Journal of Biomedical Informatics 61 (2016) 267–275

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Constraint based temporal event sequence mining for Glioblastoma survival prediction

Kunal Malhotra [a], Shamkant B. Navathe [a], Duen Horng Chau [a], Costas Hadjipanayis [b,c], Jimeng Sun [a,*]

[a] College of Computing, Georgia Institute of Technology, Atlanta, GA, USA
[b] Department of Neurosurgery, Emory University School of Medicine, Atlanta, GA, USA
[c] Winship Cancer Institute of Emory University, Atlanta, GA, USA

ARTICLE INFO

Article history:
Received 3 November 2015
Revised 5 March 2016
Accepted 25 March 2016
Available online 5 April 2016

Keywords:
Graph mining
Predictive model
Sequential pattern mining
Classification
Treatment patterns
Glioblastoma

ABSTRACT

Objective: A significant challenge in treating rare forms of cancer such as Glioblastoma (GBM) is to find optimal personalized treatment plans for patients. The goals of our study is to predict which patients survive longer than the median survival time for GBM based on clinical and genomic factors, and to assess the predictive power of treatment patterns.
Method: We developed a predictive model based on the clinical and genomic data from approximately 300 newly diagnosed GBM patients for a period of 2 years. We proposed sequential mining algorithms with novel clinical constraints, namely, 'exact-order' and 'temporal overlap' constraints, to extract treatment patterns as features used in predictive modeling. With diverse features from clinical, genomic information and treatment patterns, we applied both logistic regression model and Cox regression to model patient survival outcome.
Results: The most predictive features influencing the survival period of GBM patients included mRNA expression levels of certain genes, some clinical characteristics such as age, Karnofsky performance score, and therapeutic agents prescribed in treatment patterns. Our models achieved *c*-statistic of 0.85 for logistic regression and 0.84 for Cox regression.
Conclusions: We demonstrated the importance of diverse sources of features in predicting GBM patient survival outcome. The predictive model presented in this study is a preliminary step in a long-term plan of developing personalized treatment plans for GBM patients that can later be extended to other types of cancers.

1. Introduction

Glioblastoma (GBM) is the most lethal and biologically the most aggressive brain cancer with patients having a median survival of 12–15 months [10,29]. Understanding what factors prolong survival and promote treatment responses can be of value to patients and physicians. The Cancer Genome Atlas (TCGA) [17], a project of the National Institutes of Health (NIH), classified Glioblastoma patients into four distinct molecular subtypes affecting biological behaviors, suggesting that no single therapeutic regimen can be equally effective for all subtypes [6]. 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 correlating to 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 [20]. Krex et al. [12] and Walid [33] 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. Predictive models have been developed in the past utilizing imaging and clinical features of patients [14] and there also exists ongoing clinical trials on certain drugs to test their effect on survival [34] 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 the first line standard of care treatment, there are different treatment combinations

* Corresponding author at: College of Computing, Georgia Institute of Technology, 266 Ferst Dr, Atlanta, GA 30332, USA.
E-mail address: jsun@cc.gatech.edu (J. Sun).

http://dx.doi.org/10.1016/j.jbi.2016.03.020
1532-0464/

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 will provide insight into treatment patterns, which may be associated with greater overall patient survival. Based on our knowledge, there is no existing literature that analyzes treatment patterns that may influence survival for new GBM patients. The proposed approach is general and can be used for other clinical settings.

1.1. Contributions

Our study makes the following contributions:

1. We introduce a novel graph approach to extend existing sequential pattern mining algorithms for a clinical predictive modeling application.
2. We extended existing sequential pattern mining algorithms by incorporating two additional constraints called the 'exact-order' and 'overlap', which can generate more clinically meaningful treatment patterns.
3. We followed 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.

2. Related work

2.1. Influence of genomic factors on GBM

High dimensional gene expression profiling studies in GBM patients have identified gene signatures associated with epidermal growth factor receptor (EGFR) overexpression and survival [5,13,15,16,19,22,25–27]. Genomic abnormalities associated with TP53 and RB1 mutations have been identified in TCGA along with GBM-associated mutations in genes such as PIK3R1, NF1, and 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 this is a core requirement for GBM pathogenesis [28]. However, no one systematically tests those genomic factors together with clinical and treatment information for predicting GBM survival outcome, which is a focus of this paper.

2.2. Sequential pattern mining

Sequential pattern mining refers to the mining of frequently occurring ordered events or subsequences as patterns [11]. This technique, introduced by Agarwal and Srikant [1] in their 1995 study of customer purchase sequences, led to the development of the Generalized Sequential Pattern mining (GSP) algorithm which is based on the Apriori [35] algorithm to mine frequent itemsets. GSP uses the downward-closure property of sequential patterns and adopts a multiple-pass, candidate generation approach. Initially it finds all the frequent sequences of length one item with minimum support. Subsequently it combines every possible 1-item itemset which has the minimum support for the next pass. Besides GSP, another popular sequential mining algorithm is SPADE (Sequential PAttern Discovery using Equivalent classes) [30] which uses a vertical id-list database format data format and associates each sequence a list of transactions in which it occurs. The frequent sequences can be found by efficiently using intersection on id-lists. Bellazi et al. [31] 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 [32]. This is a fast time-interval mining method, which exploits the transitivity inherent in temporal relations. The other sequential pattern mining algorithms are based on the 'Pattern Growth' technique of frequent patterns avoiding the need for candidate generation unlike GSP and SPADE which are based on Apriori. This approach involves finding frequent single items, and condensing this information into a frequent pattern tree. PrefixSpan [8,21] is one such algorithm which exploits this approach by building prefix patterns and concatenating them with suffix patterns and concatenating them with suffix patterns to find frequent patterns. SPAM (Sequential PAttern Mining using a bitmap representation) [2] uses a depth-first traversal of the search space with various pruning mechanisms and a vertical bitmap representation of the database enabling efficient support counting. Our approach is very minimally inspired by Apriori 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 properly apply treatment pattern mining, we introduce several important constraints such as 'exact-order' and 'overlap'.

3. Approach

3.1. Data

We constructed a rich dataset of newly diagnosed GBM patients by integrating two different databases called the *TCGA* [17] and the *cBioPortal* [4,7]. TCGA consists of clinical and treatment data pooled together from different research teams, which is publicly accessible. The genomic data for the same patients was obtained from cBioPortal, a web resource of multidimensional cancer genomics data maintained by the Memorial Sloan Kettering Cancer Center.

3.1.1. Features

For our study, we analyzed data from 309 newly diagnosed GBM patients spanning over a period of 2 years from the date of diagnosis. The data was categorized into 'Clinical', 'Genomic' and 'Treatment' domains. The clinical domain includes demographic information about the patient along with 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' [28]. 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 the MGMT gene was also used for our analysis [9]. The treatment domain consists of treatment plans for each patient, which can be viewed as process data. We use sequential mining algorithms to mine significant patterns in their treatment plans and use them as features in the dataset in addition to clinical and genomic features. Table 1 summarizes the dimensions of the dataset categorized by the domain.

3.1.2. 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

Table 1
Summary of the dataset.

Feature statistics	Number of features
Clinical domain	11
Genomic domain	33
Treatment domain	49
	Total: 93
Data statistics	
Number of patients	309
Patients surviving more than a year	140
Patients surviving for less than a year	169
Race	White (243), Black (42), Asian (24)
Gender	Male (229), Females (80)

for the study consists of living patients who have already survived for more than a year in addition to the deceased patients that constitute the majority of patients.

3.2. Methodology

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

3.2.1. Predictive modeling pipeline

The predictive modeling pipeline consists of 4 modules, namely, 'Data Standardization and Cleaning', 'Sequential Pattern Mining', 'Feature Construction' and 'Prediction and Evaluation'. As shown in Fig. 1, the raw data is fed into the 'Data Standardization and Cleaning' module to filter out noisy data. The 'Sequential Pattern Mining' module extracts significant medication patterns from the treatment data including the standard of treatment. The clinical and genomic features are combined with these medication patterns to form a binary feature matrix in the 'Feature Construction' module, with each row corresponding to 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.

3.2.2. 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. E.g. 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.

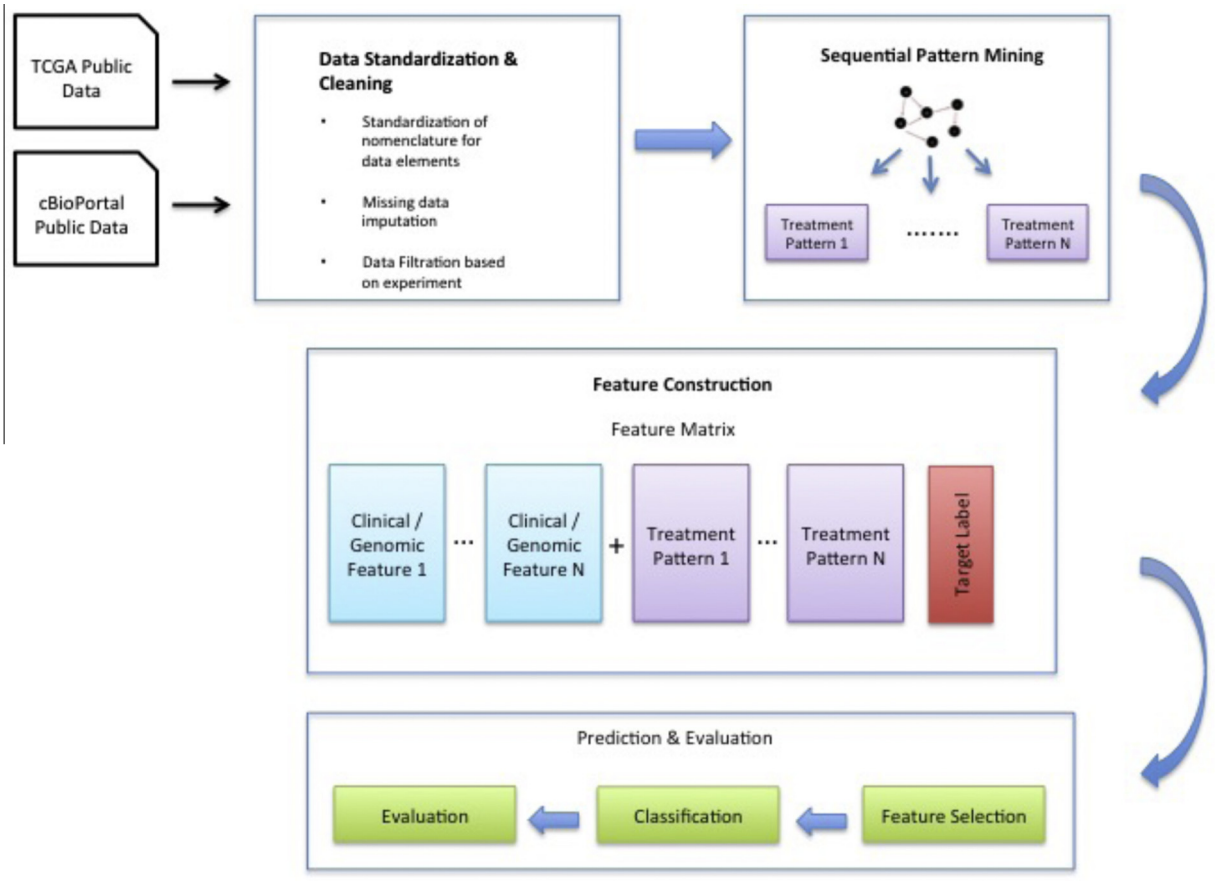

Fig. 1. Predictive modeling pipeline.

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.

3.2.3. Sequential pattern mining

The data used for this study is modeled as a graph consisting of nodes, categorized as 'patient node' and 'treatment type node', and edges categorized as, 'prescription edge' and 'sequence edge'. A graph offers a much richer model of the underlying data, and allows relationships of several types. The graph we constructed provides a good way to model event sequences and their temporal relationships [24]. For illustrative purposes, Fig. 2 shows the current representation of the data as a graph consisting of two patients. The patient nodes have properties such as 'patient id' and 'age at diagnosis'. Prescribed drugs and radiotherapy are represented as treatment type nodes with properties 'drug name' and 'radiation type' respectively. The undirected 'prescription edge' signifies the prescription of a treatment with properties corresponding to the prescription. The 'sequence edge' is a directed edge signifying the sequence in which drugs or radiation were prescribed. For example, 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' between 05/21/2007 and 06/22/2007. The edge labeled 'Followed_by' would always be between a radiation type and a drug, two drugs or two radiation types, signifying the sequence of the prescription. For example, 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_B followed Drug_A with an overlap of 24 days.

Sequential pattern mining was used to extract patterns from the treatment data to give two types of information for every patient: *the sequence of drugs/radiation prescribed* and their *time of prescription within the sequence*.

A treatment plan for a patient may consist of a combination of multiple drugs or radiation or both prescribed in a particular sequence. We define a treatment for one patient as a sequence of events, each event consisting of administration of a treatment type (drug or radiation). To mine such treatment plans, we tailor existing approaches such as GSP [1] and SPADE [30] by adding two new constraints, namely, 'exact-order' and 'overlap' constraints (explained in the following sections). 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, Drug_A → Drug_B → Radiation_D is a 2-path event set from the graph model shown in Fig. 2, consisting of two drugs and a radiation therapy forming a sequence of consecutive events (represented as nodes), which is not a hard restriction in any of the existing algorithms. Since a treatment plan for a patient may consist of a drug being prescribed more than once, an event identifier is added along with each treatment type node to satisfy the 'exact-order' constraint. The current implementation does not consider dosage of the drug or radiation therapy due to missing dosage

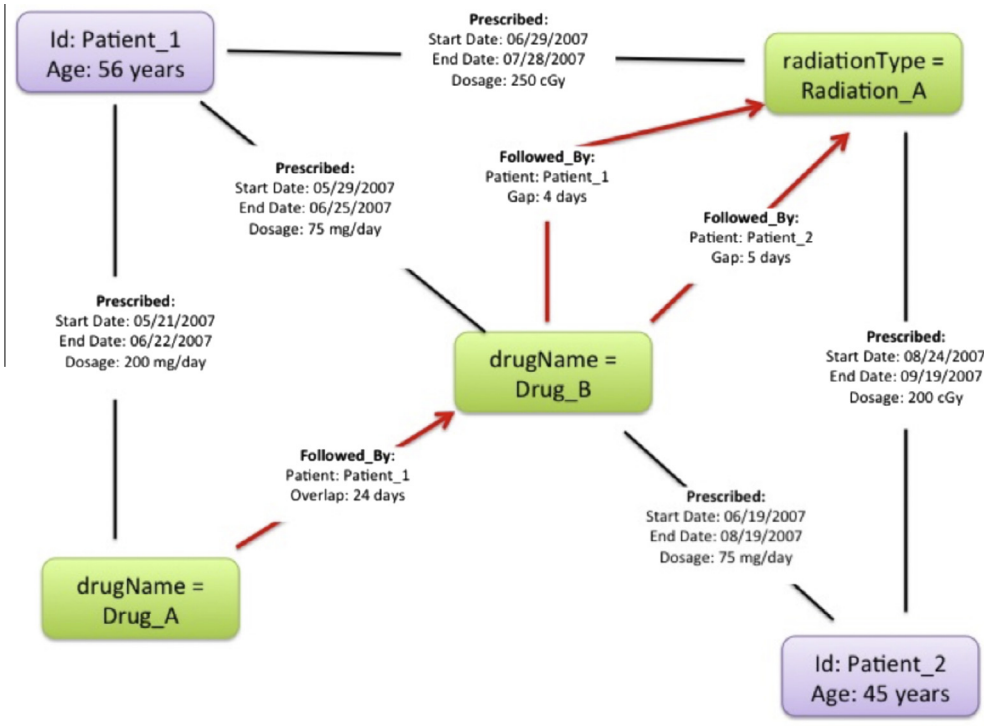

Fig. 2. Data represented as a treatment graph.

units for many therapies. The pseudo-code in Algorithm 1 provides the details of the proposed approach.

Algorithm 1 MineTreatmentPatterns

```
Input: Graph of events, minSup
Output: Frequent Sequences
1: N ← Length of path
2: minSup ← minimum support
3: N = 1
4: REPEAT Steps 5 to 12 UNTILL size of N + 1th path
   pool = 0;
5:  S ← Set of N-path sequences of treatment events
   with support ⩾ minSup
6:  for all sequence s ∈ S do
7:    for all sequence s' ∈ S - s do
8:      A ← first N treatment events of s'
9:      B ← last N treatment events of s
10:       if A = B && frequency of N + 1th
   sequence > minSup then
11:         add N + 1th sequence to N+1 path pool
12:  Increment N
```

3.2.3.1. Exact-order constraint. This constraint forces only those events to be a part of sequence, which occur consecutively. If there are multiple other drugs given between 2 drugs 'A' and 'B', then a sequence ⟨A B⟩ does not occur. Such a constraint holds a lot of significance in the treatment domain since even though the sequence ⟨A B⟩ with other drugs in between can occur frequently, clinically it may not hold any relevance since the intermediary drugs may or may not affect the outcome. This particular constraint was not implemented in traditional sequential mining algorithms. To implement this constraint, we annotate the treatment type nodes with event identifiers signifying the time of occurrence of that event.

3.2.3.2. Candidate generation. We extract path sequences of length '*N*' from the treatment graph and consider the ones prescribed to a significant number of patients for further analysis. This is followed by forming '$N+1$' path sequences, by increasing the path-length one edge at a time, and joining on the event IDs and the treatment type nodes. Fig. 3 illustrates the candidate generation step, each node representing a treatment type or a combination of treatment types. Initially a 1-path set consisting of combinations of two consecutive treatment types is extracted from the treatment graph such that the sequences should have been prescribed to a significant number of patients i.e. the sequences have a support value greater than a pre-specified threshold. From these 1-path sets, we form 2-path combinations by joining on the treatment type node and the event ID. The combinations bracketed in green have the potential to be joined since the resulting sequence has consecutive events. The ones bracketed in red are not joined since event 'A' as the fourth event is different from event 'A' as the second event. This process continues till we cannot form new combinations or they are insignificant.

3.2.3.3. Overlap constraint. A 'treatment plan' for a patient consists of all the treatment types prescribed to a patient in a sequence. When a treatment is in effect and another one starts concurrently, an overlap of treatments occurs. Two common situations are 'partial overlap' and 'total overlap' of prescriptions. We say '*n*' prescriptions ($n > 1$) have a partial overlap if all the '*n*' prescriptions are concurrently prescribed to a patient on at least one day. A total overlap is a special case of partial overlap that occurs when all the '*n*' prescriptions have the same start and end dates. Approaches called the "Single Node" and "Combination Node" approach are formulated for handling overlap constraints. In the 'Single Node' approach, a sequence considers each treatment type as a single node as shown in Fig. 4(a). If there is a partial overlap between two prescribed treatment types, the prescription that ends first becomes the source node and a directed edge connects it to the other prescription. For example, a directed edge connects 'Dexamethasone' to 'Radiation.' In the case of a total overlap between two prescriptions, both prescriptions are treated as source nodes, with directed edges leading to the next treatment

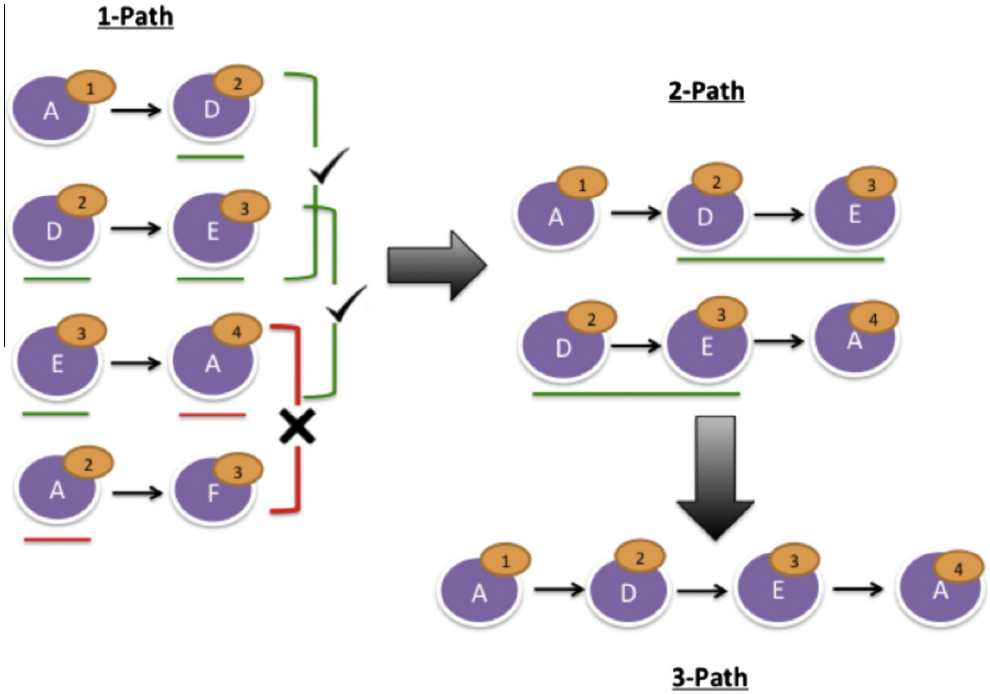

Fig. 3. Illustration of candidate generation.

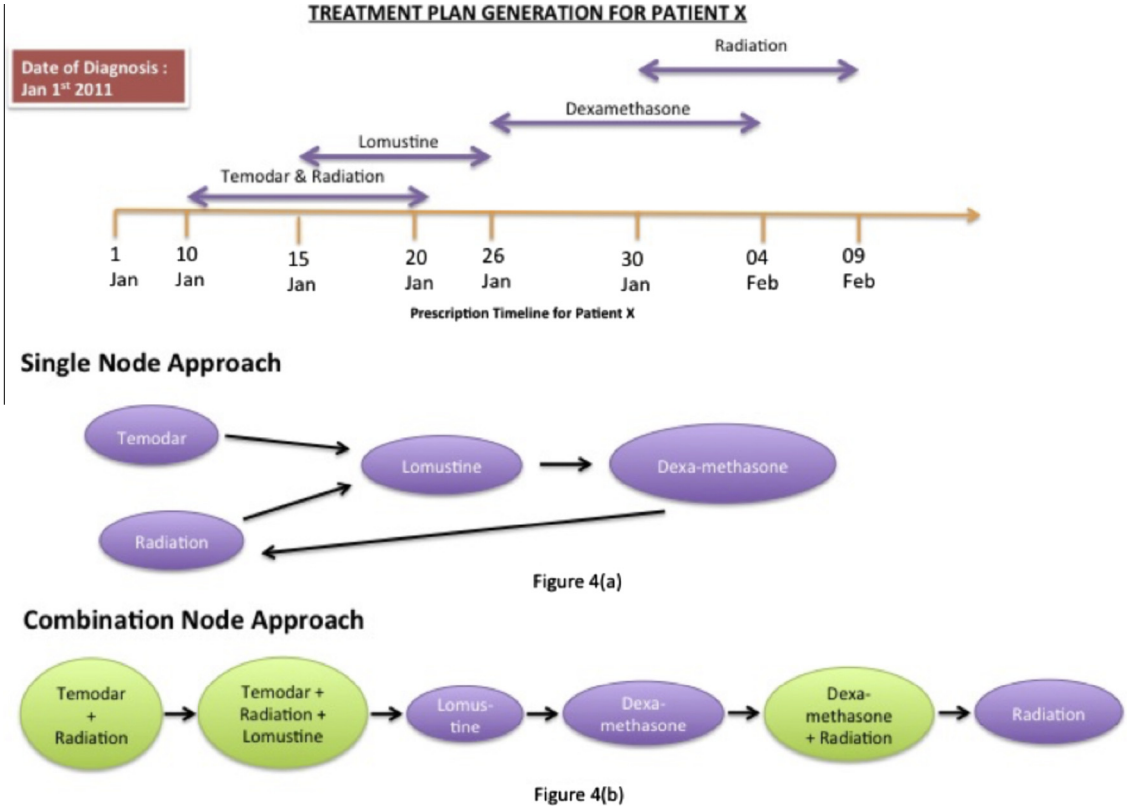

Fig. 4. Approaches for treatment plan generation. (a) Single node approach. (b) Combination node approach.

type node. For example, there are directed edges from 'Temodar' and 'Radiation' toward Lomustine.

The 'overlap' constraint refers to an overlap between multiple drug prescriptions and results in combining those drugs into a single node and treating it as a single event.

Under the approach called the 'Combination Node' approach shown in Fig. 4(b) a new node is created whether there is a partial or a total overlap between treatment type prescriptions. The significance of such a combination node approach is to retain the potential information hidden in the fact that concomitant treatment may be clinically more relevant in some situations that mono therapies. The timeline shown in the figure signifies the order of prescriptions. The purple colored nodes represent individual drugs and the green nodes are created to signify overlapping prescriptions. The combination node approach should not be used in cases where there exists a lot of variety in the combinations being created since it will result in lesser support value. For such data sets the single node approach works better since it functions only on the existing nodes. The current approach is based on data about GBM patients and would be enhanced for other diseases having extensive treatment guidelines and possibly incorporating potential complications.

3.2.4. 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. We also use time to event as the alternative outcome for Cox Regression.

3.2.5. 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. We also developed a Cox Regression model using the time lapsed between the first visit of the patient and the date of death as the survival time. The same feature selection process is applied for both Logistic Regression and Cox Regression.

4. Results

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

Table 2
Performance of various models in predicting patients surviving for >1 year using Logistic Regression (LR) and Cox Regression (CR).

	Single node approach						Combination node approach					
	C-statistic		Accuracy (%)		Precision	Recall	*C*-statistic		Accuracy (%)		Precision	Recall
	LR	CR	LR	CR			LR	CR	LR	CR		
Individual domain models												
Genomic	0.76	0.75	78.1	78.0	0.72	0.74	0.76	0.75	78.1	78.0	0.72	0.74
Clinical	0.71	0.71	72.2	72.3	0.70	0.75	0.71	0.71	72.2	72.3	0.70	0.75
Treatment	0.69	0.70	71.2	72.0	0.67	0.66	0.60	0.61	63.3	63.1	0.68	0.66
Multiple domain models												
Clinical + genomic + treatment	0.85	0.84	86.4	86.7	0.75	0.74	0.85	0.84	86.2	87.0	0.74	0.76
Treatment + genomic	0.84	0.86	84.8	84.3	0.69	0.67	0.78	0.78	81.0	81.1	0.68	0.69
Clinical + genomic	0.83	0.83	84.5	84.7	0.73	0.74	0.83	0.83	84.5	84.7	0.73	0.74
Clinical + treatment	0.78	0.79	78.6	77.8	0.67	0.68	0.75	0.76	74.5	74.3	0.66	0.70

Table 3
Predictive clinical and genomic features from the model: Clinical + genomic + treatment. The sign (+/−) shown in the table signifies positive or negative influence on the outcome.

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

Table 4
Predictive treatment patterns. The sign (+/−) shown in the table signifies positive or negative influence on the outcome.

	Predictive treatment patterns	Percentage of times selected (%)	Influence on survival >1 year	*P*-value
Single node approach	Radiation Therapy{2} → Treatment Termination	50	−	0.0061
	Lomustine{2} → Treatment Termination	40	−	0.05
	Procarbazine{2} → Treatment Termination	30	+	0.05
	Temozolomide{2} → Lomustine{3}	40	+	0.04
Combination node approach	Temozolomide{1} → [Temozolomide + Radiation]{2}	30	−	0.05
	[Temozolomide + Radiation]{1} → Temodar{2} → Lomustine{3}	30	+	0.04

4.1. Quantitative analysis

In Table 2, we report the performance of various models in which both 'single node' and 'combination node' approaches were used to construct treatment patterns which were used with different mixes of clinical and genomic features. *C*-statistic and accuracy from both logistic regression and cox regression methods have been reported. 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 (see Table 2). Among the multiple domain models, the best performance is achieved when features from all three domains are analyzed together. The average precision and sensitivity of the Logistic Regression and Cox Regression methods are also reported in the table for the individual and multiple domain models. Table 3 shows the predictive clinical and genomic features which were selected in at least 30% of the folds generated during the cross validation step along with the influence they have in prolonging overall survival beyond 1 year. The predictive treatment patterns shown in Table 4 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. E.g. 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.

4.2. Qualitative analysis

Besides accurate prediction results, the predictive features are also clinically meaningful. Methylation of the MGMT gene has been reported to be crucial for some of the standard of care chemotherapeutics such as Temozolomoide (Temodar) to be effective which in turn prolongs survival [9,23]. 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 (see supplement for score description) [18]. 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.

5. Conclusion

In this paper, 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). Novel sequential mining approach is proposed 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 are developing a treatment advisor tool to recommend treatments for a patient based on treatments given to patients having a similar clinical and genomic profile using the knowledge of treatment patterns obtained from this study. 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.

Conflict of interest

The authors declared that there is no conflict of interest.

Acknowledgment

This work was supported by the National Science Foundation, award IIS-1418511 and CCF-1533768, Children's Healthcare of Atlanta, CDC I-SMILE project, Google Faculty Award, AWS Research Award, ORNL Go! program and UCB.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2016.03.020.

References

[1] R. Agarwal, R. Srikant, Mining sequential patterns, in: Proc. International Conference on Data Engineering, ICDE, IEEE Computer Society, 1995, pp. 3–14.
[2] J. Ayres, J. Flannick, J. Gehrke, T. Yiu, Sequential pattern mining using a bitmap representation, in: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, 2002, pp. 429–435.
[4] Cerami et al., The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data, Cancer Discov. 2 (2012) 401.
[5] Nelson et al., Gene expression profiling of gliomas strongly predicts survival, Cancer Res. 64 (2004) 6503–6510.
[6] Glioblastoma and Malignant Astrocytoma: American Brain Tumor Association, 2012. Available at: <http://www.abta.org/secure/glioblastoma-brochure.pdf>.
[7] Gao et al., Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal, Sci. Signal 6 (269) (2013) pl1.
[8] J. Han, H. Pei, Y. Yin, Mining frequent patterns without candidate generation, in: Proc. ACM SIGMOD Int. Conf. on the Management of Data (SIGMOD'00, Dallas, TX), ACM Press, New York, USA, 2000.
[9] Hegi et al., MGMT gene silencing and benefit from Temozolomide in Glioblastoma, New Engl. J. Med. 352 (2005) 997–1003.
[10] E. Holland, Glioblastoma multiforme: the terminator, PNAS 97 (12) (2000) 6242–6244.
[11] M. Kamber, J. Han, Data Mining: Concepts and Techniques, second ed., Elsevier, 2006.
[12] Krex et al., Long-term survival with glioblastoma multiforme, Brain 130 (10) (2007) 2596–2606, http://dx.doi.org/10.1093/brain/awm204.
[13] Israel et al., Gene expression profiling reveals molecularly and clinically distinct subtypes of glioblastoma multiforme, Proc. Natl. Acad. Sci. USA 102 (2005) 5814–5819.
[14] M.A. Mazurowski, A. Desjardins, J.M. Malof, Imaging descriptors improve the predictive power of survival models for glioblastoma patients, Neuro-Oncology 15 (10) (2013) 1389–1394.
[15] Liau et al., Identification of molecular subtypes of glioblastoma by gene expression profiling, Oncogene 22 (2003) 2361–2373.
[16] Kouwenhoven et al., Stem cell related "self-renewal" signature and high epidermal growth factor receptor expression associated with resistance to concomitant chemoradiotherapy in glioblastoma, J. Clin. Oncol. 26 (2008) 3015–3024.
[17] T.R. Network, The Cancer Genome Atlas Data Portal, National Institute of Health, 2010.
[18] Nice.org.uk, 'Guidance on the use of temozolomide for the treatment of recurrent malignant glioma (brain cancer)'|appendix-d-karnofsky-performance-score|Guidance and guidelines|NICE, 2001, Available at: <www.nice.org.uk/guidance/ta23/chapter/appendix-d-karnofsky_performance-score> performance-score (accessed 9 Nov 2014).
[19] Batchelor et al., Gene expression-based classification of malignant gliomas correlates better with survival than histological classification, Cancer Res. 63 (2003) 1602–1607.
[20] Parsons et al., An integrated genomic analysis of human glioblastoma multiforme, Science 321 (5897) (2008) 1807–1812.
[21] Pei et al., PrefixSpan: mining sequential patterns by prefix-projected growth, in: Proc. Int. Conference on Data Engineering, (ICDE), IEEE Computer Society, 2001, pp. 215–224.

[22] Soroceanu et al., Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis, Cancer Cell 9 (2006) 157–173.
[23] Rivera et al., MGMT promoter methylation is predictive of response to radiotherapy and prognostic in the absence of adjuvant alkylating chemotherapy for glioblastoma, Neuro-oncology 12 (2) (2010) 116–121.
[24] I. Robinson, J. Webber, E. Eifrem, Graph Databases, O'Reilly Media Inc., California, 2013.
[25] Ruano et al., Identification of novel candidate target genes in amplicons of glioblastoma multiforme tumors detected by expression and CGH microarray profiling, Mol. Cancer 5 (2006) 39.
[26] Nelson et al., Gene expression profiling identifies molecular subtypes of gliomas, Oncogene 22 (2003) 4918–4923.
[27] Mischel et al., Distinct transcription profiles of primary and secondary glioblastoma subgroups, Cancer Res. 66 (2006) 159–167.
[28] R.G. Verhaak et al., Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1, Cancer Cell 17 (1) (2010) 98–110.
[29] P.Y. Wen, S. Kesari, Malignant gliomas in adults, New Engl. J. Med. 359 (5) (2008) 492–507.
[30] M.J. Zaki, SPADE: an efficient algorithm for mining frequent sequences, Machine Learning 42 (1/2) (2001) 31–60.
[31] R. Bellazi, P. Fratino, C. Cerra, L. Sacchi, S. Concaro, Mining healthcare data with temporal association rules: improvements and assessment for a practical use, in: Proceedings of the 12th Conference on Artificial Intelligence in Medicine: Artificial in Medicine, 2009.
[32] Y. Shahar, R. Moskovitch, Medical temporal-knowledge discovery via temporal abstraction, in: AMIA Annual Symposium Proceedings, 2009, pp. 452–456.
[33] M.S. Walid, Prognostic factors for long-term survival after glioblastoma, Permanente J. 12 (4) (2008).
[34] Study Finds Glioblastoma Patients Treated with Bevacizumab Experience Reduced Cognitive Function and Quality of Life. MD Anderson News Release 06/01/13. <http://www.mdanderson.org/newsroom/news-releases/2013/patients-treated-with-bevacizumab.html>.
[35] R. Agarwal, R. Srikant, Fast Algorithms for mining association rules in large databases, VLDB (1994) 487–499.

Further reading

[3] R. Beroukhim et al., Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma, Proc. Natl. Acad. Sci. USA 104 (2007) 20007–20012.