

# Temporal Event Sequence Mining for Glioblastoma Survival Prediction

Kunal Malhotra  
Georgia Institute of  
Technology  
266 Ferst Drive  
Atlanta, GA, USA  
kmalhotra7@gatech.edu

Duen Horng Chau  
Georgia Institute of  
Technology  
266 Ferst Drive  
Atlanta, GA, USA  
polo@gatech.edu

Jimeng Sun  
Georgia Institute of  
Technology  
266 Ferst Drive  
Atlanta, GA, USA  
jsun@cc.gatech.edu

Costas Hadjipanayis  
Emory University School of  
Medicine The Emory Clinic,  
Building B, 2nd floor 1365-B  
Clifton Rd., NE, Ste. 2200  
Atlanta, GA 30322  
chadjip@emory.edu

Shamkant B. Navathe  
Georgia Institute of  
Technology  
266 Ferst Drive  
Atlanta, GA, USA  
sham@cc.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, 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. We use publicly available data for 300 patients spanning a period of 2 years from The Cancer Genome Atlas Portal, which has actual de-identified patient data from multiple institutions. Information about each patient comprises a set of features from the clinical and the genomic domain. We also use sequential mining algorithms to extract treatment patterns and use the patterns themselves as additional features. A model predicting whether a patient would survive for more than a year is developed using logistic regression and the most predictive features influencing the survival period of GBM patients include mRNA expression levels of certain genes and medications given in a particular sequence. The model achieved an AUC of 0.85 with an accuracy of 86.4%. The study is a preliminary step in a long term plan of developing personalized treatment plans with GBM patients as an initial model that can later be extended to other diseases.

## Categories and Subject Descriptors

J.3 [Life and Medical Sciences]: Health; H.2.8 [Database Applications]: Data Mining

## General Terms

Algorithm, Experimentation

## Keywords

Predictive Modeling, Sequential Pattern Mining

## 1. INTRODUCTION

Glioblastoma Multiforme (GBM) is the most lethal type of brain cancer and is biologically the most aggressive subtype of malignant gliomas. From a clinical perspective, gliomas are divided into four grades and the most aggressive of these is GBM or grade IV glioma and is most common in humans. Histologically GBM may be distinguished from other grades of glioma due to the presence of necrosis (dead cells) and increase of blood vessels around the tumor. Gene expression profiling has given rise to new molecular classification schemes. This classification by gene expression profiling has also revealed molecular classes not detected by traditional methods of looking at tumor samples under a microscope. Previous studies have shown that EGFR (epidermal growth factor receptor) expression or overexpression was observed in GBM. IDH1 (Isocitrate dehydrogenase) and some IDH2 found in certain subsets of glioblastoma are being used as a diagnostic test for predicting longer survival and for evaluating the efficacy of new targeted molecular drugs. The Cancer Genome Atlas (TCGA) [22], a project of the National Institutes of Health (NIH), led to work done on classifying glioblastoma into four distinct molecular types which may lead to different treatment regimens [6].

The current standard of care for GBM patients involves surgical resection followed by radiation therapy and chemotherapy with the oral alkylating agent Temodar [18]. Most patients with GBMs survive for less than a year. This extreme mortality rate, where none have a long-term survival, has

drawn significant attention to improving treatment for these tumors. After the first line of treatment there are a pool of treatment options for a neurologist to decide from for an individual. The sequence in which the next set of drugs or therapy needs to be prescribed adds to the level of complexity since multiple drugs given in a particular sequence may work better than given in some other order. The reason GBMs are not cured by surgery is due to the topographically diffuse nature of the disease. In addition to the fact that the nature of the tumor is very complex, the location of the tumor cells within the brain is variable, resulting in the inability to completely resect this tumor [7].

GBM is a rare form of brain tumor having a median survival period of approximately 12 months, however a small percentage of patients survive for longer period of times. Clinicians have been interested in discovering the factors, which prolong survival periods. Krex et al [10] have analyzed such patients and discovered certain clinical and molecular features which play a significant role in prolonging the survival period. Predictive survival models have been developed in the past using imaging features of the MR scans of patients along with a few clinical features. The performance of the model based on both clinical and imaging features was observed to be higher than the performance of the model based on only the clinical features [12]. Based on our knowledge, there is no existing work which analyzes treatment prescribed to GBM patients and discovers medication patterns which influence survival.

## 1.1 Contributions

Our study makes the following contributions:

1. We formulate the problem of predicting patients who survive for longer than the median survival period of one year and discover factors influencing longevity.
2. We leverage the existing sequential pattern mining algorithms and tailor them to mine significant treatment patterns from the available treatment data.
3. We follow a data driven approach to build and evaluate a predictive model to address the problem mentioned above by treating significant treatment patterns themselves as features in addition to the existing clinical and genomic data.

Our long term goal is to recommend personalized treatment options for GBM patients and the first step towards addressing this goal is to ascertain whether the survival period of GBM patients is predictable and discovering the predictive features.

## 2. RELATED WORK

Sequential pattern mining refers to the mining of frequently occurring ordered events or subsequences as patterns [9]. This concept was introduced by Agarwal and Srikant [1] in 1995 based on their study of customer purchase sequences which led to the development of the first algorithm in this area called the GSP (Generalized Pattern Mining) based on the Apriori algorithm to mine frequent itemsets [1]. 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 minimum support for the next pass. SPADE(Sequential Pattern Discovery using Equivalent Classes) on the other hand uses a vertical data format and associates each itemset with an ID\_list [25] which is combination of a sequence id and an event id representing the transaction that the itemset is a part of and the time at which the itemset occurred in that transaction respectively. Length-2 sequence is formed by joining on the same sequence ids of the ID\_lists and where event ids of the itemset of the first sequence is before that of the itemset in the second sequence. 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 [16] is one such algorithm which exploits this approach by building prefix 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.

## 3. 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 [22].

The cBioPortal [4] 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 [24]. 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 [8].

## 4. METHODS

This section gives an overview of the approach used for representing the data and the predictive modeling pipeline de-

veloped to predict the survival duration of patients.

## 4.1 Data Representation

For our study we use data about 300 patients extracted from TCGA spanning over a period of 2 years. 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). Many groups have performed high dimensional profiling studies examining copy number variation (CNV) [3, 20] and gene expression profiling studies which have identified gene signatures associated with EGFR overexpression, clinical features and survival [5, 11, 13, 14, 15, 17, 21, 23]. TCGA Research Network has been established to comprehensively identify genomic abnormalities driving tumorigenesis and had made available a detailed view of genomic changes in a large GBM cohort of 576 patients. From these 91 patients were used and 601 genes were analyzed to describe the mutational spectrum of GBM, confirming previously reported TP53 and RB1 mutations and identifying GBM-associated mutations in such genes as PIK3R1, NF1, and ERBB2. CNV and mutation data on TP53, RB and receptor tyrosine kinase pathways showed that majority of GBM tumors have abnormalities in all of these pathways suggesting that this is a core requirement for GBM pathogenesis [24]

In addition to the data from the clinical and genomic domain, we also analyze 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. Since the data model has a path-oriented nature, 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 [19]. Figure 1 shows the current representation of the data as a graph. The figure 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. The structure of the graph would be enhanced for

diseases, which have extensive treatment guidelines yielding more parameters and potential complications to consider.

## 4.2 Predictive Modeling Pipeline

We give an overview of the predictive modeling pipeline which consists of 4 modules namely ‘Data Standardization and Cleaning’, ‘Sequential Pattern Mining’, ‘Feature Construction’ and ‘Prediction and Evaluation’.

### 4.2.1 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 for some data elements. Missing data is another common issue found in data sets and needs to be taken care of if data is missing for some of the data elements instrumental for our analysis. The data standardization module identifies these different data formats, missing values and creates a standardized clean data set for further analysis. Some of the examples of data standardization and missing value imputation are given below:-

1. Each drug has a start date and an end date which is crucial for extracting the information about the sequence of drug prescriptions. We imputed the missing start or end dates by calculating the mean duration of the prescription of that particular drug. If both the dates were missing then we removed that data point completely.
2. Some of the data elements such as ‘gender’ and ‘drug name’ were standardized since the data from multiple institutions had different nomenclatures
3. Living patients who had their last visit date within 365 days of their diagnosis were filtered out for this study since it was not possible to determine their survival period.

### 4.2.2 Sequential Pattern Mining

For our study we use the concept of sequential pattern mining to extract patterns from the treatment data to give two types of information: first, the sequence of drugs/radiation prescribed and second, their time of prescription. These two blocks of information are clubbed for each patient and used for further analysis.

A treatment pattern may consist of a combination of multiple drugs or radiation or both prescribed in a particular sequence. To mine such treatment patterns we use an approach motivated by GSP & SPADE algorithms. In our approach we define a concept of ‘N-path itemsets’ which consists of sequence of ‘N+1’ treatment instances consisting of ‘N’ edges. E.g Drug\_A -> Drug\_B -> Radiation\_D is a 2-path itemset from the graph model discussed previously, consisting of 2 drugs and one type of radiation therapy (represented as nodes) one immediately following the other 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 we added an event identifier along with each drug node. The event ID also

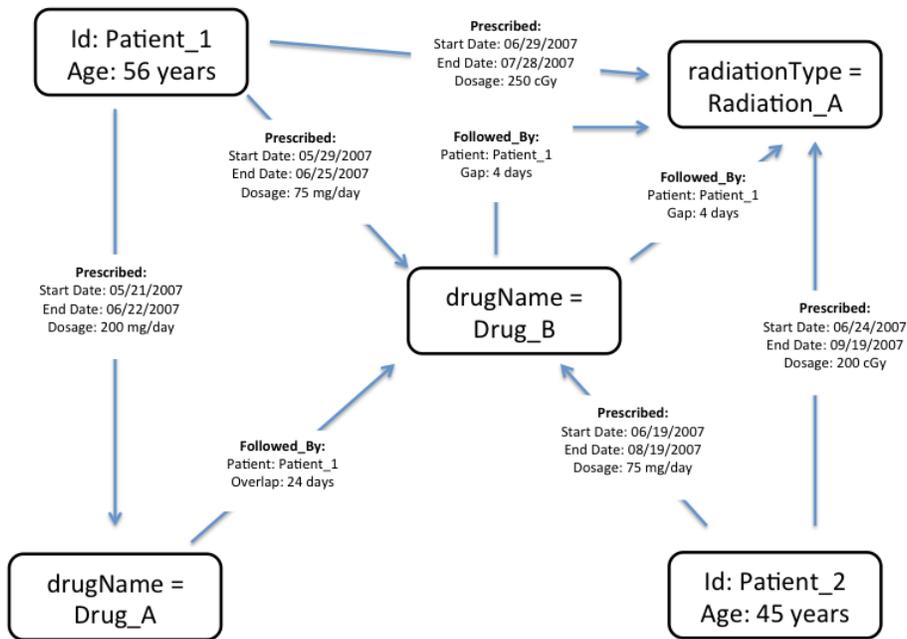


Figure 1: Data represented as a graph

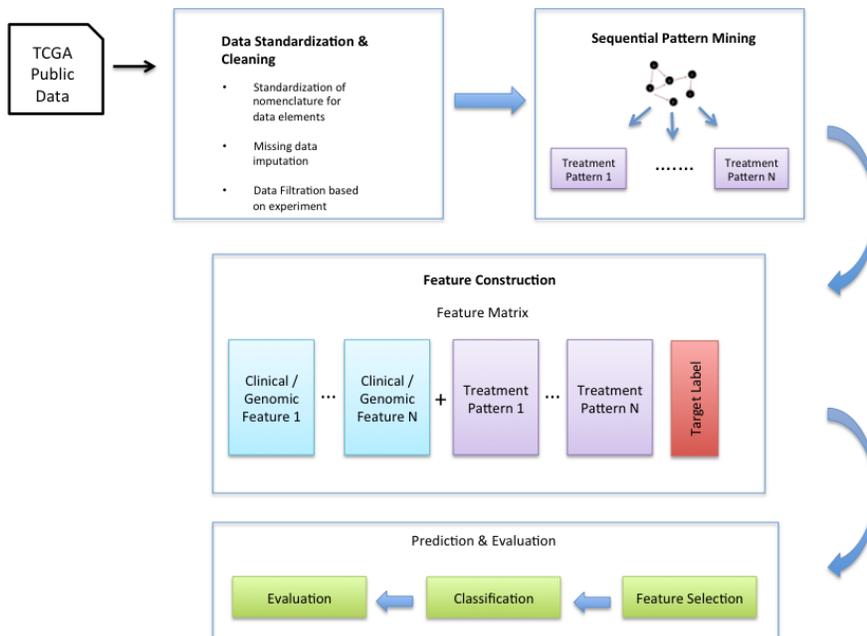


Figure 2: Predictive Modeling Pipeline

helps in executing the ‘immediate sequence’ restriction. An example shown in Fig 3 represents a treatment plan for a patient showing event ids (shown in yellow) attached to each drug node (shown in blue). According to this graph, drugs ‘A’, ‘B’, ‘C’ were prescribed on the same day followed by ‘D’ & ‘E’ followed by ‘F’, ‘D’, ‘A’ in a sequence followed by a ‘F’ again and finally ‘G’.

If there were no event ids associated with the drug nodes then the path C->D->F->G (highlighted in red) would also be considered as a pattern in spite of having intermediate drug prescriptions between ‘F’ and ‘G’. With the event ids in place, a node would be annotated as ‘NodeId.EventID’ which we call a ‘Treatment Event’. A pattern such as C.1 -> D.2 -> F.3 -> G.7 would never occur since ‘G’ at the seventh administering of treatment instance cannot immediately occur after F at the third administration. Examples of patterns which can be formed

in our model are C.1 -> D.2 ->F.3 , D.4 -> A.5 -> F.6 -> G.7, etc.

We extract path sequences of length  $N$  from the data graph and consider the ones prescribed to a significant number of patients for further analysis. This is followed by forming  $N+1$  path sequences increasing the length by one edge at a time and joining on only the event ids. The pseudo-code in Algorithm 1 explains how our algorithm works. The algorithm terminates when no more significant combinations can be formed.

---

**Algorithm 1** Sequential Mining Approach

---

```

1: procedure MINE_TREATMENT_PATTERNS
2:    $N \leftarrow$  Length of path
3:    $minSup \leftarrow$  minimum support
4:    $N = 1$ 
5: getPaths():
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  $A = B$  then
12:        add  $N+1^{th}$  sequence to  $N+1$  path pool
13:   if size of  $N+1^{th}$  path pool = 0 then
14:      $N++$ ;
15:   getPaths();

```

---

### 4.2.3 Feature Construction

Our goal is to transform the data into 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 datasets consists of both numeric and categorical data types. To standardize the data set and avoid creating a bias we converted our dataset into a binary feature matrix. This was achieved by using the categorical data values as features in the feature vector and creating bins for numeric features such as Age, KPS scores, mRNA expression z-scores, etc. E.g Age (in years) which was a numeric value was represented as 4 bins ‘Age < 25’, ‘25 < Age < 50’, ‘50 < Age < 75’ and ‘Age > 75’ which were treated as features.

In addition to these features we add the significant sequence patterns, which were obtained in sequential mining module in the feature vector. Each significant treatment pattern is treated as a feature. A value of 1 is assigned to this feature for patients who exactly received that treatment and for others it is set to 0. The target variable in our study is constructed based on the patient’s survival period. There are two data elements in the data set namely ‘days to death’ and ‘days to last follow up’. The former refers to the number of days between the date of diagnosis and the death of the patient and the latter refers to the number of days between the date of diagnosis and the date of the last follow up with the clinician. For deceased patients we use the ‘days to death’ as an indicator of the 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 the ‘days to

last follow up’ is greater than a year then we assign them a target variable of ‘1’, otherwise we discard that patient since we cannot positively conclude anything about the survival period.

### 4.2.4 Prediction and Evaluation

Our goal in this study is to analyze the clinical and genomic characteristics along with the treatment prescribed within the first year of diagnosis and predict the survival period of patients. The usual cross-validation techniques iteratively partitions the data into a training set and a test set multiple times and evaluates the classifier but in our case we cannot follow this approach since the sequential patterns which act as features are extracted only from the training set because ideally we don’t know anything about the treatment given to the test patients. The prediction module performs a 10-fold cross-validation by randomly creating a training set and test set followed by extraction of treatment patterns from each training set separately for every fold. A classifier such as Logistic Regression with a dot kernel is trained based on the clinical, genomic features along with the resulting treatment patterns before applying the model on the test set. This process is repeated for every fold. We use a greedy forward selection to pick predictive features and the prediction performance evaluated by c-statistic and accuracy is measured on these selected features. We rank the predictive features based on the number of folds in which they were selected by the greedy algorithm and their frequency of occurrence in each fold.

## 5. RESULTS

The data used in this study is extracted from TCGA and consists of approximately 300 patients. 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. The survival period for deceased patients can be calculated from the number of days before death, which is the number of days a patient survived since the day he was diagnosed till his death. From the pool of living patients we add those patients to the ‘survival greater than a year’ category if their last follow up was after a year of them being diagnosed. For our study we have three domains of features which are ‘Clinical’, ‘Genomic’ and ‘Treatment’. Before feeding the data into the predictive modeling module, we perform forward feature selection to pick only significant features for the experiment. In Table 1, we report the performance of various 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 clinical, genomic and treatment features are analyzed together. Since we use 10 fold cross validation for evaluating our models, predictive features are selected for every fold. We rank these features for each domain 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

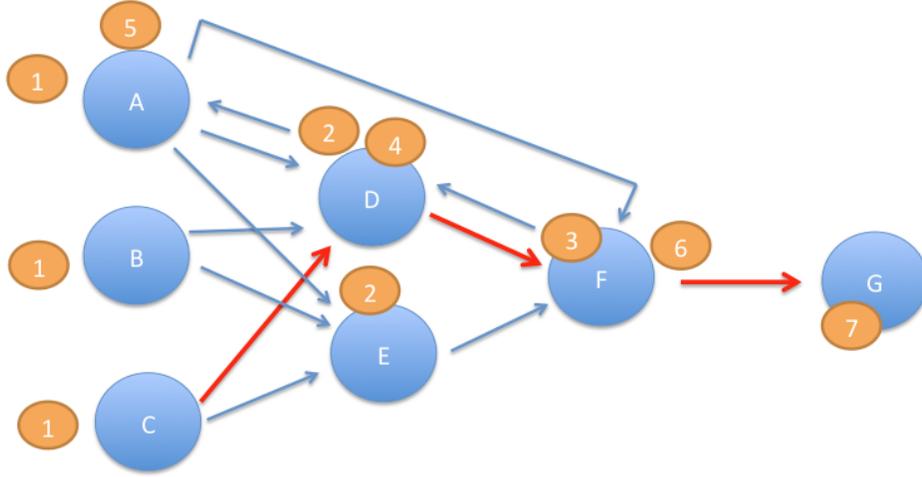


Figure 3: Treatment graph for a patient showing event ids

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%

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. In addition to the P-values showing statistical significance, we also report the nature of the influence, each feature has on the survival period.

### 5.1 Analysis

It has been observed based on the results we report, that patients survive for less than a year if the mRNA expression of the GABRA1 gene is between -1 and -1.5 and survive longer than a year if mRNA expression levels are between -1 and 1. Higher GABRA1 gene expression is associated with 'Neural' subtype of GBM and according to medical experts, patients in this category usually have longer survival periods. Hemizygous deletion of GABRA1 gene has also been observed to have a positive influence on survival. Among the clinical features, it has been observed that older patients especially the ones above the age of 75 have lesser chance of surviving for more than a year.

Most importantly our study also discovered treatment patterns, which have had both positive and negative affects on the survival period. Standard first line of treatment for GBM patients consists of chemotherapy with Temodar cou-

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

Features	No. of folds	Influence on Survival > 1 year	P-value
<b>Genomic</b>			
mRNA expression z-score for GABRA1 gene between -1.5 & -1	4	-	0.007275
mRNA expression z-score for GABRA1 gene between -1 & 1	4	+	<0.0001
Hemizygous deletion of GABRA1 gene	2	+	0.0093
<b>Clinical</b>			
Age of patient > 75 years	3	-	0.0344
<b>Treatment</b>			
External Beam Radiation Therapy.2 -> End of Treatment	5	-	0.00614
CCNU.2 -> End of Treatment	4	-	0.05001
Procarbazine.2 -> End of Treatment	3	+	0.05016

pled with radiation therapy [6]. It was observed that if treatment consists of prescribing radiation therapy using external beam or CCNU by itself as second in the treatment then the patients did not survive for more than a year. This can be explained by the fact that Radiation therapy is usually coupled with Temodar and CCNU is usually not introduced so early in the treatment. A positive influence on survival was observed with Procarbazine when prescribed second in the treatment.

## 6. CONCLUSIONS

In this paper we discuss a pipeline performing data standardization, mining sequential treatment patterns and constructing features for predicting patients surviving for longer periods. We account for patients who are already dead or have survived over a year and discard the ones in treatment for less than one year because of unpredictable outcome. The data standardization module uses averaging and association rule mining based approaches to standardize the data elements since different terminologies and nomenclatures were used to depict similar data elements. In the sequential mining module, we combine the GSP and the SPADE algorithms and tailor the approach to account for the order of administration of a drug within the treatment. The treatment patterns resulting from our approach comprise sequences of drugs, one immediately following the other. Logistic regression is used as a classifier to predict the patients who survive for longer than a year and forward feature selection is the method of choice to select predictive and significant features. Among the genomic features the mRNA expression levels and copy number variation of GABRA1 gene seem to be predictive of survival. The patient's age is the only clinical feature, which was selected by our model. Amongst the treatment patterns, prescription of radiation therapy, CCNU and Procarbazine followed by stoppage of treatment seemed to influence survival. This study is a preliminary step in providing extensive treatment guidance to oncologists / neurosurgeons about the efficacy of certain orderings of drugs and radiation therapies as part of a treatment plan. Currently the treatment patterns consist of the drug names and their event of prescription. In the future we would like to enhance this model by adding constraints such as gap between prescription of drugs, overlapping therapies, filtering clinically insignificant patterns at an early stage, etc.

## 7. REFERENCES

- [1] Agrawal R, Srikant R. "Mining Sequential Patterns". ICDE 1995: 3-14
- [2] Ayres, J., Flannick, J., Gehrke, J., Yiu, T. 2002. 'Sequential pattern mining using a bitmap representation'. In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM Press, 429-435.
- [3] Beroukhi, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., Vivanco, I., Lee, J.C., Huang, J.H., Alexander, S., et al. (2007). Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl. Acad. Sci. USA* 104, 20007-20012.
- [4] Cerami et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discovery*. May 2012 2; 401
- [5] Freije, W.A., Castro-Vargas, F.E., Fang, Z., Horvath, S., Cloughesy, T., Liao, L.M., Mischel, P.S., and Nelson, S.F. (2004). Gene expression profiling of gliomas strongly predicts survival. *Cancer Res.* 64, 6503-6510.
- [6] Glioblastoma and Malignant Astrocytoma: American Brain Tumor Association; 2012. Available at: <http://www.abta.org/secure/glioblastoma-brochure.pdf>
- [7] Holland, E. (2000). "Glioblastoma Multiforme: The terminator." *PNAS* 97(12): 6242-6244.
- [8] Hegi et. al, 'MGMT Gene Silencing and Benefit from Temozolomide in Glioblastoma', *N Engl J Med* 2005; 352:997-1003
- [9] Kamber.M., Han.J.(2006). *Data Mining: Concepts and Techniques* (2nd Edition), Elsevier.
- [10] Krex et al.'Long-term survival with glioblastoma multiforme'.*Brain* (2007) 130 (10): 2596-2606. doi: 10.1093/brain/awm204
- [11] 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 Israel, M.A. (2005). Gene expression profiling reveals molecularly and clinically distinct subtypes of glioblastoma multiforme. *Proc. Natl. Acad. Sci. USA* 102, 5814-5819.
- [12] Mazurowski M.A, Desjardins A, Malof J.M. 'Imaging descriptors improve the predictive power of survival models for glioblastoma patients', *Neuro-Oncology*, Vol.15, Issue 10, Pp. 1389-1394
- [13] Mischel, P.S., Shai, R., Shi, T., Horvath, S., Lu, K.V., Choe, G., Seligson, D., Kremen, T.J., Palotie, A., Liao, L.M., et al. (2003). Identification of molecular subtypes of glioblastoma by gene expression profiling. *Oncogene* 22, 2361- 2373.
- [14] Murat, A., Migliavacca, E., Gorlia, T., Lambiv, W.L., Shay, T., Hamou, M.F., de Tribolet, N., Regli, L., Wick, W., Kouwenhoven, M.C., et al. (2008). 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, 3015-3024.
- [15] Nutt, C.L., Mani, D.R., Betensky, R.A., Tamayo, P., Cairncross, J.G., Ladd, C., Pohl, U., Hartmann, C., McLaughlin, M.E., Batchelor, T.T., et al. (2003). Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res.* 63, 1602-1607.
- [16] Pei et al. 'PrefixSpan: Mining Sequential Patterns by Prefix-Projected Growth' ICDE, page 215-224, IEEE Computer Society, (2001)
- [17] Phillips, H.S., Kharbanda, S., Chen, R., Forrester, W.F., Soriano, R.H., Wu, T.D., Misra, A., Nigro, J.M., Colman, H., Soroceanu, L., et al. (2006). Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell* 9, 157-173
- [18] D. Williams Parsons, S. J., Xiaosong Zhang, Jimmy Cheng-Ho Lin, Rebecca J. Leary, Philipp Angenendt, Parminder Mankoo, Hannah Carter, I-Mei Siu, Gary L. Gallia, Alessandro Olivi, Roger McLendon, B. Ahmed Rasheed, Stephen Keir, Tatiana Nikolskaya, Yuri Nikolsky, Dana A. Busam, Hanna Tekleab, Luis A. Diaz Jr, James Hartigan, Doug R. Smith, Robert L. Strausberg, Suely Kazue Nagahashi Marie, Sueli Mieko Oba Shinjo, Hai Yan, Gregory J. Riggins, Darell D. Bigner, Rachel Karchin, Nick Papadopoulos, Giovanni Parmigiani, Bert Vogelstein, Victor E. Velculescu, Kenneth W. Kinzler (2008). "An Integrated Genomic Analysis of Human Glioblastoma Multiforme." *Science* 321(5897): 1807-1812.
- [19] Robinson.I , Webber.J, Eifrem.E (2013). Graph

Databases. California, O'Reilly Media Inc.

- [20] Ruano, Y., Mollejo, M., Ribalta, T., Fiano, C., Camacho, F.I., Gomez, E., de Lope, A.R., Hernandez-Moneo, J.L., Martinez, P., and Melendez, B. (2006). Identification of novel candidate target genes in amplicons of glioblastoma multiforme tumors detected by expression and CGH microarray profiling. *Mol. Cancer* 5, 39.
- [21] Shai, R., Shi, T., Kremen, T.J., Horvath, S., Liao, L.M., Cloughesy, T.F., Mischel, P.S., and Nelson, S.F. (2003). Gene expression profiling identifies molecular subtypes of gliomas. *Oncogene* 22, 4918-4923.
- [22] Network, T. R. (2010). The Cancer Genome Atlas Data Portal, National Institute of Health.
- [23] Tso, C.L., Freije, W.A., Day, A., Chen, Z., Merriman, B., Perlina, A., Lee, Y., Dia, E.Q., Yoshimoto, K., Mischel, P.S., et al. (2006). Distinct transcription profiles of primary and secondary glioblastoma subgroups. *Cancer Res.* 66, 159-167.
- [24] 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.
- [25] Zaki.M.J. "SPADE: An Efficient Algorithm for Mining Frequent Sequences". *Machine Learning* 42(1/2): 31-60 (2001)