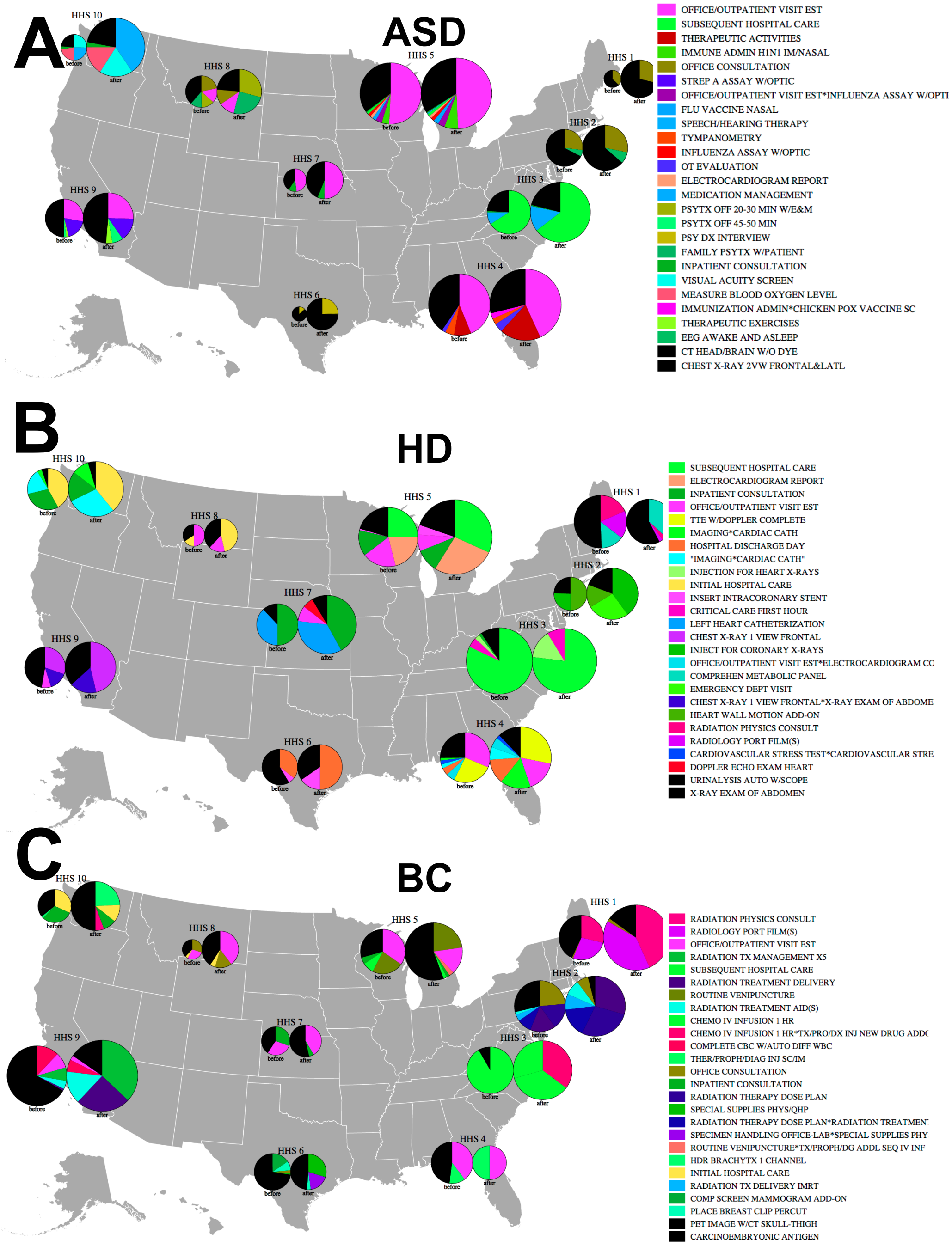


Figure 9: A visual summary of commonly observed clinical procedure sequences detected from sequential mining approach for (A) ASD, (B) HD and (C) BC for the ten HHS regions shown as pie charts. Two pie charts are shown per HHS region, with corresponding labels namely, before diagnosis and after primary diagnosis. The size of the pie charts are proportional to the total number of claims before and after the diagnosis. Every slice in the pie chart corresponds to one of the top twenty (most commonly observed) clinical procedures using the procedure plan generation process and is proportional to the overall costs incurred in administering that procedure. The legend on the right hand side illustrates the details of the commonly observed procedures.

Sequential mining was performed on the procedure claims filed for ASD, HD and BC with the aim of extracting common frequent procedural event sequences. In addition to analyzing the sequences we were also interested in analyzing the variation in costs of such procedural patterns across USA. The analysis was done across ten HHS regions since practitioners in different regions may have different protocols for treating patients. In addition to this since there may be variation in procedural patterns before and after primary diagnosis of patients, we categorize the claims data into *before primary diagnosis* and *after primary diagnosis*, where the primary diagnosis can be either ASD, HD or BC. Our analysis shows that: (1) The procedural patterns observed in the *after primary diagnosis* subset were observed more frequently than the ones in the *before primary diagnosis* subset. (2) The *after primary diagnosis* subset was also found to have new procedural sequences which were not observed in the *before primary diagnosis* subset.

We represent a summary of the top twenty CPT codes used to treat the patients across the ten HHS regions as pie charts (see figure 9), where each slice represents a certain fraction of clinical procedure sequences used. The individual CPT procedures are shown in the legend, with a brief description of the procedure followed.

Lack of consensus in clinical procedures for ASD diagnoses and subsequent treatment across HHS regions : The sequential mining of clinical procedures prior to autism diagnosis reveals significant heterogeneity in the clinical procedures followed (see figure 9A). Surprisingly, even though we examined ASD related claims, we observed that within HHS regions III, V and VI common procedural sequential patterns involved the administration of flu vaccine immediately after a chest x-ray procedure or after a regular outpatient visit. On the other hand, within HHS regions VI, VIII and IX, psychotherapy procedures such as interviews with the patient and his/her family are commonly observed. Speech therapy and evaluation coupled with oxygen level measurement was also observed as a common sequence in HHS regions II and X whereas HHS-IX (surprisingly) shows rapid strep test for streptococcal infections.

CPT patterns administered often vary across HHS regions even after diagnoses. After ASD diagnoses, in HHS regions such as HHS-I, CPT codes associated with neuromuscular reeducation, comprehensive metabolic panel and psychotherapy are dominant, whereas in the other regions such as HHS-II and HHS-X procedures involving speech and hearing evaluation, ECGs, and psychotherapy dominate. Further, in HHS-I, CPT patterns involving comprehensive metabolic panel (CMP) in combination with thyroid testing as one of the first procedures performed followed by another round of CMP, implying that patients were monitored for progress or for further diagnoses. We also found that EEG monitoring (which was one of the dominant procedures in HHS-I for patients before being diagnosed with autism) is very frequently administered in HHS-I after diagnosis. Speech and hearing therapy, which were performed before diagnosis in HHS-X, were more dominant in HHS-II after diagnosis. Other procedures involving psychotherapy interviews following their discharge from the hospital, speech therapy followed by oxygen level measurements amongst diagnosed patients have been observed to be significant in regions IX and X.

Consensus in CPT sequences after HD diagnoses Similar to ASD, we present a visual summary of claims for HD, consisting of both the before and after scenarios (Figure 9B). In contrast to ASD, we observed approximately 4 to 5 procedural sequences common across all the HHS regions. For example, some of the topmost CPT sequence patterns involved electrocardiograms and insertion of intra-coronary stents in approximately 40% of the patients. Further, 35% of the claims of emergency visits were observed after being diagnosed with heart disease along with a proportional increase in the number of patients in critical care. For most of the HHS regions we observed new CPT patterns prescribed after the primary diagnosis was made. These CPT patterns include platelet count in combination with troponin test, cardiac catheterization immediately following an ER visit, across most HHS regions, reflecting a general consensus amongst medical practitioners in how HD patients are treated.

Consensus in CPT sequences before and after BC diagnoses In breast cancer we observed a lot of homogeneity amongst the different HHS regions in terms of the sequence in which the procedures were administered (Figure 9C) although there was a marked difference between the frequency of procedures administered before and after diagnosis. Certain procedures such as comprehensive metabolic panel followed by an infusion of chemotherapy were observed in HHS-I in approximately 15% of the patients before diagnosis which increased to 40% after diagnosis. Another procedure commonly found in approximately 55% of the patients in HHS-IV both before and after diagnosis was an injection of therapeutic prophylactic immediately following an outpatient visit. The prophylactic injection was also predominant in HHS-III in combination with chemotherapy but only after the primary diagnosis was made. Many of the procedures which emerged as significant in HHS-II and HHS-V after diagnosis were setting of the radiation therapy field followed by a dose plan of radiation therapy or special radiation dosimetry. Radiation therapy field setting coupled with radiation treatment in turn followed by radiation management was found to be predominant in HHS-IX in approximately 60% of the patients.

3. **Findings related to cost of treatment prescribed** The details of the commonly occurring procedure sequences as well as their costs across the HHS regions are summarized in Tables 7, 8 and 9. Within the tables, we highlight only those HHS regions that had a statistically significant number of CPT procedure sequences extracted from our sequential mining approach. The number beside each clinical procedure indicates at which part of the sequence the procedure was administered (e.g., [Neuromuscular reeducation]2 indicates that the procedure 'neuromuscular reeducation' was performed at position 2 in the sequence). A right arrow (→) is used to indicate progress towards the next step in the clinical procedure. We also highlight combination nodes using a '*' in between the two individual procedures.

   Based on our analysis, the costs incurred to treat patients increases significantly after the primary diagnosis is made. For ASD, the procedural sequences involving speech and hearing therapy increased by nearly 7 to 8 fold (see Table 7). The cost of psychotherapy interviews of patients and family was also observed to be 3 times more than the cost incurred before diagnosis (Table 7). The cost of performing MRI and EEG procedures seemed to increase only by 1 to 2 fold implying that two procedures do not incur a large change before/after the primary diagnoses (Table 7). Another procedure which showed a moderate increase in cost was 'Neuromuscular Reeducation' which showed a 3 fold increase.

   For HD, some of the procedures incur a very high cost after the primary diagnoses (12-15 fold increase in the Costs (after) column in Table 8). Examples of such procedures include intravascular ultrasound and cardiac catheterization, which are usually treatments that have to be administered upon diagnosis. It is also possible that such procedures are often carried out in emergency situations (as evidenced by the total number of emergency visits and hospitalization events, which is significantly higher in HD cases) and therefore incur higher costs. On the other hand, routine tests such as cardio-vascular stress tests and chest x-rays showed minimal increase in costs incurred.

For BC, there is a very steep increase in the costs (12 to 15 fold) associated with procedural patterns involving ‘Radiation Physics’, ‘Radiation port films’ and ‘Chemotherapy’ after the primary diagnoses. These procedures are important clinical intervention steps and therefore it is reasonable to expect an increase in the costs incurred. Similar to the HD case, the costs of the routine procedures such as cell blood count and therapeutic prophylactic injections showed only a moderate increase of 2 to 3 fold.

The cost increase observed from comparing the before and after diagnoses columns in Tables 7, 8 and 9 suggests which clinical procedure sequences may be more important. For example, the cost of the procedure sequence involving ‘Cardiac catheterization’ followed by ‘Doppler echocardiography’ increases many fold after primary diagnosis and is also very relevant in heart disease patients. Additionally, we can also observe that a majority of the procedures for HD and BC appear only after the primary diagnosis, implying that once the diagnosis has been made, the clinical procedure patterns followed share significant similarity.

4.5 Discussion

Our sequential mining approach showed that it is feasible to identify commonly observed sequences of clinical procedures administered to patients across the entire country using EHRC data. Although the length of these clinical procedure sequence patterns were usually not more than three events in length, we found significant heterogeneity in the way clinical procedures are administered across different HHS regions. In spite of this heterogeneity in the claims data, we were able to identify consensus CPT sequence patterns across different regions. Furthermore, we observed that the consensus patterns were significantly higher for HD and BC, than for ASD patients. This may reflect the diversity in treatment options and medical opinions that exist in current medical practice about autism and its effects on people, but also reflects the difficulty in diagnosing and treating ASD in general. An allied finding from our study also showed a distinct difference in the costs incurred for the CPT procedure patterns before and after the primary diagnoses (i.e., ASD, HD and BC). Further, several of the CPT sequences became more prominent after the primary diagnoses, especially for HD and BC, indicating that once the practitioners make a diagnosis, procedure plans tend to follow a similar trajectory for a majority of the patients.

Our study also highlights some of the practical limitations and emerging challenges for analyzing EHRC data. (Note that our study only tracked CPT sequence patterns). However, physicians perform procedures in response to a particular diagnosis (e.g., ICD9 code). Thus, to obtain better insights into clinical decision making processes and how procedures are subsequently administered, we have to incorporate richer semantics into our graph construction process to include both ICD9 and CPT codes. Such an extension to the algorithm itself is straightforward, and we plan to pursue such an extension in the immediate future.

Currently, we have not quantitatively understood whether the information contained within EHRCs is similar to EMRs. Although both EMRs and EHRCs contain similar diagnostic and procedural information, the unavailability of open EMR datasets and other EHRC datasets hamper this important, quantitative comparison. In addition to diagnostic and procedural codes, EMRs also have richer content (e.g., doctor's notes, prescription records, etc.) and encode a 'clinical and medical context' that is very important for clinical decision making, which is not recorded within EHRCs.

Our study did not assign any meaning to the clinical procedure sequence patterns described here. We observed these patterns from analyzing the data, but there is no intrinsic feature within the claims that informs us about whether a particular sequence of procedures is novel or whether these sequences of procedures administered are valid for the patient (resulting in a cure or recovery). In addition, for our analyses, we did not consider any co-morbidities that may play a significant role in interpreting these results. The fact that the procedure sequence patterns extracted are short in length (typically up to a maximum of five CPT codes chained together) suggests that there is an inherent diversity in the way patients are treated by physicians. Perhaps within individual hospitals and clinics, there are well established guidelines and protocols for what sequences of clinical procedures are administered to patients, but we have not attempted to discover such patterns using our approach. Further, the lack of a database of clinical medicine practices and protocols makes it difficult to critically evaluate our results.

Finally, although we analyzed over a billion claims, we did not include any patient specific information (such as co-morbidities) or other clinically relevant parameters that may be useful for discovering these clinical procedure sequence patterns. Integrating such clinically relevant parameters (from both the claims and external data sources) and extracting meaningful results will be critical for advancing data-driven decision making processes within our healthcare delivery and practice. This work has been published in [95, 4].

Table 7: Summary of most commonly observed CPT sequences and associated costs with ASD treatment

HHS	Procedure Sequence	Cost (before)	Cost (after)
I	[Patient Evaluation]1→ [Neuromuscular Reeducation]2	14,456	55,619
	[Office Consultation]1→[EEG Monitoring]2	-	39,344
	[Office Consultation]1 → [Thyroid Stimulating Hormone Assay]2	-	16,541
	[Comprehensive Metabolic Panel * Thyroid Stimulating Hormone Assay]1→ [Comprehensive Metabolic Panel]2	-	9,524
II	[Office Consultation]1→ [EEG Awake and Sleep]2	37,006	100,116
	[Office Consultation]1→ [MRI of Brain w/o contrast]2	38,850	66,998
	[Hemoglobin Test]1→ [Office Consultation]2	8,607	12,200
	[Speech/ Hearing Evaluation]1→ [Office Consultation]2	-	25,938
	[Tuberculin Testing]1→[Office Consultation]2	-	11,612
IV	[Outpatient Visit]1→ [Typanometry]2	47,327	97,321
	[Occupational Therapy]1→ [Therapeutic Activities]2	17,6837	84,2534
	[Outpatient Visit]1→ [Chicken Pox Vaccine * Immunization and Administration]2	28,892	73,985
	[Outpatient Visit]1→ [CT Scan of Brain]2	30,277	55,754
	[Outpatient Visit]1→ [Electrocardiogram]2	NA	63,887
VI	[Psychotherapy Interview of patient]1→ [Family Pscychotherapy w/out patient]2	20040	65247
	[Hospital Discharge Day]1→ [Psychotherapy Interview of patient]2	-	11894
	[Psychotherapy Interview of patient]1→ [Psycho testing by technician]2	-	10071
X	[Speech / Hearing therapy]1→[Visual acuity screening]2	37,288	277,568
	[Speech / Hearing therapy]1→ [After hours medical services]2→ [Speech / Hearing therapy]3	25536	224905
	[Speech / Hearing therapy]1→ [Psycho testing]2	37,553	54,235
	[Initial hospital care]1→ [Inpatient con-	8,644	33,650

Table 8: Summary of most commonly observed CPT sequences and associated costs with HD treatment

HHS	Procedure Sequence	Cost (before)	Cost (after)
II	[Heart wall motion]1 → [Heart catheterization]2	32,971	87,538
	[Emergency dept visit]1→[Heart catheterization]2	-	388,749
	[Thromboplastin time partial *Assay of troponin]1 → [Assay of Troponin]2	-	14,648
	[Emergency dept visit]1 → [Heart wall motion]2	-	58,277
	[Heart catheterization]1 → [Doppler echocardiography]2	-	56,356
IV	[Doppler echocardiography]1 → [Cardiovascular stress test]2	49,024	120,435
	[CT scan of Head]1 → [Doppler echocardiography]2	33,729	75,402
	[Cardiac catheterization]1 → [Doppler echocardiography]2	33,262	515,622
	[Cardiac catheterization]1 → [ECG * Outpatient Visit]2	-	99,241
V	[Inpatient consultation]1 → [Subsequent hospital care]2	1,014,776	2,747,024
	[ECG]1 → [Subsequent hospital care]2	272,887	956,383
	[ECG]1 → [Inpatient consultation]2	113,244	252,084
	[Insertion of intracoronary stent]1 → [Subsequent hospital care]2	-	2,223,258
VI	[Insertion of intracoronary stent]1 → [Hospital discharge day]2	24,357	217,819
	[Echo transesophegal]1 → [Hospital discharge day]2	8,987	37,147
	[Intravascular coronary ultrasound]1 → [Hospital discharge day]2	8,418	104,663
	[Hospital discharge day]1 → [Glycosylated Hemoglobin Test]2	-	8,862

Table 9: Summary of most commonly observed CPT sequences and associated costs with BC treatment

HHS	Procedure Sequence	Cost (before)	Cost (after)
I	[Radiology port films]1 → [Radiation physics]2	35,380	590,310
	[Comprehensive metabolic panel]1 → [Chemotherapy infusion]2	6,435	84,093
III	[Chemotherapy infusion* Therapeutic prophylactic diagnostic injection]1 → [Chemotherapy infusion]2	-	187,291
IV	[Outpatient visit]1 → [Therapeutic prophylactic diagnostic injection]2	24,646	68,672
V	[Simple radiation therapy field setting]1 → [Complex radiation therapy field setting]2	94,685	297,814
	[Routine Venipuncture]1 → [Routine Venipuncture * Addition of drug]2	-	17,577
	[Set radiation therapy field]1 → [Radiotherapy dose plan]2	-	244,694
	[Set radiation therapy field]2 → [Special radiation dosimetry]3	-	40,527

Chapter V

TREATMENT BASED MODELS FOR EPILEPSY: PATHWAYS & EARLY IDENTIFICATION OF NON RESPONDERS

5.1 Introduction

Epilepsy is one of the most serious and common neurological disorders, with incidence rates of 50 per 100,000 individuals per year with sky rocketing effects for infants and the elderly [15, 55]. In epilepsy, the normal pattern of neuronal firing activity is perturbed, resulting in strange sensations, emotions, and convulsions [67]. The overall annual incidence of epilepsy cases ranges from 50-70 cases per 100,000 in industrialized countries to 190 per 100,000 in developing countries [128, 133]. Patients with epilepsy (PWE) suffer from decreased in quality of life, productivity, and life expectancy, and side effects of chronic medication. Comorbidities of epilepsy include but are not limited to depression, osteoporosis, fractures and increased mortality from suicide, accidents, vascular diseases,etc [88]. Limited clinical guidelines have been published to determine the common treatment pathways for epilepsy patients. Observational studies have been performed in which epileptic patients have been studied and analyzed for a period 11-12 years but no extensive understanding exists on how different AEDs are actually used in practice and potential reasons to switch to different treatment regimens [91].

Approximately 50% of PWE achieve seizure control with the first AED prescribed. Among those who continue having seizures after trying one AED, over the next 2-5 years another 13% become seizure-free with a second AED, another 4% after a second AED, and another 4% after trying 3 or more drugs [87]. Based on these observations, the International League Against Epilepsy (ILAE) has recently defined refractory epilepsy as “failure of an adequate trial of two tolerated, appropriately chosen and used AED schedules (whether as monotherapies or in combination) to achieve sustained seizure freedom.” Unfortunately, nearly one third of patients remain drug resistant or refractory to currently available AEDs. It is evident from some of the literature study that all patients should be treated in a personalized manner [40, 37, 81]. In case of epilepsy, patients are broadly characterized into refractory and non-refractory subtypes based on the response to antiepileptic medication. Non-refractory patients are the ones which have reduced seizure frequency with the first antiepileptic drug prescribed, whereas refractory patients fail to get respite from seizures even with multiple treatment regimens [57]. Refractory patients also known as drug resistant patients represent the minority of the epileptic population and bear the greatest economic and psychosocial burdens [88]. They endure side effects of multiple different AEDs for long periods compared with the majority of PWE who achieve seizure control with the first AED prescribed. Early identification of drug refractoriness allows patients and physicians to consider alternative more effective interventions early (e.g. surgery, or drug trials). Thus it important to identify the patients likely to be refractory as soon as possible [87].

We identify two main gaps in the study of epilepsy patients which are the following:

1. There does not exist extensive guidelines on making decisions with respect to AED choices when treating epilepsy patients. There are 20 different AEDs for clinicians to choose from when deciding course of treatment. The treatment choice depends on multiple factors such as patient’s age, side effects, type of epilepsy, etc [78]. A thorough multidimensional treatment pathway analysis is required to understand how epilepsy patients are treated in practice.

2. Refractory patients endure side effects of multiple different AEDs for long periods compared with the majority of PWE who achieve seizure control with the

first AED prescribed. Early identification of drug refractoriness allows patients and physicians to consider alternative more effective interventions early (e.g. surgery, or drug trials). Thus it important to identify the patients likely to be refractory as soon as possible [87].

To address the aforementioned problems, we propose the use of sequential pattern mining technique to generate frequent treatment pathways for epilepsy patients across different age groups and type of epilepsy and perform an exploratory analysis of the variations that exist in care given across the United States. We also propose a predictive model to detect whether or not a patient is likely to transition to refractory status at the time when they fail the first AED based on the medical history available at that time. The study makes use of an integrated healthcare dataset containing demographics, medications, diagnoses, procedures and encounter data for 1,376,756 epilepsy patients over a period of 10 years. The study follows a predictive modeling pipeline consisting of constructing an appropriate cohort, followed by feature construction and selection. Finally we evaluate and compare three candidate prediction algorithms: Logistic Regression, Linear Support Vector Machines (SVM), and Random Forest. We evaluate prediction models using cross-validation to obtain unbiased estimates of model performance

5.2 Data

Medical claims data including diagnosis, procedures and pharmacy claims spanning a period of 10 years ranging from January 2006 till December 2015 was collected from different regions of the United States by IMS Health Surveillance Data Incorporated (SDI) medical database. This data was the best fit for the study since it incorporates patients from geographically dispersed regions along with third party and government payers. Since the database does not require patients to be continuously associated with a single plan, we can keep track of a patient even if he switches to a different

Table 10: Data Statistics

Metric	Count
No. of Patients	1,376,756
No of Pharmacy Claims	28,403,939
No. of Diagnosis Claims	173,570,273
No. of Inpatient Encounters	201,202
No. of Outpatient Encouters	707,332
No. of ER Encounters	432,915
No. of Anti-epileptic Drugs	20

plan due to varying socioeconomic status [46]. Table 10 shows some basic statistics calculated based on the raw data for epileptic patients. The data consists of 23 AEDs out of which 3 are considered rescue medications and would not be treated as anti-epileptic drugs. Table 11 shows the complete list of AEDs used in the study.

5.2.1 Data Processing

Data processing and standardization is one of the most important and time consuming steps when building predictive models since we need to be certain that the data being used for analytics is as accurate as possible. For the purpose of this study, we removed clinically irrelevant gaps between consecutive prescriptions of the same medication or merge them if they are overlapping. We also group diagnosis, medical procedures and medications into clinically meaningful categories for better interpretation. We elaborate on these processes below.

5.2.1.1 Prescription processing

We eliminate clinically insignificant gaps between consecutive prescriptions of the same medication and also get rid of irrelevant AED prescriptions which may have been introduced in the treatment regimen to treat potential minor side effects. The scenarios for such a step have been formulated after an extensive review carried out by the clinicians. Here we explain the steps required for processing prescriptions which need to be carried out in the sequence in which they have been laid out. We also

Table 11: Anti-epileptic Drugs and Rescue Medications

Anti-Epileptic Drugs	**Generation**
Carbamazepine	I
Primidone	I
Valproate Sodium	I
Phenobarbital	I
Ethosuximide	I
Phenytoin	I
Ethotoin	I
Methsuximide	I
Ezogabine	II
Lacosamide	II
Oxcarbazepine	II
Felbamate	II
Lamotrigine	II
Rufinamide	II
Vigabatrin	II
Levetiracetam	II
Tigabine HCL	II
Zonisamide	II
Topiramate	II
Clobazam	II
Rescue Medications	
Diazepam	I
Lorazepam	I
Clonazepam	I

show illustrations of the same in figure 10 :

1. **Elimination of Small Gaps [figure 10(a)]:** A small gap refers to gap between two consecutive prescriptions of the same drug which is less than twice the days of supply of the former. In this case the former prescription is extended to end on the service date of the latter.

2. **Elimination of Overlapping Prescriptions [figure 10(b)]:** This is the case when there are two consecutive prescriptions which overlap for certain number of days. The two prescriptions are merged by shortening the former so that it ends on the service date of the latter. Once the overlap is removed it becomes a continuous prescription which can be further processed as explained in Step 4.

3. **Elimination of Adjacent Gaps [figure 10(c)**]: This refers to the case when there are two consecutive gaps between prescriptions of the same drug within 90 days or less. The former gap is closed by extending its last prescription.

4. **Merging of Continuous Prescriptions [figure 10(d)**]: This is the case when two consecutive prescriptions occur without a gap.i.e the end date of the former prescription is the same as the start date of the latter. We merge the prescriptions so that the former prescription ends on the end date of the latter.

5. **Elimination of Short Prescriptions [figure 10(e)**]: After the aforementioned steps have been executed and all the irrelevant prescription gaps have been closed, we eliminate the prescriptions which are less than 30 days since in epilepsy treatment, a prescription given for such a short period is considered irrelevant.

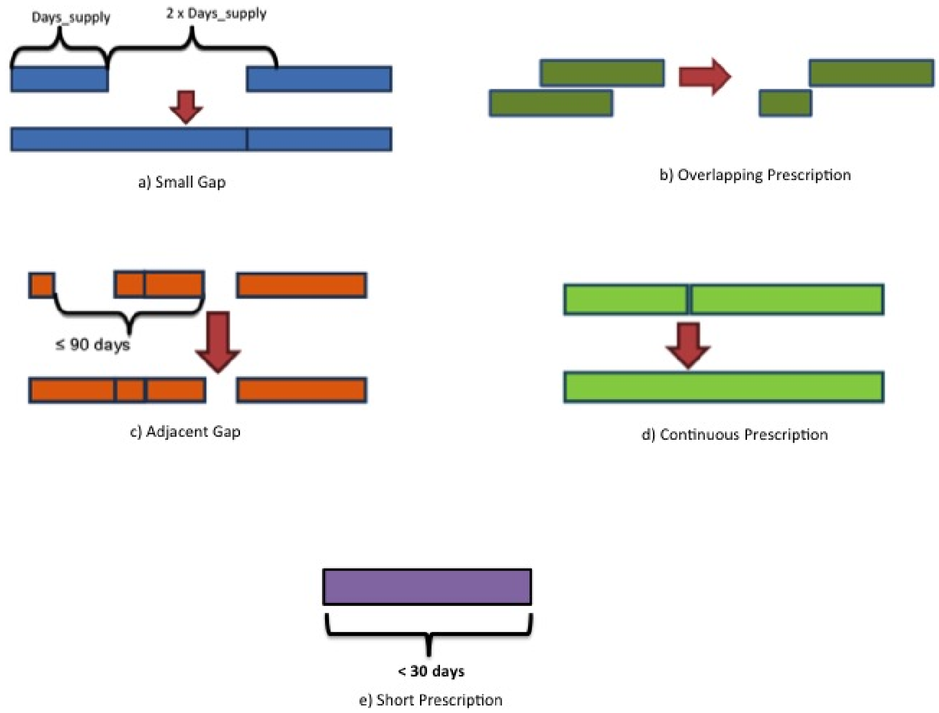


Figure 10: An illustration of prescription processing scenarios.

5.2.1.2 Grouping of Diagnosis, Procedure Codes and Medications

Most of the healthcare datasets have diagnosis and medical procedures coded by standard systems of classification such as the International Classification of Diseases and Related Health Problems (ICD) and Current Procedural Terminology (CPT) [138]. Both CPT and ICD-9 codes help in communicating uniform information to the physicians and payers for administrative and financial purposes but for analytics we group these codes into clinically significant and broader codes presented by another scheme of classification named Clinical Classification Software (CCS) maintained by Agency for Healthcare Research and Quality (AHRQ) [39]. The single level scheme consists of approximately 285 mutually exclusive diagnosis categories and 241 procedure categories. We map all the ICD-9 and CPT codes in our dataset to corresponding CCS codes and use them for constructing appropriate features for our model. Some of the codes which do not have a corresponding CCS code are discarded. For grouping medications, we refer to the USP Medicare Model guidelines v6.0 which makes use of the pharmacotherapeutic evidence within the context of FDA approved indications to classify medications into broader categories. The USP scheme consists of 146 classes of medications.

5.3 Methods

5.3.1 Sequential Pattern Mining for Epilepsy Treatment Pathways Construction

Clinicians can choose from 20 different AEDs when deciding treatment regimens for epilepsy patients. Although majority of the patients become seizure free with the first AED, there exist patients who are prescribed more than one AEDs depending on the type of epilepsy, patient's age, side effects of medications and other comorbidities [41]. The phenomena of prescription of subsequent AED is categorized as a failure of the former AED.

Definition of AED Failure: An AED treatment for a patient is said to have failed if the patient is prescribed another AED as a replacement of the current AED or as an addition to the ongoing treatment.

In this study we develop popular treatment pathways consisting of AED prescriptions as monotherapy . A treatment pathway consists of two or more AEDs prescribed commonly in a particular sequence. We perform this analysis to explore AED failure patterns across different age groups and and types of epilepsy to assess the variation in treatment routes. We visualize frequent routes of treatment using sequential pattern mining to mine patterns from data occurring above a predetermined threshold frequency. In case of epilepsy treatment , we are interested in analyzing AED prescription patterns to understand how AEDs were prescribed in practice by clinicians. To accomplish this, we use constraint based sequential mining [97] to restrict the extraction of frequent treatment patterns consisting of consecutive occurrence of

AEDs following a pattern in a minimum threshold number of patients.. The approach represents the treatment data as a directed graph explained in detail in the previous chapters with patients and AEDs as nodes and edges between the AED nodes signifying the sequence of prescribed drugs. The generation of a treatment pattern from such a graph is guided by the number of patients who are prescribed that particular pattern.

5.3.2 Predictive Modeling - Cohort Construction

Cohort Construction is the foremost step when developing a predictive model which involves defining a sample of patients to be studied which meet some criteria relevant to the problem at hand. The criteria in our study is carefully designed by the domain experts and involves setting an index date for every patient. The choice of index date is crucial for any study in predictive analytics since it defines a dividing point in the timeline of a patient, the period before which qualifies to be the observation period and the period after, becomes the evaluation period. Our goal in this study is to find the patients who have not benefited from the 1st AED prescribed to them and are likely to refract . To accomplish this we need to set the index date for a patient to be the date of failure of his first AED and the patient's medical history is scanned till one year prior to this date. The population we are studying consists of an adult epileptic set of patients with some additional inclusion and exclusion criteria carefully and extensively formulated by clinical experts and are laid out below. We also show the impact on the cohort size as result of applying these criteria in figure 12.

1. **Epilepsy Diagnosis criteria:**

   (a) The patient should have at least one diagnosis claim of 345.* (icd 9 code for epilepsy diagnosis) or 2 claims of 780.39 (icd 9 code for convulsions) at anytime in the timeline of the patient.This criteria is to make sure we exclude all the patients which have not been diagnosed with any form of

Table 12: Impact of inclusion & exclusion criteria on the cohort size.

Inclusion Criteria	**Percentage of Patients Retained**
Epilepsy Diagnosis Criteria	100 %
AED Failure Criteria	50.6 %
Age Requirement	44.4 %
Data Quality Criteria	13.3 %

epilepsy and may have had less than one convulsions which does not constitute substantial evidence to categorize a patient to be epileptic patient.

(b) The patient should have at least one AED prescription at anytime in the timeline.

(c) The first AED prescribed should be a monotherapy and should have been given for at least 60 days.

2. **AED failure criteria:** The patient should have at least one failure of AED i.e a patient should have at least two AED prescriptions.

3. **Age requirement:** Infants and teenagers in their early teens have been excluded from the study by enforcing a minimum age criteria of 16 years at the time of their first AED failure.

4. **Data quality criteria:** To avoid including patients not complying with the drug prescriptions, it is required for the patient should be active with respect to pharmacy claims in every quarter of the year.

5.3.3 Target Variable Definition

The aim of this study is to predict patients with refractory epilepsy at an early stage to discover factors contributing to such a state. Ideally the best possible way to categorize a patient as refractory is by monitoring the seizure frequency over time but since seizures are not captured in the claims data, we use the number of AEDs tried on the patient as a proxy measure for refractory status [45, 21]. he raw data after being processed and funneled through the aforementioned multiple inclusion and exclusion criteria has 183,291 patients who have failed at least 1 AED. To maintain a clean distinction between refractory and non refractory we categorize refractory patients to be the ones who have failed at least 3 distinct AEDs.

Case patient definition: We consider patients as cases or refractory who failed 3 or more AEDs. With this criteria we have 28,516 patients as cases.

Control patient definition: Control patients are the ones who have failed exactly one AED i.e they should have exactly two distinct AED prescriptions. There exists 46,285 such patients.

5.3.4 Feature Selection and Engineering

The initial set of features consist of 1,438 features extracted from the observation period of every patient excluding any information about the first prescribed AED. The observation period refers to the 1 year period before the index date. A comprehensive list of features has been included in the supplementary section. The features have been categorized into 5 different types:

1. **Demographics:** This set has all the features representing basic demographics of the patient such as age, gender and the geographic information of the patient.
2. **Comorbidities:** This particular set of features includes features corresponding to the different comorbidities associated with epilepsy such as migraine, sleep related disorders, disorders, different kinds of mental disorders etc. Some of these comorbidities are very specific such as ‘Migraine’ which is trivial to determine by looking for the appropriate diagnosis code in the data. There also

exists some generic comorbidities such as ‘Serious Mental Illness’ which is determined by the presence or absence of mental illness related disorders such as psychosis, bipolar disorders, etc. which in turn may have a range of diagnosis codes associated with them. This feature set also involves comorbidity index scores such as the Charlson Comorbidity Index [34] and Epilepsy Comorbidity Index [144] which are quantitative indications of the health of the patients.

3. **Policy**: This category of features includes the insurance payer information of patients since the type of payer represents the socio-economic status of the patients which in turn may affect the care provided to them.

4. **Treatment:** Factors representative of the treatment regimen and medical procedures undertaken by patients are included in this domain. The medications have been grouped into higher level categories based on their therapeutic categories laid down by the U.S.Pharmacopeial Convention which is one of the most popular drug groupers used in the recent times.

5. **Hospital Encounters**: Medical encounters and details about patient visits are represented by this domain such as type of visit, length of stay, etc. Various checks for occurrence of seizures using diagnosis codes as proxies and monitoring of hospital and pharmacy activity of every patient are also recorded and used as features.

A comprehensive list of features are shown in table 13

In this study we build a predictive model which trains on the observation period before the index date and predicts whether the patient would eventually become refractory or remain non-refractory at the point when he fails his first AED. Since the target variable in our study is a binomial variable, we use machine learning classifiers such as Linear Support Vector Machine (SVM), Logistic Regression and Random Forest tuned appropriately for the purpose of training the model. A subset

Table 13: Feature Summary and Statistics for the predictive model. All the features are calculated in the 1 year period before the index date. The letter 'X' represents a particular value associated with the corresponding feature

Category	**Feature_Desc**	**Data type of feature**	**No of features**
Demographics	1st digit of zip code	boolean	10
	Age at the time of first AED failure	boolean	3
	Gender	boolean	1
	Total		**14**
Comorbidity	Affective Disorder	boolean	1
	Neurological comorbidity	boolean	1
	Substance abuse	boolean	1
	Epilepsy Comorbidity Score	integer	1
	Cardio condition	boolean	1
	Diagnosis CCS code X	boolean	283
	Sleep disorder	boolean	1
	Porphyrin metabolism disorder	boolean	1
	Osteoporosis	boolean	1
	Other mental disorder	boolean	1
	Autoimmune disorder	boolean	1
	Charleson Comorbidity X	boolean	16
	Obesity	boolean	1
	Mental Retardation	boolean	1
	Liver Condition	boolean	1
	Diabetes	boolean	1
	Renal Insufficiency	boolean	1
	Serious Mental Illness	boolean	1
	Epilpesy related comorbidity X	boolean	6
	CCI Score value	integer	1
	Total		**322**
Policy	Payer X	boolean	4
	Total		**4**
Treatment	Medical procedure performed within 30 days before the index date	boolean	1
	Treatment with medication class X within 30 days before the index date	boolean	3
	Prescription of medication class X	boolean	140
	Procedure CCS code X	boolean	241
	Procedure CPT code (which has no corresponding CCS code)	boolean	706
	Total		**1091**
Hospital Encounters	Occurrence of seizure based on icd 9 code 345.X or 780.39	boolean	1
	Occurrence of seizure based on icd 9 code 345.X only	boolean	1
	Hospital encounter	boolean	1
	Hospital encounter within 30 day period before the index date	boolean	1
	Number of months of diagnosis claims activity	integer	1
	Number of months of pharmacy claims activity	integer	1
	Number of months of hospital activity	integer	1
	Total		**7**

Table 14: Case - Control Statistics

Type of Dataset / Class	Case	Control
Training	23,777	44,151
Hold Out Test	4,739	2,134

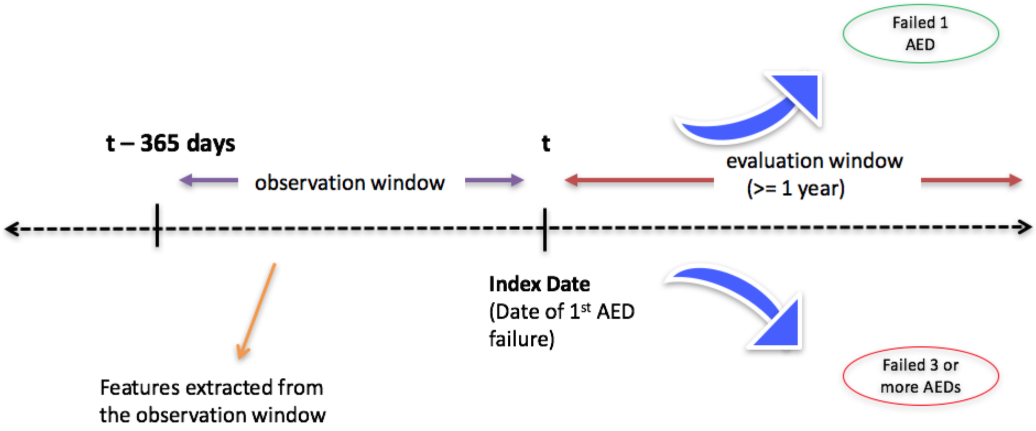

Failed 1 AED
t – 365 days
t
observation window
evaluation window (>= 1 year)
Index Date (Date of 1st AED failure)
Features extracted from the observation window
Failed 3 or more AEDs

Figure 11: Experimental setup for predictive modeling

of patient data called the training set is used to train the model. The rest of the data is used as a hold out test set which is used to objectively assess the predictive power of the trained model and is never used for training at any point in time including the feature selection phase. Table 14 shows the number of case and control patients in each of the 2 sets. The feature matrix consisting of both raw and engineered features is subjected to a feature selection process using ANOVA F-value which scores the features based on a univariate F-test. We only select a subset of the high scoring features found to be sufficient for prediction during parameter tuning to be used by the classifier. The evaluation period for this predictive model begins immediately after the index date and extends up to the last record of the patient. We use 3-fold cross validation on the training set to tune the parameters and finally test the best model from cross validation on the hold out test set. The figure 11 shows a visualization of the experimental setup.

5.4 Results

In this section we report the insights obtained upon construction and analysis of treatment pathways for anti-epileptic drugs across various age groups, providers, and type of epilepsy. We also present the details of the experimental predictive model developed using the aforementioned setting using features from different categories as explained before. We tune the appropriate parameters of the classifiers used in our study along with the number of top scoring features for the model to find the optimum experimental parameter setting having the best performance. To assess the performance of this predictive model, we compare it with a baseline model consisting of only demographic information such as age, gender and geographic location of patients.

5.4.1 Analysis of AED Treatment Pathways

In this analysis we select only those patients which have been conclusively diagnosed with epilepsy based on the epilepsy diagnosis criteria mentioned in section 5.3.2 and are at least 16 years of age at the time of their first visit. They are also required to have at least two distinct AED prescriptions with the first AED prescription being a monotherapy which is prescribed for at least 60 days. Figure 12 shows a sunburst visualization of the frequent treatment pathways based on an extensive analysis of 529,483 patients satisfying the aforementioned selection criteria. The drugs shown in the area between the inner circle and the first concentric circle constitute the first line of treatment. The drugs that follow the first line of treatment are shown between the first and second concentric circle as denoted in the figure. By analyzing the first line of treatment, there does not seem to exist any one particular drug which distinctly stands out and can be categorized as the treatment of choice irrespective of the patient's age and type of epilepsy which corroborates the fact that there is no universally accepted standard of care for epilepsy [41]. However the top 3 most

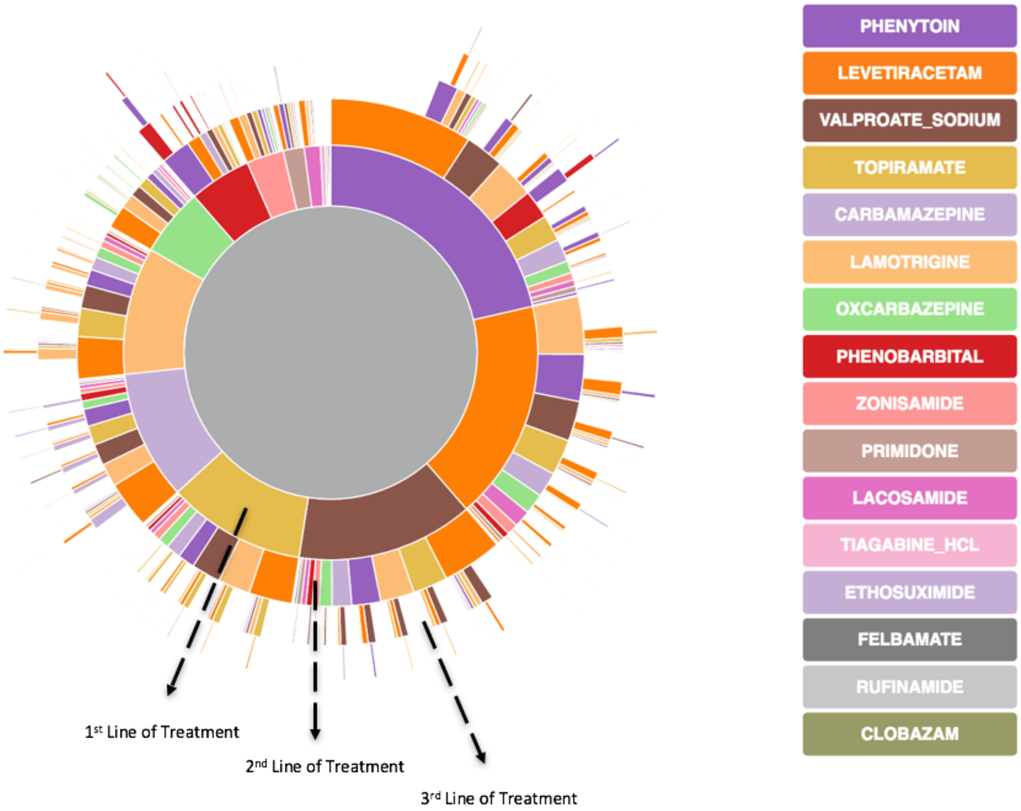


Figure 12: Treatment pathways of epilepsy patients

frequently used 1st line of care consists of AEDs Phenytoin, Levetiracetam and Valproate Sodium. Levetiracetam, Lamotrigine, Valproate Sodium and Topiramate are the popular choices of drugs in the 2nd line of treatment. Majority of the patients prescribed Phenytoin or Valproate sodium as the first drug are observed to switch to Levetiracetam, whereas most patients prescribed Levetiracetam in the first line of treatment switch to Lamotrigine or Phenytoin.

5.4.1.1 Treatment pathways across age groups

Based on existing treatment guidelines for epilepsy patients we know that some of the important factors to consider when making treatment choices for epilepsy patients is their age and the type of epilepsy diagnosed [41]. The adult epileptic patients primarily fall into 3 different age groups: 16-45 years, 45 - 65 years , 65 years and above. There are times when AEDs that work well as the 1st line of treatment for

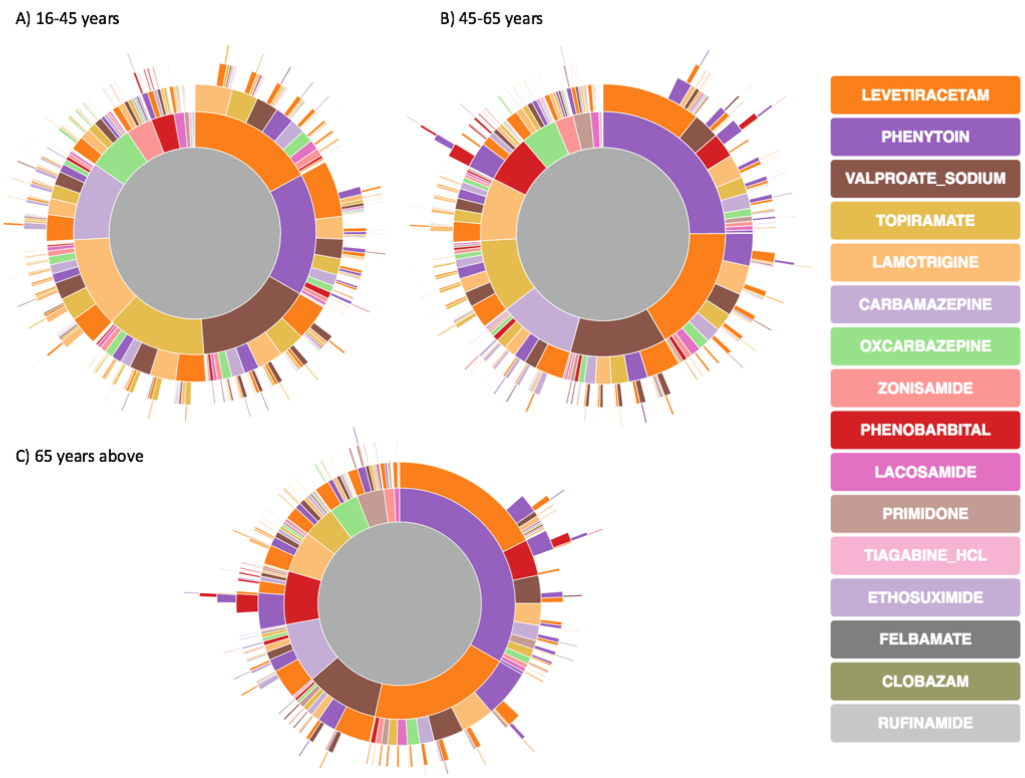


Figure 13: Treatment pathways for various age groups A) 16-45 years B) 45-65 years and C) 65 years and above

young adults may not be the best choice for the older epileptic population. Figure 13 shows the treatment pathways for the aforementioned age groups.

The visualization suggests that Levetiracetam is the most popular choice amongst the AEDs as the first line of treatment for patients in the age group of 16-45 years. With higher age groups clinicians prefer to begin treatment with Phenytoin, whereas Levetiracetam is the second most popular choice as the first drug followed by Valproate Sodium. Lamotrigine and Topiramate are also used as the first line drug for younger patients in the age group of 16 - 45 but for older patients Carbamazepine is preferred over them. For the second line of treatment, Levetiracetam, Valproate Sodium and Lamotrigine are the common choices of AEDs.

5.4.1.2 Treatment pathways based on type of epilepsy

The type of epilepsy diagnosed for patients has also been reported to be an influential factor in determining the treatment plans for patients. Idiopathic Generalized Epilepsy (IGE) is diagnosed when patients experience electrical impulses throughout the entire brain whereas Symptomatic Localization Related Epilepsy (SLRE) epilepsy involves seizures affecting only one hemisphere of the brain [62]. We categorize the patients into 2 cohorts based on the type of epilepsy diagnosed which is identified by the first occurrence of the corresponding ICD-9 diagnosis code. Figure 14 shows the sunburst visualization of the treatment pathways for the two types of epilepsies.

For patients diagnosed with IGE the clinicians prefer to recommend Valproate Sodium over Lamotrigine or Topiramate. In the visualization shown, it is observed that Valproate Sodium is amongst the top 3 AED choices for the first line of treatment preceded by popular choices Phenytoin and Levetiracetam whereas Lamotrigine and Topiramate are less preferred than aforementioned drugs which corroborates the expert recommendations. For second line of treatment for IGE patients clinicians have preferred choices of AEDs based on what drugs were prescribed in the first line. From the data we observe that best choice of AEDs after prescription of Valproate Sodium are Levetiracetam and Lamotrigine. Majority of the patients on Lamotrigine in the first line of treatment are observed to be treated with Levetiracetam, while comparable number of patients are treated with either Valproate Sodium or Topiramate. For patients prescribed Topiramate in the first line of treatment, the top contender for second line treatment is Levetiracetam succeeded by Lamotrigine and Valproate Sodium.

In the case of SLRE, the clinicians prefer Carbamazepine, Levetiracetam, Oxcarbazepine, Phenytoin, Topiramate and Valproic Acid when deciding the first line of treatment. In figure 14 the visualization for SLRE shows the use of the aforementioned medications as the preferred choices for the first line of treatment although a

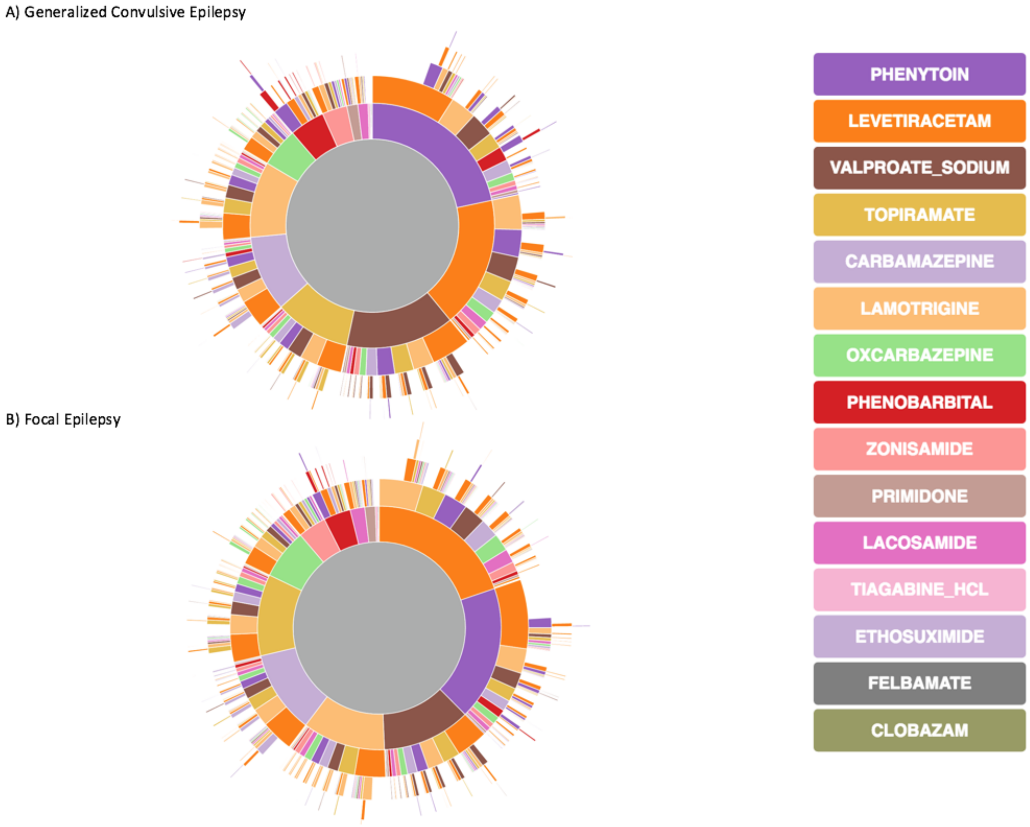


Figure 14: Treatment pathways for A) Generalized Convulsive Epilepsy and B) Focal Epilepsy

lot of variation in the second line of treatment. It has been observed that Levetiracetam which is the most popular choice as the first prescribed AED in case of SLRE is followed primarily by Lamotrigine, whereas Phenytoin, Carbamazepine, Lamotrigine and Valproate Sodium are all followed primarily followed by Levetiracetam with Lamotrigine being the next best choice as a second line drug which is in alignment with recommendation from experts as well.

5.4.2 Temporal Variation of Monotherapies

The antiepileptic drugs are categorized as 1st generation or 2nd generation depending on the time when they were approved to be used. The 1st generation drugs as shown in table 2 have been in the market for a very long time whereas the 2nd generation drugs were approved in the early nineties. The 1st generation drugs were found to

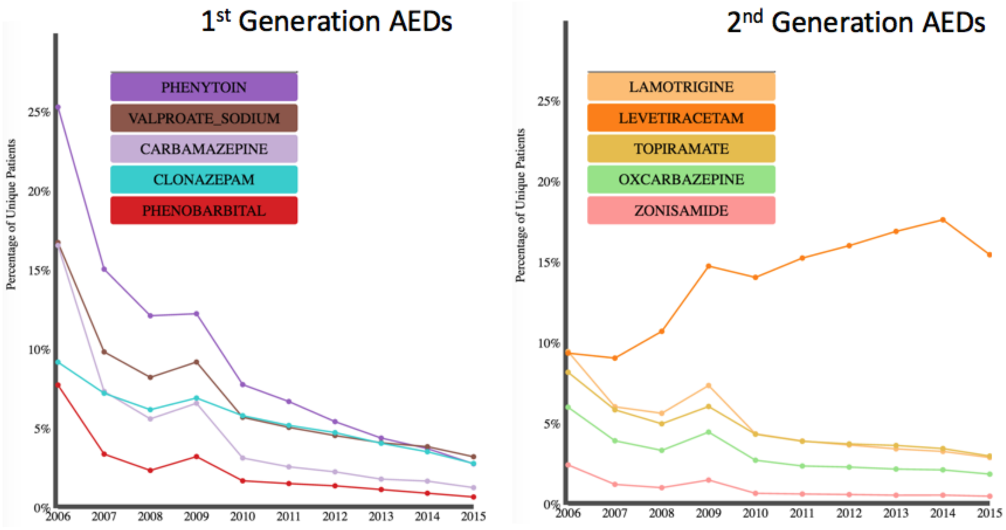


Figure 15: Temporal Variation of 1st and 2nd generation drugs as the first and 2nd line of treatment

have low efficacy towards refractory epilepsy which led to the development of newer AEDs which were termed as the 2nd generation AEDs. We analyze the temporal variation of both old and the newer drugs across the 10 year period of 2006 to 2015. Figure 8 shows such a temporal variation for top 5 frequent drugs of each generation as the first line of treatment. The first generation drugs seem to have been the popular choice of drugs in 2006 but had a steep descent in the year 2007 in which the 2nd generation drugs gained popularity due to their promising effect on refractory patients. A peak is noticed in the year 2009 for the both the generations but with increasing popularity of the 2nd generation drugs, the number of patients on the first generation drugs began to decrease after that [10].

5.4.3 Predictive Modeling - Parameter Tuning & Quantitative Analysis

The primary choice of parameter to be tuned for linear SVM and Logistic Regression is the C value which specifies the regularization strength. Random forest on the other hand is an ensemble learning method for classification and operates by constructing a multitude of decision trees based on the training data and assigns the class that is the mode of the classes of the individual trees in the forest [72]. The number of trees selected if optimally selected would increase the likelihood of obtaining accurate predictions. An important parameter which we use for both the classifiers is the class weight and use the 'balanced' value for the same. This parameter is used in studies like the current one in which the classes are highly unbalanced. We have a case to control ratio of 1:3 and this parameter helps in penalizing the assignment of the majority class. For this study we vary the C value of SVM and Logistic Regression from 0.00001 to 1 and the number of trees for random forest from 150 to 300. The other parameter we vary is the number of features used as input to the model. As mentioned before, we perform feature selection using ANOVA and rank the features in the order of importance towards predicting the target. We vary the percentage of features from 1 to 100 percent for each set of top features we vary the classifier parameters. The goal is to find the least number of top ranked features giving the best predictive performance using the most appropriate set of parameters.

The distribution of AUC values obtained by the parameter tuning process for three classifiers for the experimental model are shown in figure 16. With the SVM and the Random Forest classifier, we observe an AUC of 0.70 using top 4% of the features while with Logistic Regression, the highest AUC obtained is of 0.68. Based on our observation the AUC for Random Forest and SVM does not improve on increasing the number of features used by the model, so the maximum features we use for our model and analysis is the top 4%. With the baseline model these classifiers report an AUC of 0.58. Figure 17 shows the area under the ROC curve for the experimental and the baseline model.

To further assess the quality of the classifiers used in this study, we show calibration reliability plots in figure 18 for SVM, Random Forest and Logistic Regression with respect to the experimental model. The diagonal in the plot simulates a perfectly calibrated classifier which has predicted probabilities of classifying certain positive

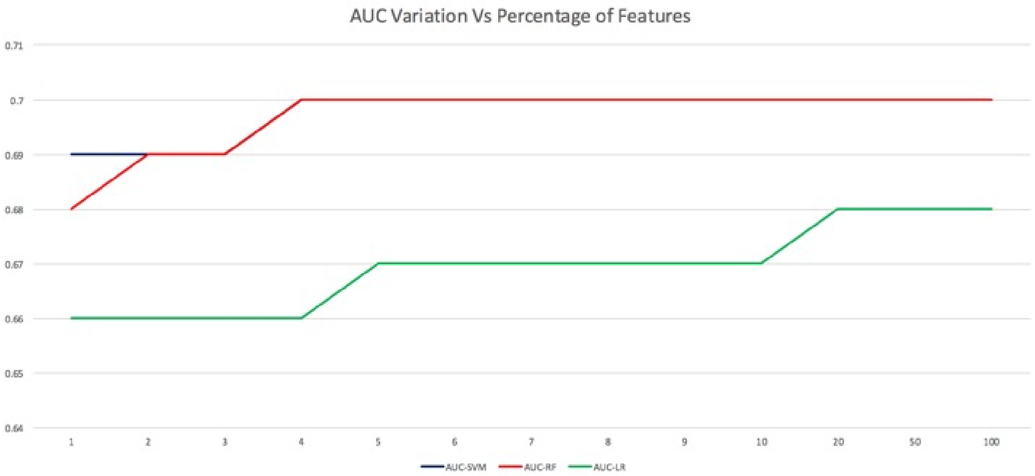


Figure 16: AUC variation with varying number of features for SVM, Random Forest and Logistic Regression

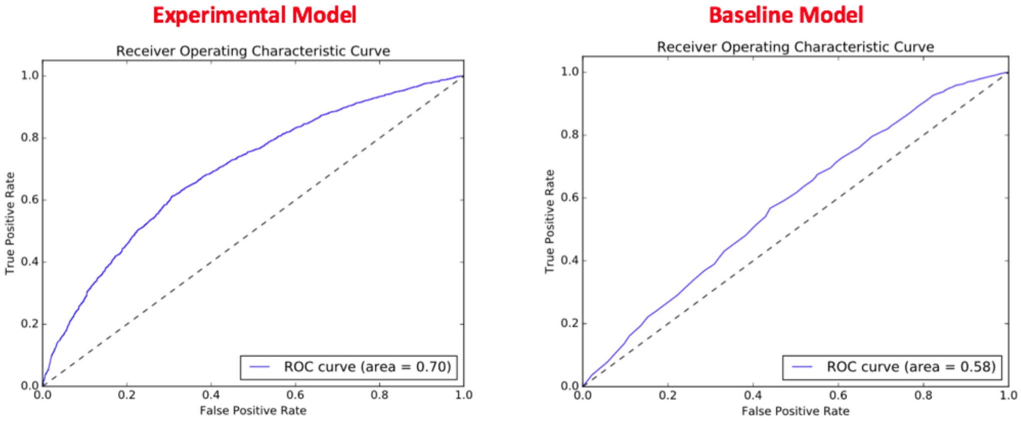


Figure 17: Area under the ROC curve for the experimental and the baseline model

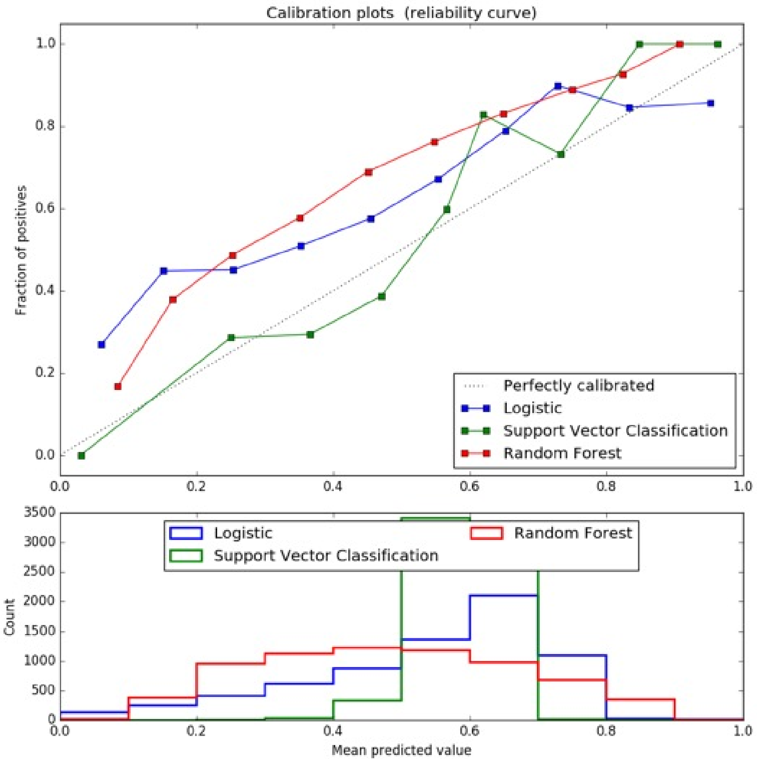


Figure 18: Area under the ROC curve for the experimental and the baseline model

instances comparable to the actual number of positive samples. A well calibrated classifier would have the curve as close as possible to this diagonal. We observe that out of the three classifiers used in the study, SVM is the best calibrated one and thus is the most reliable based on the data being used.

5.5 Conclusion

In this chapter, we perform an extensive analysis of treatment pathways of epileptic patients to gain insight into various factors that are taken into consideration when choosing AEDs for treatment. We also perform predictive analytics using claims data of epilepsy patients to develop a predictive model with the goal of not just identifying patients which do not respond to the existing anti-epileptic drugs in the market but to identify them at an early stage. We choose to predict them at the time of their first failure to prevent them from undergoing unnecessary medications which may not have any positive in influence on the patient's health. The model reported an AUC of 0.70 with features ranging from demographics and comorbidities to patient ecosystem and encounters. The results show that the model is trustworthy and has sufficient predictive power to predict refractory patients within a population of epileptics.

Chapter VI

INTERPRETABLE CLINICAL PREDICTIVE MODELS VIA ONTOLOGY GUIDED FEATURE CONSTRUCTION

6.1 Introduction

The core of the medical data extracted from claims datasets or directly from electronic health records (EHR) consists of information about patients treatment history which involves among other things, recording different diagnoses made by the patient's physician during various visits. The data recorded in a patient's electronic health record is in the most granular form to make it suitable for diagnostic, billing and reporting purposes using standard codes to accurately identify diseases, disorders, symptoms, adverse effects of drugs and chemical injuries, etc. for a patient. These codes play a significant role in characterizing a patient and assist the clinicians in deciding the future trajectory of treatment. The International Classification of Diseases (ICD) is one such standard coding system, maintained by WHO and is designed as a healthcare system with the goal of providing a set of codes to classify diseases based on minor nuances of signs, symptoms, abnormal findings, etc [138].

The ICD-9 code system is a collection of trees which represents relationships between different ICD-9 codes. Even though the ICD-9 codes have a hierarchical structure amongst themselves, they have been well placed in rich and well maintained medical ontologies such as SNOMED CT [25] and AHRQ's Clinical Classification Software (CCS). SNOMED CT is a consistent and processable representation of a wide variety of medical terms including clinical codes, synonyms, definitions, etc. interrelated with one another. The CCS classification on the other hand is focussed only on diagnosis and procedure codes and provides a way to classify them into a limited number of categories by aggregating individual ICD-9 codes into broad diagnosis and procedure groups in order to facilitate analysis and reporting. CCS provides two related classification systems. The first system called the *single-level CCS* classification groups the diagnosis and procedure codes into 285 and 231 mutually exclusive categories respectively. The other system is more detailed than the single-level CCS and is referred to as the *multi-level CCS* classification and reorganizes the single-level CCS by adding additional levels of grouping which helps clinicians in interpreting the medical condition of patients clearly [48]. Since most datasets use the Current Procedural Terminology (CPT) [13] scheme to code for procedures and services and CCS doesn?t maintain a multi-level classification system for CPT codes, we only focus on diagnosis codes in our work.

Figure 19 shows an illustration of the multi-level CCS hierarchical classification system for diagnosis codes using level 1 CCS code '7' representing diseases of the respiratory system . We have not shown the expansion of all the parent nodes shown in the figure due to space constraints. The level 1 codes offer a bird's eye view of the actual disease condition. The level 2 codes expand into multiple codes which are more specific than level 1 but generic when compared to level 3. As we go deeper in this hierarchy the codes become more specific with respect to the disease representation with the last level being level 5 consisting of the raw ICD-9 codes. In the illustration shown in figure 19, the level 1 code '7' expands into specific codes such as '7.1' which represents hypertension. The '7.1' code further branches into '7.1.1' and '7.1.2' representing essential hypertension and secondary hypertension respectively. At some levels, branching may or may not occur, e.g '7.1.1' directly splits into level 5 ICD-9 codes '401.1' and '401.9' whereas '7.1.2' further branches into '7.1.2.1' and '7.1.2.2' which in turn branch out further at level 5.

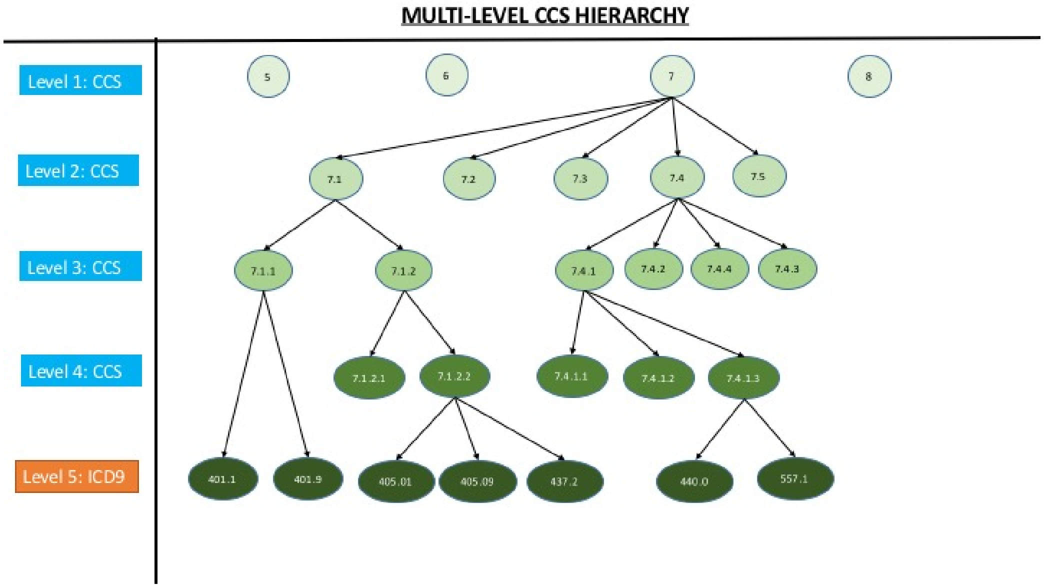


Figure 19: An illustration of Multi-level CCS classification of level 1 code '7'.

Predictive models developed in the area of healthcare do not leverage the hierarchical nature of the medical ontologies and are designed to represent ICD-9 codes in their raw form as features where each code is a binary feature indicating the presence or absence of a diagnosis code. This approach leads to unmanageably large number of features which are not feasible to analyze and interpret by clinicians. On the other hand grouping the raw ICD-9 codes using the single-level CCS classification helps in reducing the feature dimensionality but over-generalizes the features to such an extent that we lose the specificity of certain relevant diagnosis codes. The challenge in using guidance from a hierarchical knowledge base such as CCS for feature construction lies in selecting the optimal depth for each medical code in the hierarchical tree such that they are clinically interpretable and do not compromise on the predictive power in the model. To address this issue, we propose an approach to leverage the *multi-level CCS* hierarchy to optimally group the diagnosis codes up to a particular level in the hierarchy to be used in the predictive model and preserve the predictive nature of the same.

Our study and results make the following contributions:

1. We leverage the concept of information gain to exploit the hierarchical nature of medical ontologies for dimensionality reduction.

2. We evaluate our approach by developing predictive models for early identification of refractory epilepsy patients, predicting patients likely to get admitted for asthma and predicting mortality after discharge for ICU patients using the grouper codes as features selected by our algorithm and comparing their prediction power and interpretability with the model developed using raw ICD-9 diagnosis codes as baseline.

6.2 Background and Significance

ICD-9 codes have been used for analysis in a lot of work involving predictive modeling and detection of favorable events [74, 46]. Perotte et al utilize the ICD-9 hierarchy to assign appropriate diagnosis codes to discharge summary notes. They leverage the hierarchy to create an augmented label set for each document and train multiple SVM classifiers one for each code [115]. Yan et al have also exploited prior knowledge from inter code relationships in an ICD-9 hierarchy for assigning medical codes to patient visits [156]. Contrary to this work, Singh et al. leverages the hierarchical nature of ICD-9 codes for feature construction in a predictive analytics setting using a probability based approach. They analyze each subtree associated with a particular top level ICD-9 code in a hierarchy and map each code in the subtree to a conditional probability of outcome using the training data and thus for each top level code in the subtree, the code with the maximum probability is picked to be the feature [135]. Using this approach the dimensionality of the feature vector is the same as the number of top level codes in the hierarchy since it eliminates all the codes which have probability less than the maximum. This could potentially result in an inaccurate feature set since more than one diagnosis codes belonging to a subtree of a top level code could have probability values close to the maximum probability and could influence the prediction as well. There has also been research done in leveraging the structure of predictor variables by incorporating the information about the same in learning algorithms. Supervised group lasso (SGLasso) is developed on top of lasso which is an individual variable selection method and group lasso which is designed to select only cluster of covariates. SGLasso is the first of its kind which penalizes the individual variables within a cluster to select the optimal variables using lasso after which it uses group lasso to select the important clusters. This method has been used in the context of studying associations between gene expression and progression of common diseases such as heart disease or cancer [94]. The drawback of this approach is that it does not consider the hierarchical nature of the clusters for variable selection . Another method similar to SGLasso is the overlapping group lasso which factors in the possibility of groups of features having overlaps and can encode the hierarchical relationship between groups with overlapping patterns [157, 159, 135]. The hierarchical variable selection is performed via regularization in which a node in a tree is selected only if the parent node is selected. To the best of our knowledge, none of the aforementioned approaches have been applied in the context of leveraging the hierarchical nature of ICD-9 codes for feature selection.

6.3 Methodology

6.3.1 Entropy and Information Gain

Decision tree learning is one of the most popular and widely used tree based classification approach for developing predictive models. It consists of internal nodes and branches where the nodes specify a test for the value of particular feature of the data and each branch corresponds to the outcome of the test performed at the node and is eventually connected to the next node. The leaf nodes have positive and negative labels corresponding to the target variables. Classification performed by decision trees is based on the concept of *information gain.* In order to define information gain precisely, we begin by defining a concept known as *entropy* which characterizes the impurity of an arbitrary collection of examples [118].

Definition 6.3.1. *Entropy:* Given a collection S, containing positive and negative examples of the target concept, the entropy of S relative to this boolean classification is

$$H(S) = -P_{\oplus} log_2 P_{\oplus} - P_{\ominus} log_2 P_{\ominus} \tag{1}$$

where $P_{\oplus}$ refers to the proportion of positive examples in S and $P_{\ominus}$ is the proportion of negative examples which varies between 0 and 1.

To illustrate the concept of entropy, let us assume there exists a set S which is a collection of 20 samples including 5 positive and 15 negative samples. The entropy of S relative to the boolean classification is

$$H(S) = \frac{-5}{20} log_2 \frac{5}{20} - \frac{15}{20} log_2 \frac{15}{20} \tag{2}$$

The entropy is 0 if all members of S belong to the same class which is a reflection of complete purity of the set whereas an entropy of 1 indicates the equal of mix of positive and negative samples and is called an impure set.

In the process of classifying data using decision trees, it is important to measure the effectiveness of attributes in performing such a classification. The measure used for this purpose is called *information gain* which calculates the reduction in entropy as a result of partitioning the data on a particular attribute.

Definition 6.3.2. *Information Gain:* Information Gain of an attribute A with respect to a collection of samples S is defined by the reduction in the entropy when splitting the data over a given attribute value. It is given by the following equation

$$IG(S, A) = H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} H(S_v) \tag{3}$$

where

- *H(S):* The entropy of the original set S
- *Values(A):* The set of all possible values for attribute A
- S_v*:* The subset of S for which A has value v

The attribute with the maximum information gain is chosen as the splitting attribute over others and is considered the most informative feature in the feature set. After the first split the next attribute to split the data on is chosen by comparing the information gain of all the remaining attributes and picking the one which would result in maximum reduction of entropy from the previous state. This process is continued till the data can no longer be split any further.

When considering a feature matrix consisting of feature set F_N consisting of N raw diagnosis codes as features, each feature $f_i \in F_N$ has a certain level of entropy associated with it with respect to the target variable *t*. As mentioned before the first attribute to be split on is chosen by the amount of entropy reduced from the previous state which is the target variable itself. Instead of deciding the order of the features for splitting using information gain, we leverage this concept to decide the grouping of the features into a higher level based on a given hierarchy.

Problem Definition: Given a hierarchy of diagnosis codes TD with l levels where l is the lowest level consisting of N raw icd-9 codes constituting a feature set FN for a predictive model, the goal is to find the optimal grouping level for each raw code by maximizing the information gain with respect to the target t to be finally used as a feature in a predictive model without compromising the prediction power of the same.

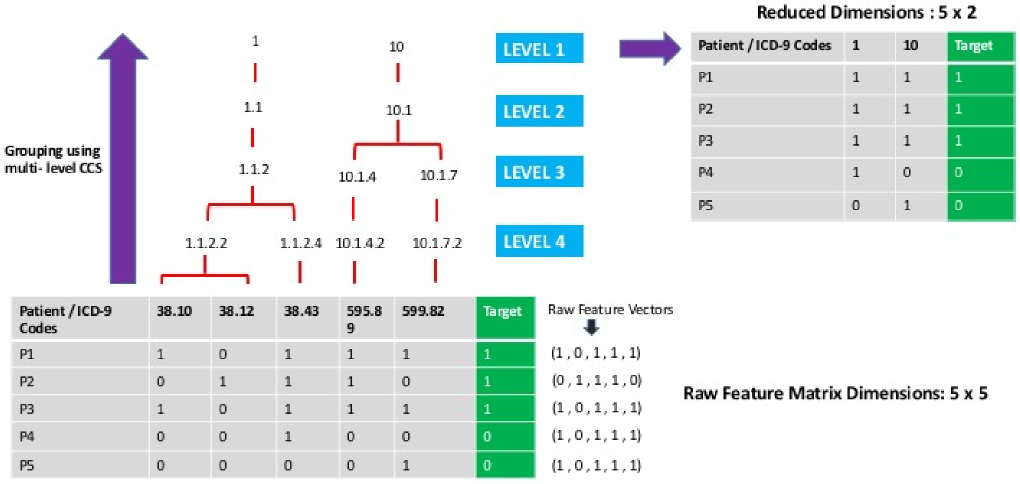
6.3.2 Information Gain in Multi-level CCS Hierarchy

Most of the datasets acquired from various EHRs or claims data from private vendors such as IMS health consists of raw ICD-9 codes which are primarily recorded for billing and insurance purposes. When developing a predictive model, we tend to use these codes in their raw form and project them as binary features in a feature matrix with values ‘1’ representing the presence of a particular ICD-9 code for a particular patient and ‘0’ signifying otherwise. Figure 20 shows an illustration of a feature matrix with 5 patients and 5 diagnosis codes as features along with a target outcome variable. The dimensions of the raw feature matrix is 5 x 5. According to the multi-level CCS classification the level 5 diagnosis codes ‘38.10’ and ‘38.12’ are child nodes of the same parent node ‘1.1.2.2’ which belongs to level 4. Similarly ‘38.43’ is the child node of ‘1.1.2.4’ which is a level 4 node. These two level 4 nodes can in turn be grouped into level 3 nodes and finally into a single level 1 node ‘1’ which is the topmost level in the CCS hierarchy. A similar structure can be observed with level 5 codes ‘595.89’ and ‘599.82’ which can be grouped into a single parent node at level 1 i.e ‘10’. This reduces the feature matrix to 5 x 2 consisting of only 2 level 1 diagnosis codes by grouping the child nodes at each level of the hierarchy and re-populating the matrix values with respect to the reduced set of features using the following equation.

$$B(C_p)_{L_i} = \vee_{k=1}^{n} B(C_k)_{L_{i-1}} \tag{4}$$

where

- *B(x):* Boolean value of x which can be either 0 or 1
- C_p*:* Parent code
- L_i: The i^th^ level in the multi CCS hierarchy
- C_k*:* The k^th^ child node with k ranging from 1 to n



Reduced Dimensions : 5 x 2

Patient / ICD-9 Codes	1	10	Target
P1	1	1	1
P2	1	1	1
P3	1	1	1
P4	1	0	0
P5	0	1	0

Patient / ICD-9 Codes	38.10	38.12	38.43	595.8 9	599.82	Target	Raw Feature Vectors
P1	1	0	1	1	1	1	(1 , 0 , 1 , 1 , 1)
P2	0	1	1	1	0	1	(0 , 1 , 1 , 1 , 0)
P3	1	0	1	1	1	1	(1 , 0 , 1 , 1 , 1)
P4	0	0	1	0	0	0	(1 , 0 , 1 , 1 , 1)
P5	0	0	0	0	1	0	(1 , 0 , 1 , 1 , 1)

Raw Feature Matrix Dimensions: 5 x 5

Figure 20: An illustration of Multi-level CCS classification of level 1 code '7'.

One of the solutions to the problem of reducing feature dimensionality is to use the top level CCS codes as features such as codes '1' and '10' as shown in figure 20. But our goal is 2-forked, i.e we would like to reduce the feature set in such as way such that the features are not too generic i.e they are clinically interpretable as well as preserve the prediction power of the model that they are used for. This led us to using the concept of information gain to measure the entropy reduction achieved with respect to the target variable by grouping codes at every level. If grouping the codes resulted in higher reduction in entropy i.e higher information gain is achieved, then grouping was performed by keeping the parent nodes and removing the child nodes. However if grouping resulted in less information gain than any of the child nodes then the parent node was removed from further consideration and the appropriate child nodes were retained and promoted to the higher level to be considered for grouping at the next higher level. Figure 21 shows an illustration of this process using the matrix shown in figure 20.

The entropy of the target using the matrix from figure 2 is 0.97 using equation 1. The numbers in brackets below every diagnosis code represents the information

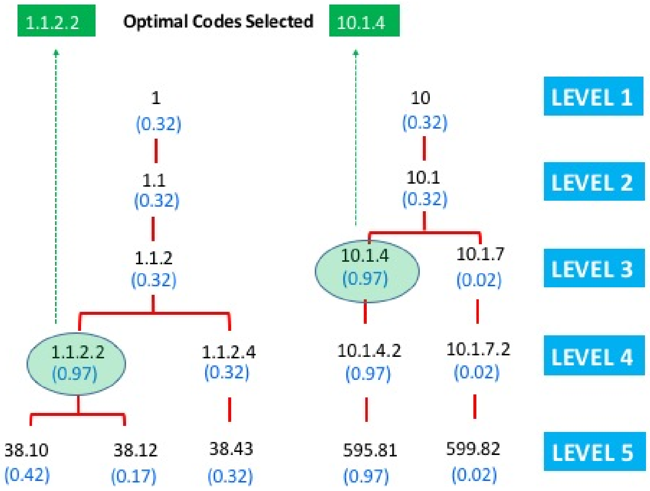


Figure 21: An illustration of optimal code selection process. The information gain is reported in brackets for every diagnosis code

gain achieved by using the corresponding diagnosis code as a splitting attribute in a decision tree. We observe that the diagnosis codes '38.10' and '38.12' have information gain 0.42 and 0.17 respectively and both are child nodes of '1.1.2.2' at level 4 which has an information gain of 0.97. This shows that using explicit codes in a predictive model at level 5 results in higher level of impurity as opposed to using the corresponding parent code at level 4 which results in a pure set if split in a decision tree. The other level 5 code '38.43' does not have any siblings and thus is replaced with its parent '1.1.2.4' at level 4 which has the same information gain as the child. The information gain of the level 4 codes '1.1.2.2' and '1.1.2.4' are compared with their parent code at level 3 and we observe that by grouping the codes at level 3 does not result in higher information gain than the maximum gain achieved by any of its children. Instead, grouping at level 3 reduces the information gain which means we get an impure dataset after grouping. In such scenarios grouping is rejected and all the child nodes with information gain higher than the parent are retained and promoted to level of the parent while the remaining child nodes along with the parent are discarded. In the next iteration the retained child nodes would be compared with the siblings of their parent code and the process continues till level 1. In this illustration we observe that diagnosis code '1.1.2.2' with an information gain of 0.97 is never overridden by a higher level code and thus becomes one of the features in the final feature set. Similarly diagnosis code '10.1.4' having higher information gain than its siblings and parent pushes its way all the way up to level 1 and finds a place in the final feature set. The pseudo code for the algorithm is given below. The input for the algorithm is a feature matrix with dimensions i (patients) x j (raw ICD-9 diagnosis codes), target label for each patient and a multi-level CCS hierarchy. The output is a set of diagnosis codes in which each code may belong to any of the 5 levels of the CCS hierarchy but the size of the set is less than or equal to 'j'.

6.4 Experiment Results

We apply the HCS algorithm in multiple different predictive models which have been trained on data from different sources. In this chapter we present 3 predictive tasks for which the observation period is used to extract the features while the prediction window is used to evaluate the outcome. For each predictive task we build an experimental model which consists of diagnosis code features generated by HCS and a baseline model consisting of raw ICD-9 codes as features. A subset of data is used for training the model using random forest classifier and tuning the model parameters and finally tested on a hold off test set. Following are the descriptions of the predictive tasks :-

1. **Predictive Task-Refractory Epilepsy : Early prediction of refractory epilepsy**

   Claims data for epilepsy patients spanning a period of 10 years is obtained from IMS Health for this task. Most of the epilepsy patients get respite from seizures from the first antiepileptic drug (AED) prescribed to them. But there exists

Algorithm 2 Hierarchical Code Selection (HCS)

```
$H(T) \leftarrow$ Entropy of Target
$L \leftarrow$ Level of multi-level CCS hierarchy
$D_{\mathrm{L}} \leftarrow$ Set of diagnosis codes at level L
$P \leftarrow$ Set of parent codes of child codes at level L
$R_1 \leftarrow$ Set of diagnosis codes retained at level L for a particular parent
$L = 5$
while $L>1$ do

    for all code $\mathrm{d} \in \mathrm{D_L}$ do
        $\mathrm{p_d}$ = $getParent$(d)
        if $\mathrm{p_d} \notin \mathrm{P}$ then
            P.add($\mathrm{p_d}$)

        for all $\mathrm{p_d} \in \mathrm{P}$ do
            $IG(p_{\mathrm{d}}) \leftarrow$ Information gain of parent
            $\mathrm{R_L} = \phi$

            for all child $\mathrm{c_d}$ of $\mathrm{p_d}$ do
                $IG(c_{\mathrm{d}}) \leftarrow$ Information gain of child $\mathrm{c_d}$
                if IG($\mathrm{c_d}$) >IG($\mathrm{p_d}$) then
                    $\mathrm{R_L}$.add($\mathrm{c_d}$)
            if $\mathrm{R_L}$.empty() == true then
                $\mathrm{R_L}$.add($\mathrm{p_d}$)
            else
                $\mathrm{D_{L-1}}$.add($\mathrm{R_L}$)
    L = L -1
```

approximately 25% of the epileptic population which does not achieve seizure freedom in spite of being treated with multiple AEDs and are categorized as refractory [18, 57]. Since refractory patients have to undergo long periods of multiple treatments, they are under a lot of financial and psychological pressure [88]. Clinicians are interested in predicting the patients likely to become refractory at the time when the 1st AED prescription fails i.e it is replaced by another AED and is referred to as the index date. The case or refractory patients are the ones which fail at least 4 or more AEDs and the controls are the ones which fail exactly one AED. The observation window extends from the index date all the way until the first visit of the patient and the prediction window is at least 6 months from the index date.

2. **Predictive Task-Asthma: Prediction of patients likely to get admitted for Asthma**

   We obtain the asthma outpatient and inpatient visit data from Electronic Health Records (EHRs) maintained by the Children's Hospital Of Atlanta (CHOA). Asthma is the most prevalent chronic disease amongst children and is one of the most difficult to diagnose at a very young age [32]. The clinical symptoms presented by children are variable and nonspecific due to coexistence of other wheezing disorders. The goal is to predict based on patient's medical history if a patient would be admitted to the hospital with asthma as the primary diagnosis. The control patients in this task have been chosen to be the ones which do not have an asthma diagnosis in any inpatient visit. The index date for case patients is the date of the asthma related inpatient visit whereas that for control patients is the most recent visit date. The observation period is chosen to be 365 days before the index date while the prediction window extend up to the last visit date after the index date.

Table 15: Data Statistics for a) Task-Refractory Epilepsy b) Task-Asthma and c) Task-Post Discharge Mortality

Prediction Tasks	**Task-Refractory**		**Task-Asthma**		**Task-Mortality**	
Dataset	**IMS Health Claims**		**CHOA EHR**		**MIMIC - III**	
	Case	**Control**	**Case**	**Control**	**Case**	**Control**
Training Set	28,485	81,984	5,816	41,694	500	441
Hold Off Set	5,671	14,670	5,869	41,757	254	379

3. **Predictive Task-Post Discharge Mortality: Prediction of patient mortality within 30 days after discharge from hospital**

   For this task, we use MIMIC-III, which is a critical care public dataset encompassing a diverse and a wide range of ICU patients [79]. Post discharge mortality prediction is an important clinical problem that clinicians are interested in solving [11]. Despite extensive care provided in an ICU which usually results a successful discharge from the hospital, there remains a risk of subsequent deterioration of the patient?s condition and death. Identification of factors which influence such an outcome would help the clinicians in providing better care to patients while they are in the ICU [31]. A predictive model is developed to predict mortality of ICU patients within 30 days of discharge from the hospital. The patients who die within this time period are the cases and the alive patients constitute the control set. The index date in this task is the date of discharge and the observation period extends up to 2000 days prior to the index date. The prediction window is the 30 day period after the index date.

The table 15 shows the statistics of the data used to develop the aforementioned models and the experimental setup is shown in figure 22.

6.5 Results

The features in the predictive models consist of diagnosis codes extracted from the data in the observation period for each data set. We use the HCS algorithm to perform the optimal diagnosis code grouping process on the baseline feature matrix to

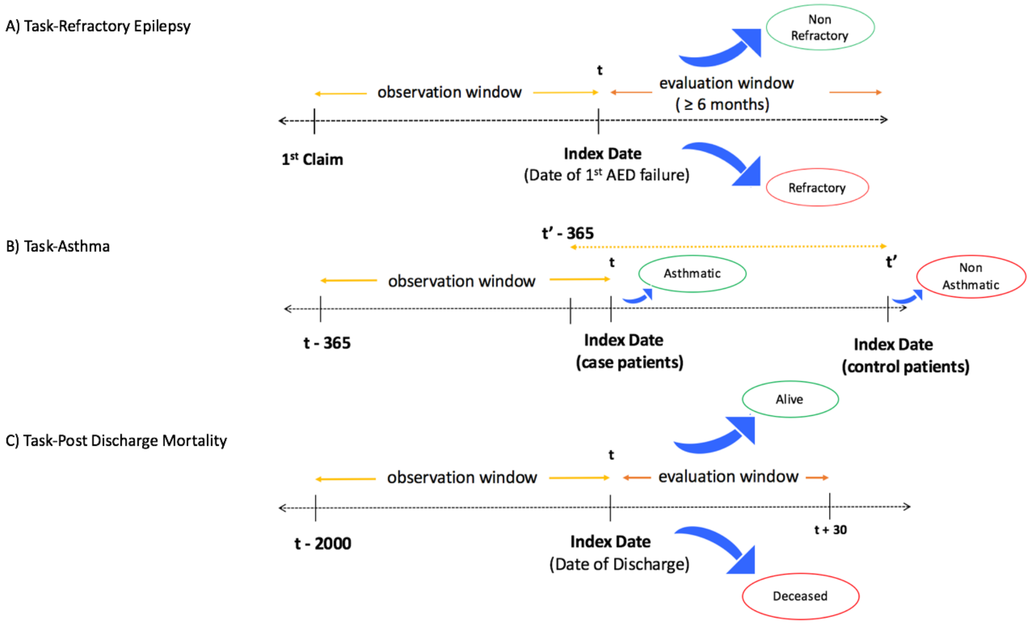


Figure 22: Experimental setup for predictive modeling tasks A) Task-Refractory Epilepsy B) Task-Asthma C) Task-Post Discharge Mortality.

come up with a reduced feature vector to develop the experimental model whereas the baseline model consists of the raw ICD-9 codes as features. Table 16 shows the percentage of codes reduced at each level of the CCS hierarchy for all the 3 datasets used. At level 5 we have all the raw ICD-9 diagnosis codes. At level 4, the HCS algorithm performs grouping of these codes and discards some of the irrelevant codes. The set of codes retained at this level ranges from level 5 to level 4 based on the information gain criteria. This process continues till level 1 and finally the set the codes retained at level 1 ranges from level 5 all the way up till level 1 of the hierarchy. After all the iterations are finished, we are able to discard more than 90% of the diagnosis codes and retain only the relevant ones for predictive modeling.

For each of the aforementioned tasks, we compare the performance of the experimental and the baseline models by varying the number of features used to perform the prediction. Since the number of features in the baseline and the HCS generated feature matrix are markedly different we use equivalent number of features for a fair performance comparison.

Table 16: Percentage of codes reduced at every level of the multi level CCS hierarchy for the 3 datasets

Model	Task-Refractory		Task-Asthma		Task-Mortality	
Dataset	IMS Health Claims		CHOA EHR		MIMIC - III	
Algorithm Consideration Level	No of Codes Retained	Percentage reduction with respect to baseline feature matrix	No of Codes Retained	Percentage reduction with respect to baseline feature matrix	No of Codes Retained	Percentage reduction with respect to baseline feature matrix
5	10,216	0 %	5,524	0 %	1,630	0 %
4	8,558	16.2 %	4,761	13.8 %	1,464	10.1 %
3	3,846	62.3 %	2,512	54.5 %	1,080	33.7 %
2	1,083	89.3 %	1,010	81.7 %	571	64.9 %
1	50	99.5 %	61	98.8 %	139	91.4 %

Figure 23 & 24 shows the performance of the experimental model along with the performance of the baseline model using the popular performance metrics; area under the ROC curve (AUC) and prediction accuracy respectively. The X- axis shows the number of features used by the two models. Since the number of features in the baseline consisting of only level 5 ICD-9 codes are much higher than the ones in the experimental model we vary the number of features in the experimental model from 1 to 100 % and use an equivalent number of features for the baseline model for fair comparison. Figure 5 shows the AUC variation as we increase the number of features in the model. As is evident from the figure, the experimental model with the HCS generated features always performs better than the baseline model. For Task-Refractory Epilepsy the top 3 features of the HCS generated set result in an AUC of 0.67 in the experimental model whereas a comparable AUC is achieved by the baseline using the top 20 raw ICD-9 codes as features. The highest AUC of 0.69 is achieved by the experimental model using the top 20 of the HCS features after which there is no further improvement. With equivalent number of features in the baseline, the AUC achieved is 0.67 which does improve further with inclusion of more features in the model but is not able to achieve the performance reported using HCS features. For Task-Asthma, the top 3 features of the experimental model report an AUC of 0.74 whereas an equivalent number of raw ICD-9 codes result in an AUC of 0.70 and as observed in the Task-Refractory Epilepsy, the highest AUC of 0.78 is achieved with the top 20 HCS generated features whereas it takes 35 raw ICD-9 codes to achieve the same AUC. In the Task-Post Discharge Mortality, the AUC reported by the experimental model is always higher than the one reported by the baseline by approximately 0.02-0.05 with the highest AUC achieved being 0.74. A similar trend is observed in the line graph showing the variation in prediction accuracy. The accuracy reported by the experimental model for Task-Refractory is 0.64 which is achieved using the top 3 HCS features whereas the baseline model reports a comparable accuracy with 50 ICD-9 codes. The accuracy values in the baseline model for Task-Asthma is shown to be high with less than 10 features although it reports the same accuracy as the experimental model using the top 20 features which reported a higher AUC as mentioned before. The accuracy value of the experimental model is shown to be constant at 0.83 whereas that of the baseline drops below 0.83 with more than 30 features. The trend observed in Task-Post Discharge Mortality with respect to accuracy is similar to that observed with respect to the AUC values. The highest accuracy achieved by the experimental model is 0.65 which is 0.04 higher than the one reported by the baseline.

The results show that not just the performance of a predictive model with HCS generated features is higher than that reported by raw ICD-9 codes, a small number of HCS features are clinically strong enough to perform a prediction with performance comparable with a model with much higher number of raw ICD-9 codes as features. This also makes it easier for the clinicians to interpret the factors influencing the prediction since they have to deal with less number of features which could be part of any of the CCS hierarchy levels ranging from level 1 to level 5 and thus may not be too generic or specific.

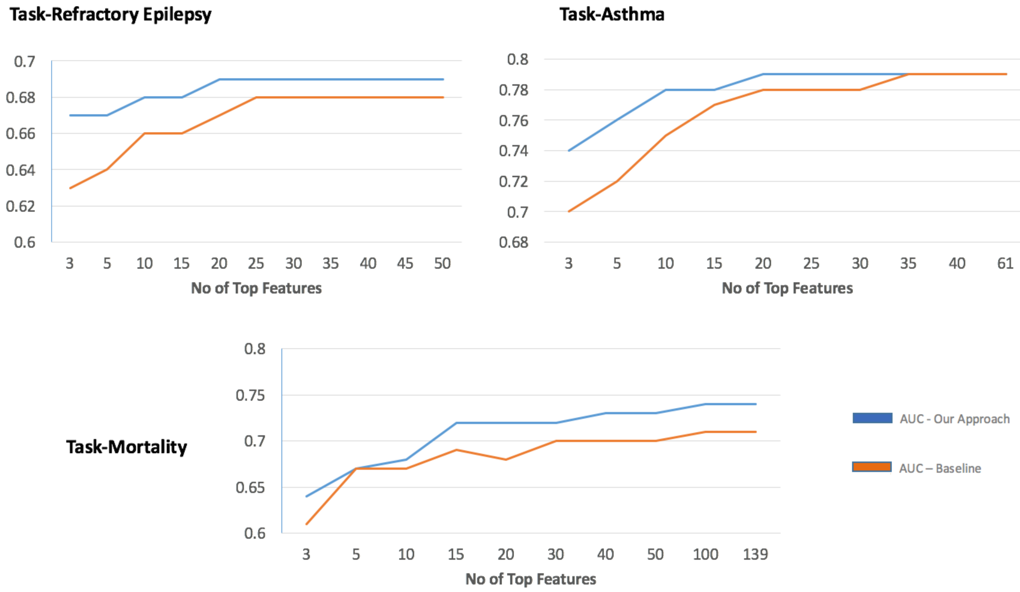


Figure 23: AUC variation of prediction tasks

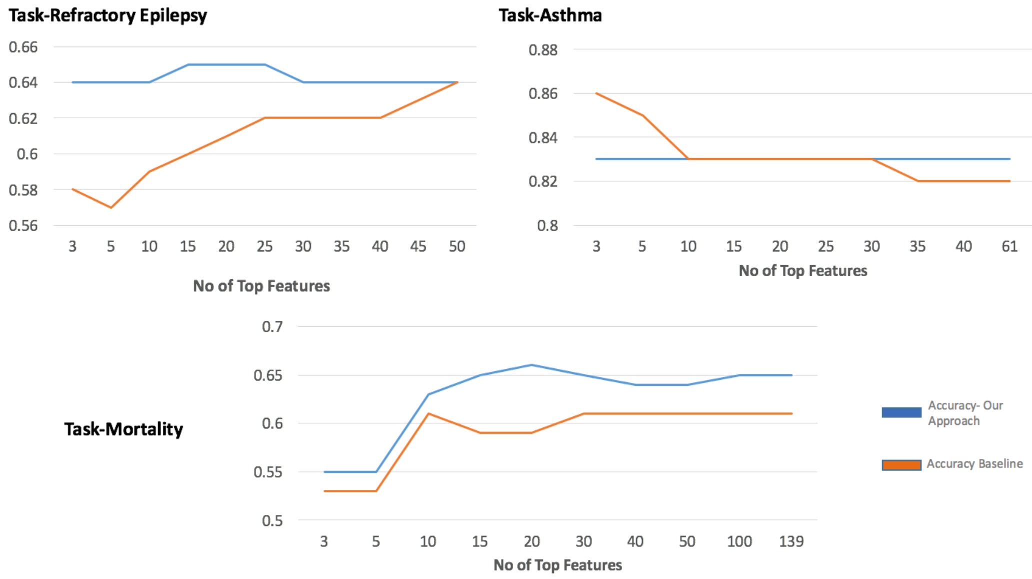


Figure 24: Accuracy variation of prediction tasks

6.6 Conclusion

In this chapter, we propose an information gain based approach to leverage hierarchical nature of medical ontologies for optimal feature engineering in the context of clinical predictive models. We use the CCS ontology which consists of multiple levels of classification of ICD-9 diagnosis codes in a hierarchy in which the topmost levels represent a generic view of the disease condition under consideration. As we go deeper into the hierarchy , the codes are classified into more specific representations with the last level being the raw ICD-9 codes themselves. The general approach in developing a predictive model is to use the the raw ICD-9 codes as binary features in a predictive model and perform some basic aggregations on these features to predict an outcome of interest. Since ICD-9 codes were originally developed with the goal of standardizing the process of recording the disease conditions of patients for billing and reporting purposes, they have been defined to be extremely granular. The drawback of such raw codes in a predictive model is that a large number of features need to be considered to achieve the highest performance metric. Using the information gain based approach we have successfully reduced the raw features by grouping them optimally into clinically relevant broader categories which are manageable in size and clinically interpretable by the domain experts. We have also shown that predictive performance can be improved over the baseline model using the optimally selected set of features in the context of 3 prediction tasks developed using EHR, claims and ICU data respectively.

Chapter VII

AUTOMATED SCHEMA EVOLUTION APPROACH VIA META-DATABASE MANAGEMENT

In data analytics, one of the foremost step deals with defining a cohort of patients being studied for a particular disease. Cohort construction helps in developing models which are focussed on a homogenous set of patients as opposed to a very broad sample of patients which are difficult to analyze. Even though some diseases have set of patient management guidelines and common data elements which have been extensively laid down by clinical experts to analyze, there are some diseases for which clinicians may want to differ from the traditional methods of treatment and record new data elements for certain set of patients [36].

Most small outfits which have less number of physicians have to incur a lot of maintenance cost in adopting expensive EMR systems which could be customized by the vendor based on needs of the user. To cater to such small scale organizations, we propose a user interface based tool to transfer the flexibility of automating the process of customizing back-end databases to the clinical researchers. The tool is based on a dialog based approach to make appropriate schema changes based on user's requirements via the front-end interface. The interface provides the user the capability to modify existing forms consisting of disease specific data elements generated by other users by adding or dropping certain data elements based on their requirements. The user can enter data with the help of these forms which gets recorded in an underlying database. The primary motivation for this work is to avoid the intervention of database system administrators to make periodic changes at the back-end as a result of evolving data. Since most medical facilities have a relational database to record patient data, we base our approach on a relational data model with a long term goal of supporting analytics.

The target audience for the underlying approach is clinical researchers with very basic knowledge of data modeling who are interested in generating cohorts of patients for analysis and developing analytical models around our tool. We have tested our implementation with a small clinical practice which deals primarily with ALIF (Anterior Lumbar Inter-body fusion) patients [130]. Clinicians at this practice are focussed on dealing with surgical procedures related to the spine and are in need for a system which can provide them the capability to record new data in a traditional relational database as and when required. Our contributions towards this work are the following:

1. We develop a meta-database called *Mbase* which has been carefully modeled to handle the dynamic nature of the front-end involving customization of template forms.

2. An approach is developed to gather the requirements of the front-end user to customize the template forms and translate the same into database modeling operations keeping the consistency of the database intact.

3. We also develop a data retrieval utility which is supported by an interactive user interface providing the capability to group underlying data elements in a format required by the user to generate patient cohorts.

Although the traditional approach of developing a relational database application begins with conceptually constructing a data model using the entity-relationship model followed by developing a relational schema[49], most of the applications in recent times undergo extensive changes during the development process. In this chapter, we present an approach which caters to data evolution in real time keeping relational constraints of the underlying database intact.

7.1 Related Work

Schema evolution has been studied extensively in the past. One such study involves an application which is focussed in the clinical dental domain and is adept at handling specific types of relationships between relational database tables [146]. They require the user to be familiar with the back-end database and provide mapping between the front-end and the back-end data elements which entails the user to revise the mapping with any change is made at the front end. They do not have the capability of encompassing all the potential changes which can be triggered in a relational model such as creation of new tables as a result of new requirements, dealing with multi-valued attributes and relationship specific attributes. Systems like Encore [136] and Orion [83] also deal with handling database evolution issues. According to Palisser et al. [9], Orion generates a version of the state of the database with every small transformation in the schema which could potentially lead to exponential number of versions. Encore on the other hand is an object oriented database system and generates versions of object types with changes in the design environment. Another system called the O2 automatically generates consistent databases in spite of heavy changes triggered by update operations on the structure of the back-end schema [50]. The drawback of such an approach is that the transformations at the schema level could be reflected immediately or deferred to a later time period which could be problematic and affect data availability. We carefully analyze the aforementioned approaches and systems and design our approach to close the existing gaps in real time schema evolution without compromising on relational consistency and availability of data.

7.2 System Architechture

Figure 25 shows the architecture of the system. The front-end consists of template forms used to store data in an underlying database (Sbase) which consists of data elements appropriately stored in relational tables. The user has the option of either using the existing forms if they satisfy the requirements or modify them appropriately. The user also has the flexibility to create new forms which can be linked to appropriate existing forms. For example, a form named ‘Procedures’ exists and is used to record different types of surgical procedures that can potentially be performed on an ALIF patient, and if a new form ‘Patient Medical Record’ needs to be generated to store details about patients, the latter could be linked to the ‘Procedures’ form indicating that patients may undergo surgical procedures and the ‘Patient Medical Record’ form may contain some data elements which are specific to the procedure that the patient would undergo, e.g. number of hours in surgery. Such use cases involve extensive changes at the front end, since multiple forms could be modified and created, along with appropriate schema level changes in Sbase to incorporate the changes made in the forms. To manage such heavy structural changes in the forms, we use a meta-database called the Mbase to track such changes. The Mbase is a collection of relational tables, which store information about the existing form definitions such as name of the form, the labels of the each field it contains, the type of fields etc. The changes made in the Sbase schema would be handled by data modeling operations based on information provided by the user about the changes made in the forms.

7.2.1 Form Structure Management using Mbase

Figure 26 shows the schema of Mbase. The ‘Form’ table stores basic annotations of a form such as the name of the form, the date it was created on, the user who created the form, etc. All the existing forms are stored in this table along with the new forms generated by the user. Any modification to an existing form results in

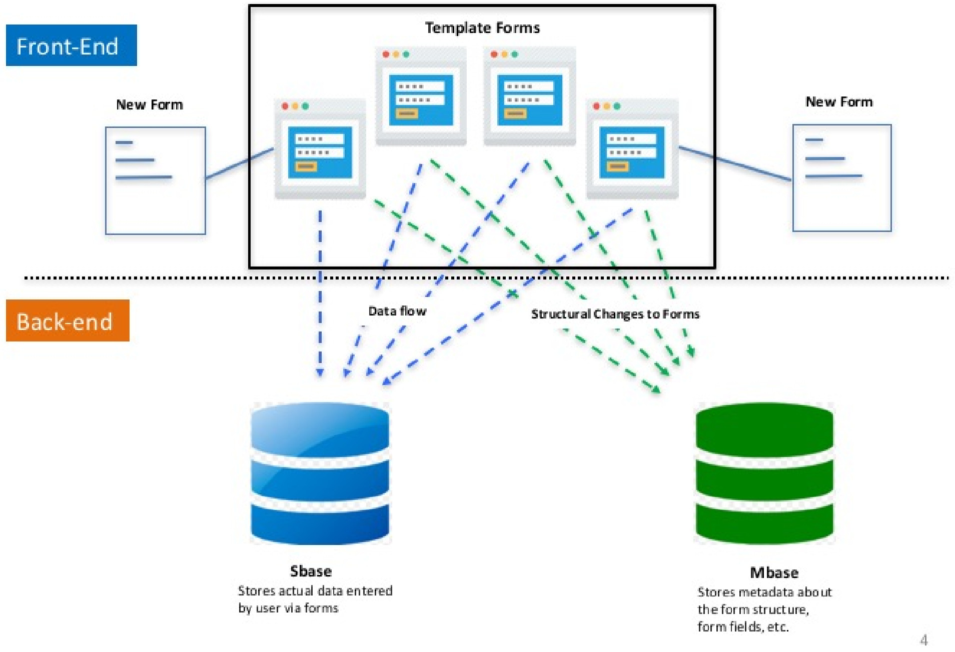

Front-End
Template Forms
New Form
New Form
Back-end
Data flow
Structural Changes to Forms
Sbase
Stores actual data entered by user via forms
Mbase
Stores metadata about the form structure, form fields, etc.
4

Figure 25: System Architecture

a new form and is assigned a different identifier which is the primary key of the 'Form' table. We also record the information about the original form which has been modified to come up with the new form and link it back to its own form identifier in this table. Since a form can consist of data fields, we have a table 'Data Field' which keeps a record of all the data fields, along with the labels that are displayed on the different forms. Since the user has the flexibility to add or delete fields from an existing form, we have to keep a record as to which fields need to be displayed for a particular form. A many to many relationship is defined between the 'Form' table and 'Data Field' table which results in another table 'Form Has Field' which records the form identifier and the field identifiers. This table can be queried to manage the content of each form. A data field can be a usual text box, a drop-down menu, checkbox, radio button, etc. This information about the type of the data field is recorded in the 'Data Field Category' table in which the field categories can be

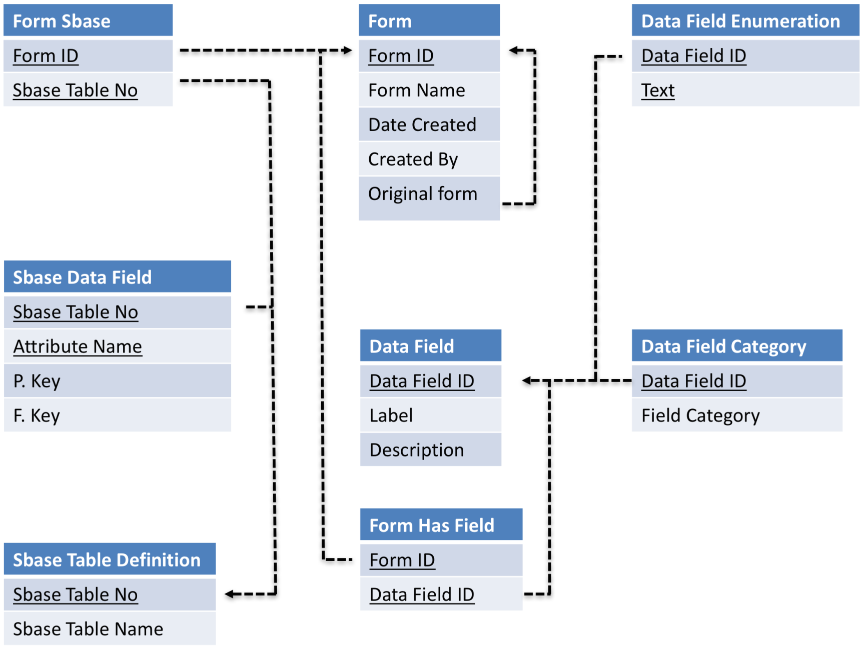


Figure 26: Schema of Mbase database. The arrows show referential integrity constraints

any of the aforementioned options. Our current implementation allows the option of having data fields which have radio buttons, checkboxes or dropdown menus and this information is recorded in another table called the 'Data Field Enumeration' table which records the field identifier along with the various text options which need to be displayed. In the current implementation, creation of a new form from scratch results in a single table or multiple ones depending on the relationships it shares with the existing forms, whereas addition or deletion of fields from an existing form results in a new version of the form, which results in appropriate changes to the corresponding table structure or new tables as well. The association between the forms and the corresponding tables in Sbase is captured in the 'Form Sbase' table. We also store the structural information of the Sbase tables along with information about the primary and foreign key attributes in the 'Sbase Data Field' and 'Sbase Table Definition' table. The arrows in the figure represent the referential integrity constraints amongst the tables in Mbase.

7.2.2 Front-end functionality

The front-end of our tool supports two main functionalities which deal with recording incoming data and querying existing data to generate data extracts. While recording data, the user can either choose to use the existing templates and modify them based on business requirements or generate a new form. Figure 27 shows the screen which displays the the list of available template forms along with the option of generating a new form from scratch. We also provide a drop down list of patients in the system for which existing data can be queried for editing or adding new information about an existing patient. On choosing the option to create a new form, the user is presented with a blank template on which he can drag appropriate form components from the 'Form Component' panel. When adding a data field the user is presented with another set of questions to gather information about the same. This information is vital to generate the correct data type of the field and help the system to make a decision as to whether this new field would result in an additional attribute in an existing table or a new table needs to be created based on the guidelines of developing relational models. Figure 28 shows the data field screen at the bottom which came up as a result of adding 'Gender' to the new form template on top. Once the information asked has been submitted, based on the type of the data field, additional questions may be asked. For example, in the case of 'Gender', the user would have to provide additional information about the options which need to be displayed since it is a drop-down field. After all the fields have been added to the form, the user would be required to enter some information about the new form and its relationship with existing forms if any. This helps the system to develop the back-end data model as appropriately as possible. Figure29 shows the form generator screen with the appropriate questions. The '?' next to each field provides the user with examples to get clarity on what is being asked. For example, the user is asked to pick a data field which can uniquely identify an instance of the data entered using the form. This would help the system

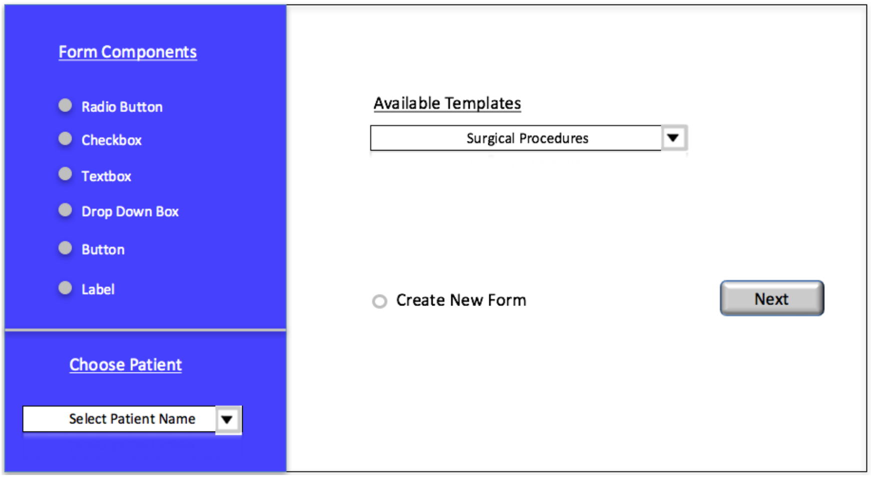
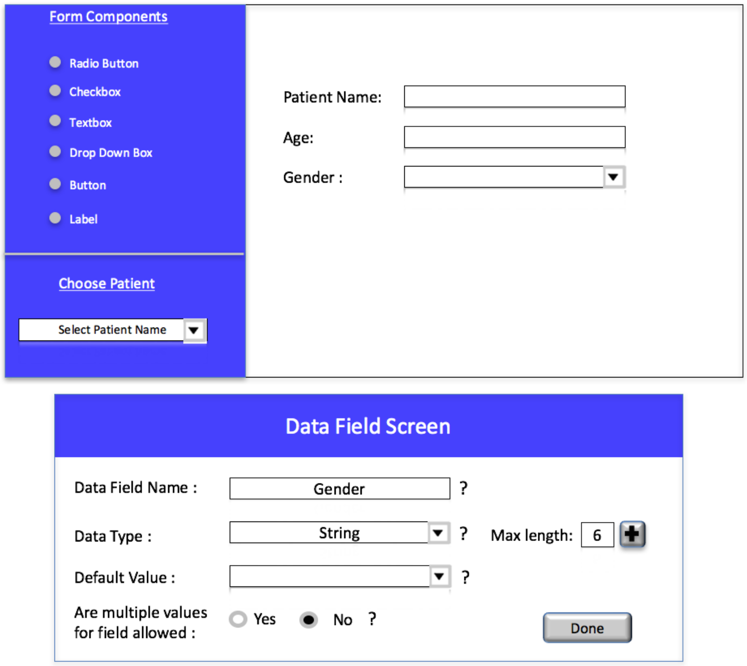


Figure 27: Screen showing list of available forms.

assign a primary key to the appropriate table. Additional questions pertain to the relationship with the existing forms to make appropriate decisions about the schema changes that would be required in Sbase.

7.2.3 Data consistency Versus User flexibility

The current approach gives the user a lot of flexibility to make changes at the front-end since he has the freedom to generate new forms as and when required. This could potentially lead to a lot of problems with data consistency and redundancy in cases when users choose to build new forms. There could be cases when the data elements required by the user differs very minimally from the ones present in the template forms in which case, the user is expected to modify the appropriate existing form but instead chooses to create a new one. This could result in multiple tables having redundant data elements which is a drawback of the current implementation of the system. For example, there could be two 'Patient Medical Record' tables at the back-end created by two different users which have majority of the attributes in common but not all. This could result in data about the same patient being entered in both tables by the two users at different times. Also, some patients may appear in one table while others appear in the other one. To merge the data into a single



Patient Name:

Age:

Gender :

Data Field Screen

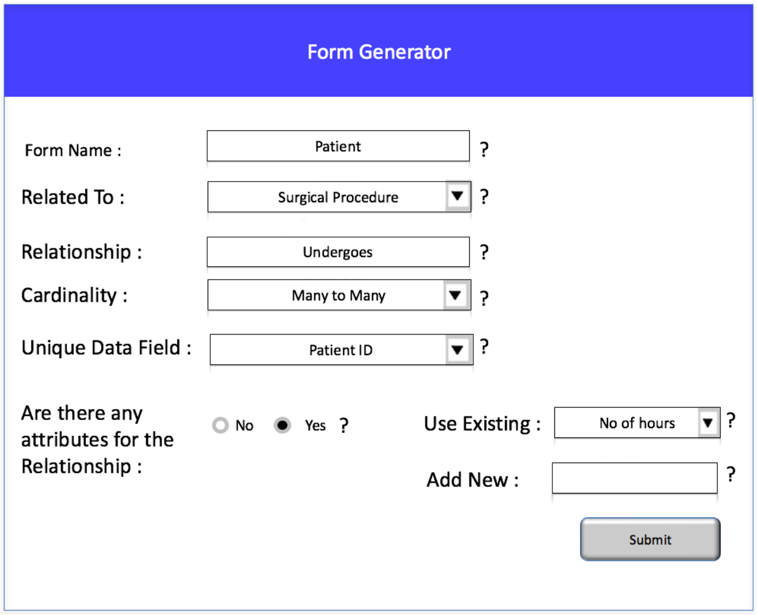
Data Field Name : Gender ?

Data Type : String ? Max length: 6

Default Value : ?

Are multiple values for field allowed : ○ Yes ● No ?

Figure 28: Data Field Screen



Form Generator

Form Name : Patient ?

Related To : Surgical Procedure ?

Relationship : Undergoes ?

Cardinality : Many to Many ?

Unique Data Field : Patient ID ?

Are there any attributes for the Relationship : ○ No ● Yes ?

Use Existing : No of hours ?

Add New : ?

Figure 29: Form Generator

table at the back-end, we need to monitor the back-end database periodically and perform the ‘full outer join’ operation followed by a ‘union’ on the two tables to achieve consistency.

7.3 Data Extract Generation

The other significant aspect of the system is to assist clinical researchers in generating extracts of data being stored in Sbase for analytical purposes. We provide a UI based data extraction feature which can be easily used to integrate data from multiple tables with the help of appropriate join operations. Only proper joins based on primary key - foreign key relationships between Sbase tables are allowed in the current implementation. The users are also provided the capability to perform basic aggregation operations supported by SQL such as counting number of data points, calculating averages, etc. In addition to aggregation, the users can group the aggregation results per patient to generate patient level summary reports. Figure 30 shows the data extract generator window. The user can choose to view multiple data fields which may belong to a single form or different forms. Once the appropriate fields have been selected, the user is given a choice to perform basic SQL supported aggregation operations on any of the selected fields or another data field of choice. This utility also supports the GROUP BY operation to aggregate data for a certain field, e.g. count of the number of surgical procedures performed per patient could be generated by using the ‘COUNT’ operator on the the ‘Procedure ID’ field and aggregating the data by ‘Patient ID’. The final aggregation results can be filtered based on a threshold using the ‘Filter by’ option which corresponds to the ‘HAVING’ operator in SQL.

Performing data extraction does not require the user to have knowledge of SQL which is a query language used to query data from relational databases. It is therefore possible to have inappropriate user requests to view multiple data fields together. On selecting the required data fields, the system confirms with Mbase if the tables from which the data fields need to be extracted from have the potential to be joined based on the referential integrity information stored in the 'Sbase Data Field' table in Mbase. If the tables cannot be joined the user is prompted to revise the selection. Aggregation operators can be chosen to perform operations on data fields which in turn can be selected by the user. In addition to the aggregators, the system allows the user to group the aggregations based on a selected field to generate instance level data e.g. counting the number of surgical procedures performed per patient. Based on the selections made by the user, the system automatically generates a sql query after performing the following consistency checks:

1. Aggregator operations such as MIN, MAX, AVERAGE are not allowed to operate on categorical data fields.

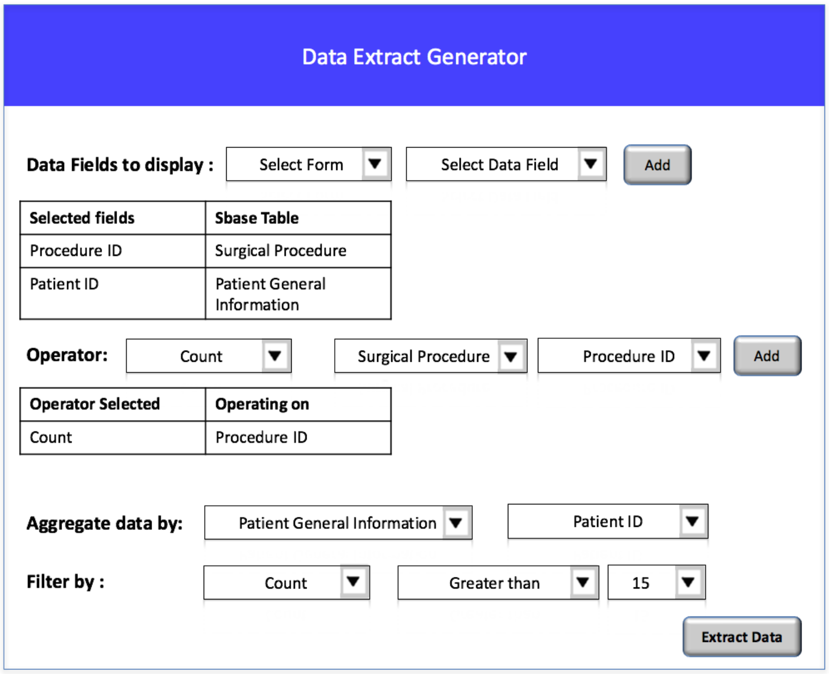
2. Referential integrity is checked when performing joins amongst tables corresponding the selected fields.

3. The data fields chosen by the user for viewing are not allowed to be operated upon by the aggregators.

The data generated based on the query is displayed to the user which can be saved as a view in Sbase with appropriate descriptive annotations for other users to use in the future.

7.4 Discussion

Traditionally, both large and small scale organizations store terabytes of data in relational databases, but with the recent advancements in database research there has been a paradigm shift from relational to NoSQL and graph databases due to their ability to deal seamlessly with unstructured and continuously evolving data[89]. Such



Data Fields to display : Select Form ▼ Select Data Field ▼

Selected fields	Sbase Table
Procedure ID	Surgical Procedure
Patient ID	Patient General Information

Operator: Count ▼ Surgical Procedure ▼ Procedure ID ▼

Operator Selected	Operating on
Count	Procedure ID

Aggregate data by: Patient General Information ▼ Patient ID ▼

Filter by : Count ▼ Greater than ▼ 15 ▼

Figure 30: Screen to generate data extracts

data models do not have a static schema and thus do not require any schema modifications with changing requirements. The primary reason for choosing relational model in our work is that our approach is focussed on schema modifications of already existing databases which are currently in use. Most small scale medical facilities have relational back-end databases, it is more feasible to develop a schema modification approach rather than transitioning to a NoSQL database. In addition to this, the CAP theorem suggests that it is possible to maintain only two of the three properties namely consistency, availability and partition tolerance [30] in a database. NoSQL approaches maintain partition tolerance but would fail at providing consistency and availability both at once. On the other hand, relational databases guarantee consistency and availability but may fail at dealing with partitions. Since the system is primarily designed for storing health related data, it becomes absolutely essential to maintain a relationally consistent system with data available at all times.

7.5 Conclusion

In this chapter, we present a carefully designed approach focussed on dealing with evolving data and frequently changing requirements which is a result of difference of opinion amongst clinicians about what data elements to record for disease specific patients. The system supports a traditional database *Sbase*, the schema of which is specific to the data being stored. There also exists a meta-database namely *Mbase* which stores the metadata about the front-end form structure used to record data, data fields within the forms, linkages between forms and information about the connections between the forms and the corresponding table in *Sbase*. The user has the flexibility to reuse the existing template forms to record additional data and modify them appropriately based on their requirements or generate new customized forms with different data fields. The aforementioned changes in the form structures and creation of new forms result in appropriate changes in *Sbase* such as addition of new attributes in tables, creation of new tables with additional linkages with existing ones to maintain referential integrity. The importance of this system lies in minimizing database administrator's role in maintaining the schema of back-end database although with the current implementation, periodic rearrangement is inevitable to avoid redundancy. In addition to this, the system also consists of a data extract generation feature which can be used to create extracts of data for a set of patients or patient level extracts for analytical purposes. The interface provides various operations which the user can choose to perform on the existing data fields in the database without any knowledge of query languages. SQL queries are automatically generated based on the user selections after performing various consistency and validation tests. This work has been published in [96].

Chapter VIII

CONCLUSION AND FUTURE DIRECTIONS

8.1 Conclusion

The healthcare system in the United States is plagued with increasing costs and increasing chronic illnesses. In spite of having fewer physician visits and lesser hospital admissions than some of the other high income countries, the cost incurred to patients is much higher which is attributed to the greater use of expensive medical equipment and technologies and overuse of laboratory tests. The use of tobacco, obesity, hypertension, risks related to heart diseases and diabetes etc. have created havoc in the lives of an average American and it has become mandatory that they be controlled. The United States has failed miserably in achieving favorable health outcomes such as reducing mortality as a result of ischemic heart disease, amputations due to diabetes, etc. and is desperately in need of healthcare analytical applications which can discover factors leading to such adverse events and improve quality of care. The Health Information Technology for Economic and Clinical Health (HITECH) Act provides incentives to healthcare providers to adopt EHR systems to store data. This has led to clinicians and data scientists to leverage the electronic data being stored to solve important clinical problems and pave the way for personalized medicine. Personalized medicine involves a deep understanding of the clinical and genomic characteristics of individual or a cohort of similar patients with the goal of recommending customized treatment plans for them. A wealth of information is being collected in EHRs and analyzed with the help of state of the art analytical algorithms and tools.

In this dissertation, we presented a non-conventional graph based approach to extract sequential patterns from data. This method is very useful in the context of analyzing treatment data by converting it into a graph consisting of nodes and edges, where nodes represent the patients and medical events. The edges represent the occurrence of a medical event with respect to a patient and also inform us about the sequence in which the medical events occur. We introduced two medically relevant constraints in this algorithm namely the 'Consecutive Occurrence' and the 'Temporal Event Overlap' constraint. The former deals with generating sequential patterns with medical events occurring immediately after one another without allowing any intermediate events to occur whereas the latter deals with overlapping treatments such as multiple drug prescriptions which may be given to a patient in combination, and incorporates them when mining patterns which eventually may consist of such combination drugs.

We presented two case studies which use this sequential mining approach in the context of survival prediction of Glioblastoma patients and comparative analysis of nation wide treatment practices for the three most popular diseases. The Glioblastoma patient survival prediction is an important clinical problem involving discovering factors which help in prolonging survival periods of patients. In addition to clinical and genomic factors, we used sequential pattern mining to mine significant treatment patterns from patient data and use them as features in a predictive model which is developed to predict if a patient would survive for longer than a year which in the context of Glioblastoma is categorized as a long term surviving patient. The other case study involves leveraging claims data for three popular disease conditions namely Autism , Heart Disease and Breast Cancer to compare how care is delivered in different parts of the country. Sequential pattern mining was used to extract patterns of medical procedures for such a comparative analysis.

We also performed an extensive analysis of the epilepsy disease which has engulfed approximately 65 million people world wide out which 2 million are in the United States. Although observational studies exist which study epilepsy patients for more than 10 years, there does not exist an extensive understanding of how the anti epileptic drugs are prescribed in practice. One of the goals of this piece of work was to understand how these drugs are prescribed to patients and what factors related to the first line of treatment or patient specific attributes can influence the second line of treatment. In addition to this analysis, we were also interested in identifying the drug resistant epileptic population called refractory patients at an early stage in the course of the treatment so that they can be prescribed the newly formulated drugs which have shown promising affects in clinical trials. Identification of such patients is extremely important since they do not respond to any of the existing anti epileptic drugs in the market and as a result undergo a financial and a psychological burden throughout the treatment process. To accomplish this, we narrowed down our focus on only the adult epileptic population since pediatric epilepsy has a number of additional subtypes of epilepsy which exclusively occur only in childhood [139]. The point of second anti-epileptic drug prescription is chosen to be the point at which the prediction should be made about a patient likely to be refractory in future or not since patients are prescribed the first anti-epileptic drug for a considerable amount of time before the clinicians switch them to a different drug. Leveraging the patient data up to the point of the first AED failure has turned out to be very informative. A predictive model is designed for this purpose using advanced engineered features from the medical history of patients including specific comorbidities, healthcare ecosystem, insurance payer information, hospital encounters, etc.

The next piece of work focusses on the feature construction aspect of predictive modeling in healthcare. An important chunk of features engineered for use in healthcare predictive models consists of patient diagnoses which are standardized and identified by codes maintained by a popular and universally accepted classification system called the ICD-9 code classification. The codes range from the most high level disease conditions to minor nuances and symptoms presented by patients and were primarily designed for billing and reporting purposes. There also exist knowledge bases such as the Clinical Classification Sotware (CCS) which organizes these codes in a hierarchical format with the goal of making them clinically interpretable. We proposed an approach based on the popular concept of entropy reduction and information gain, to optimally leverage the hierarchical information in a medical ontology such as CCS to construct clinically interpretable features which are not too generic neither too specific for the purpose of predictive modeling. We tested the approach using three different data sets used to develop different predictive models based on features selected by our approach. On comparison with the baseline model consisting of features based on raw ICD-9 codes, we found that we were able to achieve a better predictive performance using our approach than the baseline and yet present a set of clinically interpretable and manageable number of features to the clinicians to interpret and analyze.

The final segment of the thesis consists of work done with regards to data modeling and incorporating the changing requirements in the healthcare industry. We proposed an automated schema evolution system which manages the modifications required at the schema level in a database as a result of changes made in the front end applications. The clinicians usually have a difference of opinion with respect to data elements needed to characterize patients and diseases in spite of having a set of common data elements decided by the appropriate medical community. The system translates the changes made at the front end such as addition and deletion of fields, generation of new forms, etc. into set of operations performed at the back-end without compromising the relational consistency of the database.

8.2 Future Directions

This dissertation has opened up directions in which researchers can build upon the existing work. The non-conventional graph approach for sequential mining can be extended to add more medically relevant constraints prior to extracting treatment patterns. Gap constraint is one such example which restricts the temporal gap between drug prescriptions and can be used to limit the inclusion of those drugs in a pattern which are separated by extremely long gaps. Currently the overlap constraint does not require the drugs or other medical events to overlap for a minimum number of days which could be important from a clinician's perspective and can be incorporated in the future.

The epilepsy patient analysis work currently focusses on analyzing the current practice of anti epileptic drug prescriptions and early prediction of drug resistant epilepsy. This work could be extended in the direction of treatment recommendation and clinical decision support where predictive models could be developed to generate a ranked list of treatment options for non refractory epilepsy patients. For refractory patients, such a decision support system may not be of use in the near future since none of the current drugs in the market are suitable for them and the wait is on for the clinical trials for new drugs to finish. However, the determination of such refractory patients itself, may be valuable to look for alternative treatment options for such patients.

The ontology guided feature selection work also has a lot of potential. The current work compares the HCS algorithm based predictive models with baseline models consisting of raw ICD-9 diagnosis codes. Further experimentation could be done by using multiple baseline models consisting of medical codes from all the levels of the ontology together in a feature matrix. Codes belonging to each level separately could also be considered for developing a baseline model which would result in one model each for every level in the hierarchy. The current algorithm focusses on the parent-child relationships existing in ontologies such as the CCS classification system. The algorithm could be extended to leverage the other clinically significant relationships between diagnoses present in the SNOMED ontology. These relationships are semantically rich and could be used for dimensionality reduction by calculating similarity between different diagnosis codes and grouping them into a broader category if semantically similar.

The dynamic database modification system also has a lot of scope of improvement as mentioned before. The current implementation of the system is based on the assumption that users have very basic knowledge of database modeling which could be revised to make it suitable for completely novice users with no technical knowledge. The query functionality of the system also has the potential for improvement by incorporating the capability to handle complex and sophisticated queries.

Overall, we have laid a groundwork of customizing sequential mining for medical applications along with customizable database schema models and applying it to study and to solve many important clinical problems to personalize treatments and improve the quality of care.

Appendix A

SUPPLEMENTARY MATERIAL

A.1 Supplementary Material for Chapter 3

This section presents the supplementary material for 3.

A.1.1 Raw data elements used for the study

We number of features reported in 3 is after processing the raw data elements and converting most of them into binary features.

1. **Clinical Features**

   Age at initial pathologic diagnosis

   Date of death

   Last follow up date

   Ethnicity

   Gender

   Histological type

   History of neoadjuvant treatment

   Method of initial pathologic diagnosis

   Karnofsky performance score

   Performance status scale timing

   Person neoplasm cancer status

   Prior glioma

   Race

2. **Genomic Features**

   (a) **Copy number variation data of the following genes**

   CDKN2A

   EGFR

   GABRA1

   NEFL

   NF1

   NKX2-2

   OLIG2

   PDGFRA

   RELB

   SLC12A5

   SYT1

   TNFRSF1A

   TRADD

   (b) **mRNA expression levels of the following genes (z-scores)**

   CDKN2A

   EGFR

   GABRA1

   IDH1

   NEFL

   NF1

   NKX2-2

   OLIG2

   PDGFRA

   PTEN

RELB

SLC12A5

SYT1

TNFRSF1A

TP53

TRADD

MGMT

IDH1

PTEN

TP53

(c) **Methylation status of the following genes**

MGMT

3. **Treatment Patterns (used as features)**

Note: The number in the bracket denotes the time of prescription of the drug in the sequence and arrow (→) denotes the direction of sequence.

I (Temodar + Radiation) {1} →Temodar {2} →Avastin {3} →Treatment Termination

II (Temodar + Radiation) {1} →Temodar {2} →Avastin {3} →Treatment Termination

III Radiation {1} →Temodar {2} →Treatment Termination

IV BCNU {3} →Treatment Termination

V Temodar →Treatment Termination

VI Gliadel Wafer {1} →Radiation {2}

VII Radiation {1} →CCNU {2}

VIII (Temodar + Radiation) {1} →Temodar {2} →CCNU {3}

IX Dexamethasone →Treatment Termination

X Temodar {1} →(Temodar + Radiation) {2}

XI Dexamethasone {1} →(Dexamethasone + Radiation) {2} →Dexamethasone {3}

XII (Temodar + Radiation) {2} →Temodar {3} →Treatment Termination

XIII Dexamethasone {1} →(Dexamethasone + Radiation) {2}

XIV (murine81c6 + monoclonal antibody I-131) {1} →Radiation {2}

XV Temodar {3} →Treatment Termination

XVI Dexamethasone {3} →(Dexamethasone + Dexamethasone) {4}

XVII (CCNU + Vincristin + Procarbazine) {2} →Treatment Termination

XVIII Radiation {2} →Temodar {3}

XIX (Temodar + Radiation) {1} →Temodar {2}

XX (Temodar + Radiation) {1} →CCNU {2}

XXI Temodar {2} →Avastin {3}

XXII CCNU {3} →Temodar {4}

XXIII Temodar {4} →Treatment Termination

XXIV (Dexamethasone + Radiation) {2} →Dexamethasone {3}

XXV (Temodar + Radiation) {1} →Temodar {2} →Treatment Termination

XXVI (Temodar + Radiation) {1} →Radiation {2}

XXVII Avastin {3} →Treatment Termination

XXVIII (Dexamethasone + Gliadel Wafer) →Treatment Termination

XXIX (Dexamethasone + Radiation) {2} →Dexamethasone {3}

XXX (Temodar + Radiation) {2} →Temodar {3}

XXXI Radiation {1} →Temodar {2}

XXXII Radiation →Treatment Termination

XXXIII Gliadel Wafer {1} →Temodar {2}

XXXIV Temodar {2} →Treatment Termination

XXXV Radiation {1} →(CCNU + Vincristin + Procarbazine) {2} →Treatment Termination

XXXVI Radiation {1} →(CCNU + Vincristin + Procarbazine) {2}

XXXVII (Temodar + Radiation) →Treatment Termination

XXXVIII Avastin {4} →Treatment Termination

XXXIX Radiation {2} →Treatment Termination

XL (Temodar + Radiation) {1} →Radiation {2} →Treatment Termination

XLI Temodar {5} →Treatment Termination

XLII Radiation {3} →Treatment Termination

XLIII (Dexamethasone + Temodar + Radiation) {3} →(Dexamethasone + Temodar) {4}

XLIV Radiation {2} →BCNU {3}

XLV Radiation {1} →(Temodar + Radiation) {2}

XLVI (Temodar + Radiation) {1} →Temodar {2} →Avastin {3}

XLVII CCNU {2} →Treatment Termination

XLVIII Dexamethasone {1} →(Dexamethasone + Radiation) {2} →Dexamethasone {3}

XLIX Temodar {2} →CCNU {3}

L Temodar {2} →Avastin {3} →Treatment Termination

A.1.2 Karnofsky Performance Score

KPS scores ranges from 10 - 100. Table 17 shows the description of each of the scores.

Table 17: Karnofsky Performance Score Descriptions

Score	**Description**
100	Normal no,complaints; no evidence of disease
90	Able to carry,on normal activity; minor signs or symptoms of disease
80	Normal,activity with effort ; some signs or symptoms of disease
70	Cares for,self; unable to carry on normal activity
60	Requires occasional assistance, but is able to care for most of his personal needs.
50	Requires considerable assistance and frequent medical care.
40	Disabled; requires special care and assistance.
30	Severely disabled; hospital admission is indicated although death not imminent.
20	Very sick; hospital admission necessary; active supportive treatment necessary.
10	Moribund; fatal process,progressing rapidly.

REFERENCES

[1] "National hospital discharge survey." http://www.cdc.gov/nchs/nhds. [Online; accessed 1-October-2016].

[2] "Personalized medicine." http://health.usnews.com/health-conditions/cancer/personalized-medicine/overview2. [Online; accessed 1-October-2016].

[3] "Wda. what is watson?." http://www.ibm.com/smarterplanet/us/en/ibmwatson/discovery- advisor.html. . [Online; accessed 2-October-2016].

[4] "Sequential pattern mining of electronic healthcare reimbursement claims: Experiences and challenges in uncovering how patients are treated by physicians." ©[2015] IEEE. Reprinted, with permission, from [Kunal Malhotra, Tanner C. Hobson, Silvia Valkova, Laura L. Pullum, Arvind Ramanathan], 2015.

[5] Afzal, Z., Engelkes, M., Verhamme, K., Janssens, H. M., Sturkenboom, M. C., Kors, J. A., and Schuemie, M. J., "Automatic generation of case-detection algorithms to identify children with asthma from large electronic health record databases," *Pharmacoepidemiology and Drug Safety*, vol. 22, no. 8, pp. 826–833, 2013.

[6] Agrawal, R. and Srikant, R., "Mining sequential patterns," in *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*, pp. 3–14, IEEE, 1995.

[7] Aloysius, G. and Binu, D., "An approach to products placement in supermarkets using prefixspan algorithm," *Journal of King Saud University-Computer and Information Sciences*, vol. 25, no. 1, pp. 77–87, 2013.

[8] Amarasingham, R., Moore, B. J., Tabak, Y. P., Drazner, M. H., Clark, C. A., Zhang, S., Reed, W. G., Swanson, T. S., Ma, Y., and Halm, E. A., "An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data," *Medical Care*, vol. 48, no. 11, pp. 981–988, 2010.

[9] Andany, J., Léonard, M., and Palisser, C., "Management of schema evolution in databases.," in *VLDB*, pp. 161–170, 1991.

[10] Aneja, S. and Sharma, S., "Newer anti-epileptic drugs," *Indian Pediatrics*, vol. 50, no. 11, pp. 1033–1040, 2013.

[11] Arnold, A., Simoons, M., Detry, J.-M., Von Essen, R., Van de Werf, F., Deckers, J., Lubsen, J., Verstraete, M., Group, E. C. S., and others, "Prediction of mortality following hospital discharge after thrombolysis for acute myocardial infarction: is there a need for coronary angiography?," *European Heart Journal*, vol. 14, no. 3, pp. 306–315, 1993.

[12] Aspden, P., Wolcott, J., Bootman, J. L., Cronenwett, L. R., and others, *Preventing medication errors: quality chasm series*. National Academies Press, 2006.

[13] Association, A. M., *CPT 2003: current procedural terminology*. Singular, 2002.

[14] Ayres, J., Flannick, J., Gehrke, J., and Yiu, T., "Sequential pattern mining using a bitmap representation," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge discovery and Data mining*, pp. 429–435, ACM, 2002.

[15] Banerjee, P. N., Filippi, D., and Hauser, W. A., "The descriptive epidemiology of epilepsy—a review," *Epilepsy Research*, vol. 85, no. 1, pp. 31–45, 2009.

[16] Barrows, R. C. and Clayton, P. D., "Privacy, confidentiality, and electronic medical records," *Journal of the American Medical Informatics Association*, vol. 3, no. 2, pp. 139–148, 1996.

[17] Bates, D. W. and Gawande, A. A., "Improving safety with information technology," *New England Journal of Medicine*, vol. 348, no. 25, pp. 2526–2534, 2003.

[18] Begley, C. E., Famulari, M., Annegers, J. F., Lairson, D. R., Reynolds, T. F., Coan, S., Dubinsky, S., Newmark, M. E., Leibson, C., So, E., and others, "The cost of epilepsy in the united states: An estimate from population-based clinical and survey data," *Epilepsia*, vol. 41, no. 3, pp. 342–351, 2000.

[19] Bellazzi, R., Ferrazzi, F., and Sacchi, L., "Predictive data mining in clinical medicine: a focus on selected methods and applications," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 5, pp. 416–430, 2011.

[20] Benaloh, J., Chase, M., Horvitz, E., and Lauter, K., "Patient controlled encryption: ensuring privacy of electronic medical records," in *Proceedings of the 2009 ACM workshop on Cloud Computing Security*, pp. 103–114, ACM, 2009.

[21] Berg, A. T. and Kelly, M. M., "Defining intractability: comparisons among published definitions," *Epilepsia*, vol. 47, no. 2, pp. 431–436, 2006.

[22] Berlingerio, M., Bonchi, F., Giannotti, F., and Turini, F., "Mining clinical data with a temporal dimension: a case study," in *Bioinformatics and Biomedicine, 2007. BIBM 2007. IEEE International Conference on*, pp. 429–436, IEEE, 2007.

[23] Bermingham, L. and Lee, I., "Spatio-temporal sequential pattern mining for tourism sciences," *Procedia Computer Science*, vol. 29, pp. 379–389, 2014.

[24] Berwick, D. M., "Disseminating innovations in health care," *JAMA*, vol. 289, no. 15, pp. 1969–1975, 2003.

[25] Bhattacharyya, S., "Snomed ct history and ihtsdo," in *Introduction to SNOMED CT*, pp. 19–23, Springer, 2016.

[26] Blumenthal, D., "Wiring the health system—origins and provisions of a new federal program," *New England Journal of Medicine*, vol. 365, no. 24, pp. 2323–2329, 2011.

[27] Blumenthal, D. and Tavenner, M., "The "meaningful use" regulation for electronic health records," *New England Journal of Medicine*, vol. 363, no. 6, pp. 501–504, 2010.

[28] Bradley, E. and Taylor, L., *The American health care paradox: Why spending more is getting us less.* PublicAffairs, 2013.

[29] Brailer, D. J., "Interoperability: the key to the future health care system," *Health Affairs*, vol. 24, p. W5, 2005.

[30] Brewer, E., "Pushing the cap: Strategies for consistency and availability," *Computer*, vol. 45, no. 2, pp. 23–29, 2012.

[31] Campbell, A. J., Cook, J. A., Adey, G., and Cuthbertson, B., "Predicting death and readmission after intensive care discharge," *British Journal Of Anaesthesia*, 2008.

[32] Castro-Rodriguez, J. A., "The asthma predictive index: early diagnosis of asthma," *Current opinion in allergy and clinical immunology*, vol. 11, no. 3, pp. 157–161, 2011.

[33] Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., and others, "The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data," *Cancer Discovery*, vol. 2, no. 5, pp. 401–404, 2012.

[34] Charlson, M. E., Pompei, P., Ales, K. L., and MacKenzie, C. R., "A new method of classifying prognostic comorbidity in longitudinal studies: development and validation," *Journal of Chronic Diseases*, vol. 40, no. 5, pp. 373–383, 1987.

[35] Chittaro, L., Combi, C., and Trapasso, G., "Data mining on temporal data: a visual approach and its clinical application to hemodialysis," *Journal of Visual Languages & Computing*, vol. 14, no. 6, pp. 591–620, 2003.

[36] Cleve, A., Gobert, M., Meurice, L., Maes, J., and Weber, J., "Understanding database schema evolution: A case study," *Science of Computer Programming*, vol. 97, pp. 113–121, 2015.

[37] Collins, F. S. and Varmus, H., "A new initiative on precision medicine," *New England Journal of Medicine*, vol. 372, no. 9, pp. 793–795, 2015.

[38] Concaro, S., Sacchi, L., Cerra, C., Fratino, P., and Bellazzi, R., "Mining healthcare data with temporal association rules: Improvements and assessment for a practical use," in *Conference on Artificial Intelligence in Medicine in Europe*, pp. 16–25, Springer, 2009.

[39] Cost, H., Project, U., and others, "Clinical classifications software (ccs) for icd-9-cm," *Rockville, MD: Agency for Healthcare Research and Quality*, 2010.

[40] Cuthbert, B. N. and Insel, T. R., "Toward the future of psychiatric diagnosis: the seven pillars of rdoc," *BMC Medicine*, vol. 11, no. 1, p. 1, 2013.

[41] D Smith, D. C., "The management of epilepsy," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 70, no. 2, 2001.

[42] Davis, K., Stremikis, K., Schoen, C., and Squires, D., "Mirror, mirror on the wall, 2014 update: how the us health care system compares internationally," *The Commonwealth Fund*, vol. 16, 2014.

[43] DeNavas-Walt, C., Proctor, B. D., and Smith, J. C., "Us census bureau, current population reports, p60-238," *Income, poverty, and health insurance coverage in the United States: 2009*, 2010.

[44] Developmental, D. M. N. S. Y., Investigators, . P., and others, "Prevalence of autism spectrum disorder among children aged 8 years-autism and developmental disabilities monitoring network, 11 sites, united states, 2010.," *Morbidity and mortality weekly report. Surveillance summaries (Washington, DC: 2002)*, vol. 63, no. 2, p. 1, 2014.

[45] Devinsky, O., "Patients with refractory seizures," *New England Journal of Medicine*, vol. 340, no. 20, pp. 1565–1570, 1999.

[46] Devinsky, O., Dilley, C., Ozery-Flato, M., Aharonov, R., Goldschmidt, Y., Rosen-Zvi, M., Clark, C., and Fritz, P., "Changing the approach to treatment choice in epilepsy using big data," *Epilepsy & Behavior*, vol. 56, pp. 32–37, 2016.

[47] Donzé, J., Aujesky, D., Williams, D., and Schnipper, J. L., “Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model,” *JAMA Internal Medicine*, vol. 173, no. 8, pp. 632–638, 2013.

[48] Elixhauser, A., Steiner, C., and Palmer, L., “Clinical classifications software (ccs),” *Book Clinical Classifications Software (CCS)(Editor ed^ eds)*, 2008.

[49] Elmasri, R. and Navathe, S., *Fundamentals of database systems.* Pearson, 7 ed., 2016.

[50] Ferrandina, F., Meyer, T., Zicari, R., Ferran, G., and Madec, J., “Schema and database evolution in the o˜ 2 object database system,” in *VLDB*, vol. 95, pp. 170–181, Citeseer, 1995.

[51] Fisher, R. S., Harding, G., Erba, G., Barkley, G. L., and Wilkins, A., “Photic-and pattern-induced seizures: A review for the epilepsy foundation of america working group,” *Epilepsia*, vol. 46, no. 9, pp. 1426–1441, 2005.

[52] Fitz Henry, F., Murff, H. J., Matheny, M. E., Gentry, N., Fielstein, E. M., Brown, S. H., Reeves, R. M., Aronsky, D., Elkin, P. L., Messina, V. P., and others, “Exploring the frontier of electronic health record surveillance: the case of post-operative complications,” *Medical Care*, vol. 51, no. 6, p. 509, 2013.

[53] for Clinical Excellence, N. I., Britain, G., and others, *Guidance on the use of temozolomide for the treatment of recurrent malignant glioma (brain cancer).* National Institute for Clinical Excellence, 2001.

[54] for Disease Control, C., Prevention, and others, “Chronic disease overview,” *Online factsheet*, 2015.

[55] Forsgren, L., Beghi, E., Oun, A., and Sillanpää, M., “The epidemiology of epilepsy in europe–a systematic review,” *European Journal of Neurology*, vol. 12, no. 4, pp. 245–253, 2005.

[56] Freije, W. A., Castro-Vargas, F. E., Fang, Z., Horvath, S., Cloughesy, T., Liau, L. M., Mischel, P. S., and Nelson, S. F., “Gene expression profiling of gliomas strongly predicts survival,” *Cancer Research*, vol. 64, no. 18, pp. 6503–6510, 2004.

[57] French, J. A., “Refractory epilepsy: clinical overview,” *Epilepsia*, vol. 48, no. s1, pp. 3–7, 2007.

[58] Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., and others, “Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal,” *Science Signaling*, vol. 6, no. 269, p. p11, 2013.

[59] Garde, S., Knaup, P., Hovenga, E. J., and Heard, S., "Towards semantic interoperability for electronic health records–domain knowledge governance for open ehr archetypes," *Methods of Information in Medicine*, vol. 46, no. 3, pp. 332–343, 2007.

[60] Garg, A. X., Adhikari, N. K., McDonald, H., Rosas-Arellano, M. P., Devereaux, P., Beyene, J., Sam, J., and Haynes, R. B., "Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review," *JAMA*, vol. 293, no. 10, pp. 1223–1238, 2005.

[61] Garrett, B. E., Dube, S. R., Trosclair, A., Caraballo, R. S., Pechacek, T. F., for Disease Control, C., (CDC), P., and others, "Cigarette smoking—united states, 1965–2008," *MMWR Surveill Summ*, vol. 60, no. 1, pp. 109–113, 2011.

[62] Gastaut, H., Gastaut, J., Silva, G. E., and Sanchez, G., "Relative frequency of different types of epilepsy: a study employing the classification of the international league against epilepsy," *Epilepsia*, vol. 16, no. 3, pp. 457–461, 1975.

[63] Gildersleeve, R., Cooper, P., and others, "Development of an automated, real time surveillance tool for predicting readmissions at a community hospital," *Applied Clinical Informatics*, vol. 4, no. 2, pp. 153–169, 2013.

[64] Gotz, D. and Stavropoulos, H., "Decisionflow: Visual analytics for high-dimensional temporal event sequence data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1783–1792, 2014.

[65] Gotz, D., Wang, F., and Perer, A., "A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data," *Journal of Biomedical Informatics*, vol. 48, pp. 148–159, 2014.

[66] Group, U. C. S. W. and others, "United states cancer statistics: 1999-2011 incidence and mortality web-based report. atlanta, ga: Us dhhs, cdc, national cancer institute. 2014," 2015.

[67] Hachinski, V., Iadecola, C., Petersen, R. C., Breteler, M. M., Nyenhuis, D. L., Black, S. E., Powers, W. J., DeCarli, C., Merino, J. G., Kalaria, R. N., and others, "National institute of neurological disorders and stroke–canadian stroke network vascular cognitive impairment harmonization standards," *Stroke*, vol. 37, no. 9, pp. 2220–2241, 2006.

[68] Han, J., Pei, J., and Kamber, M., *Data mining: concepts and techniques.* Morgan Kaufmann, 2 ed., July 2011.

[69] Hanauer, D. A. and Ramakrishnan, N., "Modeling temporal relationships in large scale clinical associations," *Journal of the American Medical Informatics Association*, vol. 20, no. 2, pp. 332–341, 2013.

[70] Hegi, M. E., Diserens, A.-C., Gorlia, T., Hamou, M.-F., de Tribolet, N., Weller, M., Kros, J. M., Hainfellner, J. A., Mason, W., Mariani, L., and others, “Mgmt gene silencing and benefit from temozolomide in glioblastoma,” *New England Journal of Medicine*, vol. 352, no. 10, pp. 997–1003, 2005.

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

[72] Ho, T. K., “Random decision forests,” in *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, vol. 1, pp. 278–282, IEEE, 1995.

[73] Holroyd-Leduc, J. M., Lorenzetti, D., Straus, S. E., Sykes, L., and Quan, H., “The impact of the electronic medical record on structure, process, and outcomes within primary care: a systematic review of the evidence,” *Journal of the American Medical Informatics Association*, vol. 18, no. 6, pp. 732–737, 2011.

[74] Hougland, P., Nebeker, J., Pickard, S., Van Tuinen, M., Masheter, C., Elder, S., Williams, S., and Xu, W., “Using icd-9-cm codes in hospital claims data to detect adverse events in patient safety surveillance,” *Advances in Patient Safety: New Directions and Alternative Approaches*, vol. 1, 2008.

[75] Hoyt, R. E. and Yoshihashi, A. K., *Health Informatics: Practical Guide for Healthcare and Information Technology Professionals.* Informatics Education, 6 ed., 2014.

[76] Hu, J., Perer, A., and Wang, F., “Data driven analytics for personalized healthcare,” in *Healthcare Information Management Systems*, pp. 529–554, Springer, 2016.

[77] Hunt, D. L., Haynes, R. B., Hanna, S. E., and Smith, K., “Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review,” *JAMA*, vol. 280, no. 15, pp. 1339–1346, 1998.

[78] Jacobs, M. P., Leblanc, G. G., Brooks-Kayal, A., Jensen, F. E., Lowenstein, D. H., Noebels, J. L., Spencer, D. D., and Swann, J. W., “Curing epilepsy: progress and future directions,” *Epilepsy & Behavior*, vol. 14, no. 3, pp. 438–445, 2009.

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

[80] Kahn, M. G., Batson, D., and Schilling, L. M., "Data model considerations for clinical effectiveness researchers," *Medical Care*, vol. 50, 2012.

[81] Katsnelson, A., "Momentum grows to make'personalized'medicine more'precise'," *Nature Medicine*, vol. 19, no. 3, pp. 249–249, 2013.

[82] Kaushal, R., Shojania, K. G., and Bates, D. W., "Effects of computerized physician order entry and clinical decision support systems on medication safety: a systematic review," *Archives of Internal Medicine*, vol. 163, no. 12, pp. 1409–1416, 2003.

[83] Kim, W., Ballou, N., Chou, H.-T., Garza, J. F., and Woelk, D., "Features of the orion object-oriented database system," in *Object-oriented Concepts, Databases, and Applications*, pp. 251–282, ACM Press, 1989.

[84] Kim, W. and Chou, H.-T., "Versions of schema for object-oriented databases.," in *VLDB International Conference*, pp. 148–159, 1988.

[85] Kolp, M. and Zimanyi, E., "Enhanced er to relational mapping and interrelational normalization," *Information and Software Technology*, vol. 42, no. 15, pp. 1057–1073, 2000.

[86] Krex, D., Klink, B., Hartmann, C., von Deimling, A., Pietsch, T., Simon, M., Sabel, M., Steinbach, J. P., Heese, O., Reifenberger, G., and others, "Long-term survival with glioblastoma multiforme," *Brain*, vol. 130, no. 10, pp. 2596–2606, 2007.

[87] Kwan, P. and Brodie, M. J., "Early identification of refractory epilepsy," *New England Journal of Medicine*, vol. 342, no. 5, pp. 314–319, 2000.

[88] Laxer, K. D., Trinka, E., Hirsch, L. J., Cendes, F., Langfitt, J., Delanty, N., Resnick, T., and Benbadis, S. R., "The consequences of refractory epilepsy and its treatment," *Epilepsy & Behavior*, vol. 37, pp. 59–70, 2014.

[89] Leavitt, N., "Will nosql databases live up to their promise?," *Computer*, vol. 43, no. 2, pp. 12–14, 2010.

[90] Lenfant, C., "Clinical research to clinical practice—lost in translation?," *New England Journal of Medicine*, vol. 349, no. 9, pp. 868–874, 2003.

[91] Lhatoo, S., Sander, J., and Shorvon, S., "The dynamics of drug treatment in epilepsy: an observational study in an unselected population based cohort with newly diagnosed epilepsy followed up prospectively over 11–14 years," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 71, no. 5, pp. 632–637, 2001.

[92] Liang, Y., Diehn, M., Watson, N., Bollen, A. W., Aldape, K. D., Nicholas, M. K., Lamborn, K. R., Berger, M. S., Botstein, D., Brown, P. O., and others, "Gene expression profiling reveals molecularly and clinically distinct subtypes of glioblastoma multiforme," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 16, pp. 5814–5819, 2005.

[93] Lobach, D. F. and Hammond, W. E., "Computerized decision support based on a clinical practice guideline improves compliance with care standards," *The American Journal of Medicine*, vol. 102, no. 1, pp. 89–98, 1997.

[94] Ma, S., Song, X., and Huang, J., "Supervised group lasso with applications to microarray data analysis," *BMC Bioinformatics*, vol. 8, no. 1, p. 1, 2007.

[95] Malhotra, K., Hobson, T. C., Valkova, S., Pullum, L. L., and Ramanathan, A., "Sequential pattern mining of electronic healthcare reimbursement claims: Experiences and challenges in uncovering how patients are treated by physicians," in *IEEE International Conference on Big Data*, pp. 2670–2679, IEEE, 2015.

[96] Malhotra, K., Medhekar, S., Navathe, S. B., and Laborde, M. D., "Towards a form based dynamic database schema creation and modification system," in *International Conference on Advanced Information Systems Engineering*, pp. 595–609, Springer, 2014.

[97] Malhotra, K., Navathe, S. B., Chau, D. H., Hadjipanayis, C., and Sun, J., "Constraint based temporal event sequence mining for glioblastoma survival prediction," *J Biomed Inform*, vol. 61, pp. 267–75, Jun 2016.

[98] Mandl, K. D., Markwell, D., MacDonald, R., Szolovits, P., and Kohane, I. S., "Public standards and patients' control: how to keep electronic medical records accessible but privatemedical information: access and privacydoctrines for developing electronic medical recordsdesirable characteristics of electronic medical recordschallenges and limitations for electronic medical recordsconclusionscommentary: Open approaches to electronic patient recordscommentary: A patient's viewpoint," *BMJ*, vol. 322, no. 7281, pp. 283–287, 2001.

[99] Mazurowski, M. A., Desjardins, A., and Malof, J. M., "Imaging descriptors improve the predictive power of survival models for glioblastoma patients," *Neuro-oncology*, vol. 15, no. 10, pp. 1389–1394, 2013.

[100] McDonald, C. J., "The barriers to electronic medical record systems and how to overcome them," *Journal of the American Medical Informatics Association*, vol. 4, no. 3, pp. 213–221, 1997.

[101] Miller, G. A., "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[102] Miller, R. H. and Sim, I., "Physicians' use of electronic medical records: barriers and solutions," *Health Affairs*, vol. 23, no. 2, pp. 116–126, 2004.

[103] Mischel, P. S., Shai, R., Shi, T., Horvath, S., Lu, K. V., Choe, G., Seligson, D., Kremen, T. J., Palotie, A., Liau, L. M., and others, "Identification of molecular subtypes of glioblastoma by gene expression profiling," *Oncogene*, vol. 22, no. 15, pp. 2361–2373, 2003.

[104] Moskovitch, R. and Shahar, Y., "Medical temporal-knowledge discovery via temporal abstraction.," in *AMIA*, 2009.

[105] Murat, A., Migliavacca, E., Gorlia, T., Lambiv, W. L., Shay, T., Hamou, M.-F., De Tribolet, N., Regli, L., Wick, W., Kouwenhoven, M. C., and others, "Stem cell–related "self-renewal" signature and high epidermal growth factor receptor expression associated with resistance to concomitant chemoradiotherapy in glioblastoma," *Journal of Clinical Oncology*, vol. 26, no. 18, pp. 3015–3024, 2008.

[106] Murphy, D. R., Laxmisan, A., Reis, B. A., Thomas, E. J., Esquivel, A., Forjuoh, S. N., Parikh, R., Khan, M. M., and Singh, H., "Electronic health record-based triggers to detect potential delays in cancer diagnosis," *BMJ quality & safety*, vol. 23, no. 1, pp. 8–16, 2014.

[107] Murphy, S., Xu, J., and Kochanek, K., "Deaths: Final data for 2010. national vital statistics reports," *National Center for Health Statistics*, vol. 61, no. 4, 2013.

[108] Network, T., "The cancer genome atlas data portal," *National Institute of Health*, 2010.

[109] Nutt, C. L., Mani, D., Betensky, R. A., Tamayo, P., Cairncross, J. G., Ladd, C., Pohl, U., Hartmann, C., McLaughlin, M. E., Batchelor, T. T., and others, "Gene expression-based classification of malignant gliomas correlates better with survival than histological classification," *Cancer Research*, vol. 63, no. 7, pp. 1602–1607, 2003.

[110] Organization, W. H. and others, "World cancer report 2014 (epub)," *World Health Organization Press: Lyon, France*, 2014.

[111] Ouh, J.-Z., Wu, P.-H., and Chen, M.-S., "Experimental results on a constraint based sequential pattern mining for telecommunication alarm data," in *Web Information Systems Engineering, 2001. Proceedings of the Second International Conference on*, vol. 2, pp. 186–193, IEEE, 2001.

[112] Parsons, D. W., Jones, S., Zhang, X., Lin, J. C.-H., Leary, R. J., Angenendt, P., Mankoo, P., Carter, H., Siu, I.-M., Gallia, G. L., and others, "An integrated genomic analysis of human glioblastoma multiforme," *Science*, vol. 321, no. 5897, pp. 1807–1812, 2008.

[113] Patnaik, D., Butler, P., Ramakrishnan, N., Parida, L., Keller, B. J., and Hanauer, D. A., "Experiences with mining temporal event sequences from electronic medical records: initial successes and some challenges," in *Proc. KDD Conf.*, pp. 360–368, ACM, 2011.

[114] Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., Dayal, U., and Hsu, M.-C., "Mining sequential patterns by pattern-growth: The prefixspan approach," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 11, pp. 1424–1440, 2004.

[115] Perotte, A., Pivovarov, R., Natarajan, K., Weiskopf, N., Wood, F., and Elhadad, N., "Diagnosis code assignment: models and evaluation metrics," *Journal of the American Medical Informatics Association*, vol. 21, no. 2, pp. 231–237, 2014.

[116] Phillips, H. S., Kharbanda, S., Chen, R., Forrest, W. F., Soriano, R. H., Wu, T. D., Misra, A., Nigro, J. M., Colman, H., Soroceanu, L., and others, "Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis," *Cancer cell*, vol. 9, no. 3, pp. 157–173, 2006.

[117] Protection, P. and Act, A. C., "Public law 111-148," *Title IV, x4207, USC HR*, vol. 3590, p. 2010, 2010.

[118] Quinlan, J. R., "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[119] Ramanathan, A., Pullum, L., Steed, C., Quinn, S., and Chennubhotla, C., "Oak ridge bio-surveillance toolkit," in *IEEE VAST Workshop on Public Health's Wicked Problems: Can InfoVis Save Lives*, 2013.

[120] Ramanathan, A., Pullum, L. L., Hobson, T. C., Stahl, C. G., Steed, C. A., Quinn, S. P., Chennubhotla, C. S., and Valkova, S., "Discovering multi-scale co-occurrence patterns of asthma and influenza with oak ridge bio-surveillance toolkit," *Frontiers in Public Health*, vol. 3, 2015.

[121] Ramanathan, A., Pullum, L. L., Hobson, T. C., Steed, C. A., Quinn, S. P., Chennubhotla, C. S., and Valkova, S., "Orbit: Oak ridge bio-surveillance toolkit for public health dynamics," *BMC Bioinformatics*, vol. 16, no. 17, p. 1, 2015.

[122] Rivera, A. L., Pelloski, C. E., Gilbert, M. R., Colman, H., De La Cruz, C., Sulman, E. P., Bekele, B. N., and Aldape, K. D., "Mgmt promoter methylation is predictive of response to radiotherapy and prognostic in the absence of adjuvant alkylating chemotherapy for glioblastoma," *Neuro-oncology*, pp. 116–121, 2009.

[123] Robinson, I., Webber, J., and Eifrem, E., *Graph Databases: New Opportunities for Connected Data.* " O'Reilly Media, Inc.", 2015.

[124] Rohan, M., "Healthcare analytics market worth $18.7 billion by 2020." Retrieved from marketsandmarkets. com: http://www. marketsandmarkets. com/PressReleases/healthcare-data-analytics. asp.

[125] Ruano, Y., Mollejo, M., Ribalta, T., Fiaño, C., Camacho, F. I., Gómez, E., de Lope, A. R., Hernández-Moneo, J.-L., Martínez, P., and Meléndez, B., "Identification of novel candidate target genes in amplicons of glioblastoma multiforme tumors detected by expression and cgh microarray profiling," *Molecular Cancer*, vol. 5, no. 1, p. 1, 2006.

[126] Saeed, M., Lieu, C., Raber, G., and Mark, R. G., "Mimic ii: a massive temporal icu patient database to support research in intelligent patient monitoring," in *Computers in Cardiology, 2002*, pp. 641–644, IEEE, 2002.

[127] Salcman, M., "Glioblastoma and malignant astrocytoma," *Brain Tumors*, vol. 449, 1995.

[128] Sander, J. and Shorvon, S., "Epidemiology of the epilepsies.," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 61, no. 5, pp. 433–443, 1996.

[129] Sanders, D., Burton, D. A., and Protti, D., "The healthcare analytics adoption model: A framework and roadmap," *Health Catalyst*, 2013.

[130] Sasso, R. C., Reilly, T. M., Resnick, D., and Haid, R., "Anterior lumbar interbody fusion: threaded bone dowels versus titanium cages," *Surgical Management of Low Back Pain. Rolling Meadows, IL: American Association of Neurological Surgeons*, pp. 103–116, 2001.

[131] Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., and Chute, C. G., "Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, 2010.

[132] Schmidt, D., "Drug treatment of epilepsy: options and limitations," *Epilepsy & Behavior*, vol. 15, no. 1, pp. 56–65, 2009.

[133] Scott, R. A., Lhatoo, S. D., and Sander, J. W., "The treatment of epilepsy in developing countries: where do we go from here?," *Bulletin of the World Health Organization*, vol. 79, no. 4, pp. 344–351, 2001.

[134] Shai, R., Shi, T., Kremen, T. J., Horvath, S., Liau, L. M., Cloughesy, T. F., Mischel, P. S., and Nelson, S. F., "Gene expression profiling identifies molecular subtypes of gliomas," *Oncogene*, vol. 22, no. 31, pp. 4918–4923, 2003.

[135] Singh, A., Nadkarni, G., Guttag, J., and Bottinger, E., "Leveraging hierarchy in medical codes for predictive modeling," in *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 96–103, ACM, 2014.

[136] Skarra, A. H. and Zdonik, S. B., "The management of changing types in an object-oriented database," in *ACM Sigplan Notices*, vol. 21, pp. 483–495, ACM, 1986.

[137] Sledge Jr, G., Miller, R., and Hauser, R., "Cancerlinq and the future of cancer care.," in *American Society of Clinical Oncology educational book/ASCO. American Society of Clinical Oncology. Meeting*, pp. 430–434, 2012.

[138] Slee, V. N., "The international classification of diseases: ninth revision (icd-9)," *Annals of Internal Medicine*, vol. 88, no. 3, pp. 424–426, 1978.

[139] Solomon, N. and McHale, K., "An overview of epilepsy in children and young people," *Nursing Children and Young People*, vol. 24, no. 6, p. 28, 2012.

[140] Sperling, M. R., "The consequences of uncontrolled epilepsy," *CNS spectrums*, vol. 9, no. 02, pp. 98–109, 2004.

[141] Squires, D., "The global slowdown in health care spending growth," *JAMA*, vol. 312, no. 5, pp. 485–486, 2014.

[142] Squires, D. and Anderson, C., "Us health care from a global perspective: spending, use of services, prices, and health in 13 countries.," *Issue brief (Commonwealth Fund)*, vol. 15, pp. 1–15, 2015.

[143] Srikant, R. and Agrawal, R., "Mining sequential patterns: Generalizations and performance improvements," in *International Conference on Extending Database Technology*, pp. 1–17, Springer, 1996.

[144] St Germaine-Smith, C., Liu, M., Quan, H., Wiebe, S., and Jette, N., "Development of an epilepsy-specific risk adjustment comorbidity index," *Epilepsia*, vol. 52, no. 12, pp. 2161–2167, 2011.

[145] Svanström, H., Callréus, T., and Hviid, A., "Temporal data mining for adverse events following immunization in nationwide danish healthcare databases," *Drug Safety*, vol. 33, no. 11, pp. 1015–1025, 2010.

[146] Taylor, D., Naguib, R., and Boulton, S., "A dynamic clinical dental relational database," *IEEE Transactions on Information Technology in Biomedicine*, vol. 8, no. 3, pp. 298–305, 2004.

[147] Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., and Jemal, A., "Global cancer statistics, 2012," *CA: a cancer journal for clinicians*, vol. 65, no. 2, pp. 87–108, 2015.

[148] Tso, C.-L., Freije, W. A., Day, A., Chen, Z., Merriman, B., Perlina, A., Lee, Y., Dia, E. Q., Yoshimoto, K., Mischel, P. S., and others, "Distinct transcription profiles of primary and secondary glioblastoma subgroups," *Cancer Research*, vol. 66, no. 1, pp. 159–167, 2006.

[149] Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., Miller, C. R., Ding, L., Golub, T., Mesirov, J. P., and others, "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1," *Cancer Cell*, vol. 17, no. 1, pp. 98–110, 2010.

[150] Walid, M. S., "Prognostic factors for long-term survival after glioblastoma," *The Permanente Journal*, vol. 12, no. 4, pp. 45–48, 2008.

[151] Weatherspoon, D. and Chattopadhyay, A., "International classification of diseases codes and their use in dentistry," *Journal of Dental, Oral and Craniofacial Epidemiology*, vol. 1, no. 4, p. 20, 2013.

[152] Wen, P. Y. and Kesari, S., "Malignant gliomas in adults," *New England Journal of Medicine*, vol. 359, no. 5, pp. 492–507, 2008.

[153] Williams, G. J. and Simoff, S. J., *Data mining: Theory, methodology, techniques, and applications*, vol. 3755. Springer, 2006.

[154] Wright, G., "Data modeling in healthcare industry." [White Paper] Retrieved from http://www.teradata.com/resources/white-papers/data-modeling-in-the-healthcare-industry-eb5301/?type=WPLangType=1033LangSelect=true, 2016.

[155] Wu, P.-H., Peng, W.-C., and Chen, M.-S., "Mining sequential alarm patterns in a telecommunication database," in *International Workshop on Databases in Telecommunications*, pp. 37–51, Springer, 2001.

[156] Yan, Y., Fung, G., Dy, J. G., and Rosales, R., "Medical coding classification by leveraging inter-code relationships," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 193–202, ACM, 2010.

[157] Yuan, L., Liu, J., and Ye, J., "Efficient methods for overlapping group lasso," in *Advances in Neural Information Processing Systems*, pp. 352–360, 2011.

[158] Zaki, M. J., "Spade: An efficient algorithm for mining frequent sequences," *Machine Learning*, vol. 42, no. 1-2, pp. 31–60, 2001.

[159] Zhao, P., Rocha, G., and Yu, B., "The composite absolute penalties family for grouped and hierarchical variable selection," *The Annals of Statistics*, pp. 3468–3497, 2009.

[160] Zhao, Q. and Bhowmick, S. S., "Sequential pattern mining: A survey," *ITechnical Report CAIS Nayang Technological University Singapore*, pp. 1–26, 2003.

[161] Zhou, J., Wang, F., Hu, J., and Ye, J., "From micro to macro: data driven phenotyping by densification of longitudinal electronic medical records," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 135–144, ACM, 2014.

[162] Zhou, L., Rundensteiner, E. A., and Shin, K. G., "Schema evolution for real-time object-oriented databases," technical report, University of Michigan, Computer Science and Engineering Division, Department of Electrical Engineering and Computer Science, 1994.