A Personalized Decision Support System for Medical Treatment Planning

Kunal Malhotra

College of Computing, Georgia Institute of Technology, Atlanta, Georgia, USA
kmalhotra7@gatech.edu

Abstract. 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 and at most hospitals, a lot of variation exists in the treatment plans subsequently chosen. We propose to extensively study recommended treatment guidelines and evaluate historical treatment data for selected rare and chronic diseases. As representative diseases we study Glioblastoma Multiforme (brain cancer) which is classified as a rare disease, and Diabetes Mellitus, which is a nationally and globally widespread chronic disease. A graph model is designed to capture the data pertaining to the treatment options and actual treatments administered and further analyzed to discover sequential treatment patterns based on different outcome classes based on longevity, complications etc. The notion of 'Patient Similarity' would be explored to form cohorts of clinically similar patients. The treatment patterns would be ranked, and highly ranked patterns would be ordered depending on expected outcomes before being assigned to cohorts of patients. A prototype decision support system is planned for recommending treatment options based on a patients clinical and possibly genomic data when available.

Keywords: Graph Data, Sequential Pattern Mining, Patient Similarity

1 Introduction

Evidence-based medicine refers to the explicit and exhaustive use of available medical evidence to improve quality of care provided. It involves integration of individual clinical expertise with the best available external clinical evidence from systematic research [17]. When deciding on the treatments for patients, medical practitioners consult both clinical evidence and use their own clinical judgement and experience to inform their decisions and recommendations to patients. Variability in such recommendations among multiple practitioners could directly affect a patient's wellbeing and recovery. A key challenge faced by a physician is determining the optimal treatment for a given patient. While protocols exist for first and second line treatments for the vast majority of diseases, tailoring treatments to an individual patient presents a huge diagnostic challenge. With the ever-expanding treatment options, there is an increasing need for a system to identify treatment patterns and new drugs that would cure the patient in the most efficient manner. Treatment is also a dynamic process, and must evolve as new information becomes available such as response to a particular treatment approach, which may be positive or adverse during the course of treatment.

In acute conditions, the goal is to find a "cure" or treatment to reverse or at least arrest the progression of disease typically manifested in a variety of aggressive cancers. On the other hand, in chronic diseases like diabetes or renal disease, the goal is to manage the disease state and prevent further deterioration or progression to more serious conditions. Part of chronic disease management is to identify and prescribe measures for improving the quality of life. Toward this end, we have selected two specific diseases for analysis, namely Glioblastoma Multiforme (GBM), and Diabetes Mellitus. GBM is the most lethal type of brain cancer and is biologically the most aggressive subtype of malignant gliomas. The current standard of care for GBM patients involves surgical resection followed by radiation and chemotherapy with an oral alkylating agent Temodar [14]. The median survival period for GBM patients is one year after diagnosis. This extreme mortality rate, where none have a long-term survival, has drawn significant attention to improving treatment for patients with these tumors. Chronic diseases are among the worlds leading causes of death. Diabetes is a complex disease, often found to co-occur with other chronic conditions such as hypertension and depression, which not only complicates diabetes management and but also increases the risk of developing diabetes by 60% [2]. Disparities in treatment exist among diabetes patients in the U.S and majority of them don't achieve recommended guidelines. [18].

With the enforcement of electronic medical records(EMRs), a vast amount of healthcare data is being captured. Using various techniques in data mining we can improve decision-making. The results of healthcare data analysis can influence cost, revenue and quality of care [8]. Decision trees have been used extensively based on extensive medical data and clinical evidence to develop decision support systems [20]. For this dissertation, we aim at using techniques of information processing and data mining to make use of historical data about patients, guidelines of treatments given to them, health outcomes, complications, etc to develop a set of algorithms and implement them in a decision support tool.

2 Problem Definition and Goals of Research

We propose to develop a framework to evaluate efficiency of treatments in terms of outcomes using clinical patient data and make personalized treatment recommendations to patients. Outcomes of interest include likelihood of survival, longevity, probability of hospital readmission within 30 days, etc. Short and long term economic outcomes are equally important. However we are not likley to be able to incorporate cost data into our analaysis, we identify characteristic treatment patterns and associate these patterns with patients having similar outcomes.

So far we have been working with GBM which is a rare disease. We would like to extend our scope to chronic diseases like Diabetes Mellitus to investigate the validity and accuracy of prediction in both cases. After preliminary discussions with domain experts we anticipate that our general approach will be equally applicable in multiple acute and chronic clinical situations where preliminary and secondary treatments are well understood, but subsequent treatment may have too many options. We identify clinical and behavioral similarities among patients and create patient cohorts based on the similarities found. A clinically relevant distance measure needs to be developed to perform patient profiling [23]. This 'patient similarity assessment needs extensive exploration for studying the treatments prescribed to different cohorts and help in treatment comparison, management of patients in groups and prediction. Heuristics would be developed to rank the identified treatment patterns for an individual patient based on the extent of similarity a new patient shares with classes of patients. The model would also be enhanced to incorporate various responses to treatments. The goals of my research are as follows:-

1. **Understanding the disease:** A comprehensive understanding of how the disease affects people, the preferred first and second line of treatments, variations in standard of care, treatment options other than the standard of care, and the circumstances under which a particular course of treatment is prescribed.
2. **Analysis of treatments administered:** Historic treatment data would be represented in a graph and extensively analyzed to develop a decision support model.
3. **Creating a decision support tool:** A tool would be built to assist medical professionals in deciding the most probable treatment patterns for individual patients by learning from historical data. There are three subgoals here: a) Recognition of treatment patterns characteristic of a particular outcome or multiple outcomes; b) Defining a similarity measure for patients based on their clinical characteristics; and c) Developing a ranking algorithm to order the treatment patterns to achieve the best outcomes for individual patients or patient cohorts based on their similarity to classes of patients.
4. **Evaluation of the decision support tool:** Evaluation of the tool would be done by training and testing our algorithms on datasets provided by our partner institutions in this study. The treatments prescribed by the doctors would be compared with the one recommended and ranked by our tool. Validation of our approach would be done using various measures such as receiver operating characteristic curve, accuracy, precision, and recall.

3 Methods and Techniques

3.1 Data Representation

A prototype of the tool is being developed for GBM patients using clinical and genomic data from a public portal called 'The Cancer Genome Atlas Portal' [10] and cBioPortal [3]. The clinical domain includes demographic information about the patient along with some basic clinical features, e.g Karnofsky performance score, histological type, survival duration, prior glioma information and most importantly the vital status of the patient (Living / Dead). Studies show that GBM patients can be classified into four subtypes namely Classical, Mesenchymal, Proneural and Neural based on the expression levels of a particular set of genes [6, 22]. For our study we considered these set of genes and used their mRNA expression levels, copy number varation data and methylation status. Additional information includes drugs prescribed along with their dosage, therapy type, radiation type, radiation dosage, and start and end dates for the treatment. We model this data as a graph where nodes are of two types: 'patient node' & 'treatment type node' and edges are also of two types: 'prescription edge' & 'sequence edge'. A graph offers a much richer picture of a network, and relationships of several types. The majority of path-based graph database operations are highly aligned with the way in which the data is laid out hence increasing the efficiency [16]. Figure 1 shows a graph consisting of two patients just for illustrative purposes. In the graph patient nodes have properties such as 'patient id', 'age', etc. Drugs and radiation prescribed are represented as treatment type nodes with properties such as 'drug name' and 'radiation type' respectively. The 'prescription edge' signifies the prescription of treatment with properties such as 'start date of prescription', 'end date of prescription', 'dosage', etc. The 'sequence edge' signifies the sequence in which drugs or radiation were prescribed. E.g., The edge labeled 'Prescribed' between the patient node with 'id = Patient_1' and the drug node with 'drugName = Drug_A' signifies that 'Patient_1' was prescribed 200 mg/day of 'Drug_A' on 05/21/2007 till 06/22/2007. The other type of edge labeled 'Followed_by' would always be between two drugs or two types of radiation or between a radiation type and a drug signifying the sequence of the prescription. E.g., the 'Followed_by' edge between source node 'Drug_A' and target node 'Drug_B' with properties 'patient' and 'overlap' signifies that for 'Patient_1', Drug_A was followed by Drug_B and there was an overlap of 24 days. The graph shown in the above figure is based on the data available for GBM patients. We expect that we will need a more complex graph structure that accounts for comorbidities such as hypertension and clinical depression yielding more parameters and potential challenges for graph mining.

3.2 Approach

Our approach is driven by the outcome of the treatment. The objective is to increase the survival period of the patient as much as possible. For GBM patients we analyze the treatment data of patients and identify treatment patterns, which

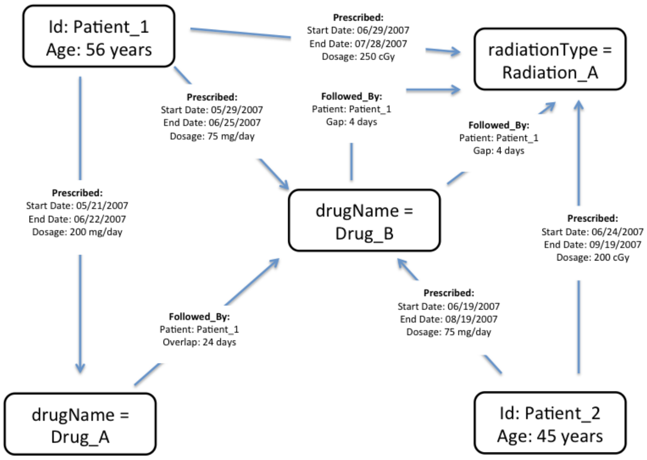

Fig. 1. Data represented as a graph

are characteristic of survival for a certain range of time (e.g., 6-12 months). The notion of 'patient similarity' would be explored to either use existing similarity measures or develop a new metric for the purpose of identifying similar patients, which would play a significant role in recommending treatment for a cohort of patients. Significant work has been done at the Healthcare Systems and Analytics Research department of IBM in the area of patient similarity involving physician feedback as an important parameter to group similar patients [21]. Chan et al.(2010) [4] have proposed a new patient similarity algorithm named SimSVM, which does a binary classification and outputs the predicted class which is survival greater than 12 months or less than 12 months and degree of similarity or dissimilarity. Their approach only considers a single outcome and a few similarity measures as input. We believe that our approach will be a significant improvement since we plan to consider multiple outcomes together as we believe that a single outcome based approach could be misleading. For each individual patient we would rank the treatment patterns based on historical experience with similar patients. Heuristics would be developed to do the ranking.

We categorize the patients based on some outcome variable as a range of values: e.g., survival period of 'less than 6 months', '6-12 months', '12-18 months', etc. The treatments for all the patients in each period would be represented as a graph. Sequential pattern mining techniques such as GSP (Generalized Sequential Pattern mining) [1] & SPADE (Sequential Pattern Discovery using Equivalence classes) [24] have been tailored to come up with "predominant" treatment patterns for every period. Thes patterns may consist of a combination of multiple drugs following a sequence and would be characteristic of a particular survival period representing the commonly used treatments associated with that period. These techniques are motivated by association rule mining techniques such as the Apriori algorithm. Initially a combination of N=2 drugs following a sequence are accounted for and the ones prescribed to a significant number of patients are considered for further analysis where 'N' is the number of drugs. This is followed by a combination of N+1 drugs and so on and so forth. The algorithm terminates when no more significant combinations can be formed. These patterns are used as features to classify patients based on survival periods.

The goal is to find patterns best suited for a particular class of patients sharing similar clinical and/or genomic characteristics. For a recently diagnosed patient we would determine the "nearest" classes based on "distance" of this patient from each predefined class of patients. Due to the difference in the extent of patient similarity, we would assign weights to the treatment patterns prescribed to patients belonging to a particular class. These weights would be dependent on the following parameters:

1. Distance between the test patient and the candidate patient class: The weight assigned to treatment patterns characteristic of a particular patient class is inversely proportional to the distance of the test patient from that class.
2. Number of candidate patients following a particular type of treatment pattern with respect to a particular class: The larger the number of patients following a particular treatment pattern, the higher the weight. A particular treatment pattern may have different weights for different patient classes.
3. Other criteria could include the degree of side effects, the extent of adverse reactions in patients and several other clinical considerations as determined by some consensus among physicians

4 Preliminary Results

Preliminary analysis on the GBM data was performed which consisted of approximately 300 deceased patients extracted from TCGA. We classify the data into two classes based on survival period which are i) patients surviving less than a year and ii) patients surviving more than a year. A linear classifier such as Logistic Regression was trained based on the clinical and the genomic features recorded before the treatment was started. In addition to these, significant sequential patterns prescribed within one year of their diagnosis were extracted from the treatment graph and used as features. For our study we have three domains of features which are 'Clinical' , 'Genomic' and 'Treatment'. We perform forward feature selection to pick significant features before running the classifier. In Table 1, we report c-statistic [13] and the accuracy of various predictive models with features from individual domains followed by models with a combination of the domains to analyze the predictive power of the same.

By analyzing the different combination of models in our experiment, we observe that among 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 as is evident from Table 1. Among the multiple domain models the best performance is achieved when

Table 1. Performance of various models in predicting patients surviving for >1 year

Single Domain Models	**c-statistic**	**Accuracy**
Genomic	0.76	78.1%
Clinical	0.71	72.2%
Treatment	0.69	71.2%
Multiple Domain Models		
Clinical + Genomic + Treatment	0.85	86.4%
Treatment + Genomic	0.84	84.8%
Clinical + Genomic	0.83	84.5%
Clinical + Treatment	0.78	78.6%

clinical, genomic and treatment features are analyzed together. We use 10 fold cross validation for evaluating our models and predictive features are selected for every fold. We rank these features based on the number of folds they are picked in (shown in Table 2). The treatment features shown in the table are in the form of treatment events, which consist of the drug/radiation type appended with the event number in the treatment. The arrow in the treatment patterns indicates the sequence in which the drugs were prescribed. The third column in the table shows if a feature positively or negatively influences survival.

5 Related Work

Some work has been done in the area of developing models for predicting treatment plans for patients. Research groups have developed models to predict the various drug interventions as well as drugs coupled with lab interventions that would work best for a particular disease. These models do not include important parameters like symptoms, results of investigations, laboratory test results, etc and are only limited to prediticting drugs that may be effective [15]. We consider a very comprehensive definition of a treatment plan and the approach outlined previously would rank the treatment patterns for a given patient, which implies selection of drugs/interventions, their dosages, their ordering, etc. Based on the models built by Kim. et al (2004) [7] for chronic heart failure (CHF) treatment, significant factors improving the plasma BNP levels were discovered, which were validated by large-scale trials. Similar work has been done in the area of heart disease diagnosis reporting fairly good accuracy [19]. Neuvirth et al (2011) [11] present a prototype for a data-driven risk assessment system for Diabetes patients and claim to identify physicians who can deliver optimal care to such patients and also identify patients requiring emergency care services. The decision support model developed by Chen et al(2012) for Diabetes [5] uses a case based reasoning approach to find patient cases similar to the one queried and is not very robust since the approach used by authors to find similar cases is not very granular and eventually the same line of treatment which is given to these similar cases is recommended for the new patient. In our approach we will take into consideration all the cohorts of patients similar to the test patient and

Table 2. Predictive Features from the Model: Clinical + Genomic + Treatment

Features	No. of folds	Influence on Survival >1 year
Genomic		
mRNA expression z-score for GABRA1 gene between -1.5 & -1	4	-
mRNA expression z-score for GABRA1 gene between -1 & 1	4	+
PTEN gene with R130* mutation	3	-
mRNA expression z-score for TP53 gene between -1.5 and -1	3	-
mRNA expression z-score for TP53 gene between 1 and 1.5	3	-
Clinical		
karnofsky performance score between 20 & 40	4	-
Age of patient >75 years	3	-
Treatment		
Dexamethasone.1 ->End of Treatment	7	-
External Beam Radiation Therapy.2 ->End of Treatment	5	-
CCNU.2 ->End of Treatment	4	-
Temodar.1 ->Avastin.2	3	+
Temodar.2 ->CCNU.3	3	+
Procarbazine.2 ->End of Treatment	3	+

then assign weights to the treatment pattern in each cohort, which we believe would be more accurate than the case based reasoning approach. We extensively tease out the different treatment patterns that are characteristic of a particular outcome and plan to come up with a meaningful measure of patient similarity to build patient profiles based on clinical and possibly behavioral variables, especially for diabetes. We believe our approach to assign weights to different treatment patterns for a particular patient profile is also very comprehensive and unique especially because we will be relying on consensus from domain experts for individual diseases.

6 Expected Outcomes

Most of the work mentioned below will feed into the prototype development work. Actual testing will be done for the validation of algorithms with available data. Overall testing of our treatment suggestion system will be done in conjunction with the prototype implementation.

Decision Support Framework: Our long-term goal is to develop a decision support framework primarily focusing on analyzing sequential patterns in the treatment prescribed and making personalized recommendations for different patients. This would involve the following:

1. **Enhancing the graph:** The current graph structure [9] is restricted to nodes representing therapies. Treatment of chronic diseases such as Diabetes involves management of social behavior, physical activities, etc in addition to medication. We would also need to represent periodic monitoring of the patient and patient preferences to come up with an optimal treatment. New technologies based on parallel processing like Hadoop and graph databases such as Neo4j [12] may be possible choices to use.

2. **Algorithms:** A variety of algorithms will be designed including sequential pattern mining algorithms to analyze treatment data with multiple parameters, clustering algorithms to cluster patients into patient cohorts , algorithms to define a patient clinically in terms of patient similarity or to define the distance of a patient from a patient class. In addition to these, weighting and ranking algorithms for treatments would also be designed.

7 Contribution of Dissertation

The dissertation is focused on developing a decision support framework for treatment of rare and chronic diseases and finding treatment patterns which can influence a particular outcome or multiple outcomes. In the paper we presented the first step in mining significant sequential treatment patterns and correlating the patterns with survival period of patients. From a clinical perspective, it gives an insight to the medical practitioner about the significance of prescribing a set of drugs in a particular sequence. The contribution of the thesis would be the following :-

1. The treatment selection problem would be modeled using a graph formalism.
2. Enhancements would be made to sequential mining algorithms to incorporate parameters pertinent to the disease under consideration.
3. A spophisticated predictive decision support model would be developed for the healthcare domain to make personalized treatment recommendations which would have the potential to be enhanced for other domains.

References

1. Agrawal R. et al. “Mining Sequential Patterns”. ICDE 1995: 3-14
2. Centers for Disease Control and Prevention. National diabetes fact sheet: national estimates and general information on diabetes and prediabetes in the United States, 2011. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2011
3. Cerami et al. ‘The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. Cancer Discovery’. May 2012 2; 401.

4. Chan L.W.C et al (2010). “Machine Learning of Patient Similarity: A Case Study on Predicting Survival in Cancer Patient after Locoregional Chemotherapy” 2010 IEEE International Conference on Bioinformatics and Biomedicine Workshops.
5. Chen J.X et al (2012). “Diabetes Care Decision Support System”. 2nd International Conference on Industrial & Information Systems.
6. Hegi et al (2005).“MGMT Gene Silencing and Benefit from Temozolomide in Glioblastoma.” The New England Journal Of Medicine 352(10): 997-1003.
7. Kim.J et al (2004).“A Novel Data Mining Approach to the Identification of Effective Drugs or Combinations for Targeted EndpointsApplication to Chronic Heart Failure as a New Form of Evidence-based Medicine.” Cardiovascular Drugs and Therapy 18(6): 483-489.
8. Kon H.C et al (2005). “Data mining applications in healthcare.” Journal of Healthcare Information Management 19(2): 64-72.
9. Mooney C.H et al. “Sequential Pattern Mining - Approaches and Algorithms”. ACM Computing Surverys, Vol.45, No.2, Article 19, Feb 2013. Sequential Pattern Mining Survey paper Carl H Mooney and John F. Roddick.
10. Network, T. R. (2010). The Cancer Genome Atlas Data Portal, National Institute of Health.
11. Neuvirth et. al (2011).“Toward personalized care management of patients at risk: the diabetes case study”. KDD: 395-403.
12. Neo4j.org, (2014). Neo4j - The World’s Leading Graph Database. [online] Available at: http://www.neo4j.org/.
13. Park S.H et al (2004). “Receiver Operating Characreristic (ROC) Curve: Practical Review for Radiologists.” Korean Journal of Radiology 5(1): 11-18.
14. Parsons et. al. “An Integrated Genomic Analysis of Human Glioblastoma Multiforme.” Science 321(5897): 1807-1812, 2008.
15. Razali A.M et al (2009). “Generating Treatment Plan in Medicine : A Data Mining Approach” American Journal Of Applied Sciences 6(2): 345-351.
16. Robinson.I et al (2013). Graph Databases. California, O’Reilly Media Inc.
17. Sackett D.L et al.(1996). “Evidence based medicine: what is it and what is isnt.” BMJ 312(7023): 71-72.
18. Saaddine J.B et al ,“Improvements in diabetes processes of care and intermediate outcomes: United States”, 1988-2002.[summary for patients in Ann Intern Med. 2006 Apr 4;144(7):I12; PMID: 16585656]. Annals of Internal Medicine.2006;144(7):465-74.
19. Shouman M et al (2012). ”Using data mining techniques in heart disease diagnosis and treatment” Japan-Egypt Conference on Electronics, Communications and Computers
20. Skevofilakas M.T et al (2005).“A decision support system for breast cancer treatment based on data mining technologies and clinical practice guidelines.” Conference Proceedings IEEE Engineering in Medicine and Biology, Shanghai China.
21. Sun.J et al (2012). “Supervised Patient Similarity Measure of Heterogenous Patient Records.” SIGKDD Explorations 14(1): 16-24.
22. Verhaak R.G et al (2010). “Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1.” Cancer Cell 17(1): 98-110.
23. Wang.F et al (2011). “iMet: Interactive Metric Learning in Healthcare Applications”. SDM: 944-955.
24. Zaki.M.J. “SPADE: An Efficient Algorithm for Mining Frequent Sequences”. Machine Learning 42(1/2): 31-60 (2001)