CHAPTER - 3

ANALYSIS OF CLINICAL DATA STORED IN THE WAREHOUSE:
ASSOCIATION MINING BASED STUDY FOR
IDENTIFICATION OF CLINICAL PARAMETERS
AKIN TO OCCURRENCE OF BRAIN TUMOR
3.1 Introduction

The characteristics of clinical data as it originates during the process of clinical documentation, includes issues of data availability and complex representation models, which can make data mining applications challenging. Henceforth, data preprocessing and transformation are required before one can apply data mining to clinical data. The application of classical data warehousing process should be thus able to answer the queries being raised. It should also be able to mitigate issues like appropriate storage structure of clinical data, varied sources of data, reduce the dimensionality constraint, and handle multiple data variables. Thus it would make it easier for researchers and data analysts to acquire the data and information they need.

The stored data in the warehouse would provide a basis for the analysis of risk factors for the disease. For example, we can compare tumor with non-tumor patients to find patterns associated with occurrence of brain tumor. This method has been common practice in evidence-based medicine, which is an approach in which a clinician is aware of the evidence in support of clinical practice, and its associated strength [1]. The interpretability of results is a requirement for designing a data mining method for medical applications. In general, medical practitioners and researchers do not care how sophisticated a data mining method is, but they do care how understandable its results are [1]. Rules are a type of the most human-understandable knowledge, and therefore it is most suitable for deciphering of new rules corresponding to data associated with medical applications. Association rule mining [2] is a general purpose rule discovery scheme that has been widely used for discovering rules in medical applications [3-5]. It retrieves all frequent patterns in a data set and forms interesting rules among frequent patterns [6]. Association rule mining has been used to find disease-disease, disease-finding, and disease-drug co-occurrences in electronic health record data [7-8], demonstrating the importance of finding disease co-occurrences. Association rule mining using objective measures and transitive inference for pruning, have been used in the clinical domain, to find associations between medications and clinical problems using electronic health record data highlighting some of the challenges in identifying valid associations [9]. Studies made by Brossette et al. [3], Paetz and Brause[5], Ohsaki et al .[4], Ordonez et al.[10], based on association mining technique, states about associative rules corresponding to hepatitis, heart diseases, etc. which has made a path breaking impact in the healthcare sector.

The objective of this study is to propose for a data mining process, that can be used for assessment of patients for brain tumor (primary stage) and discover associative rules based on clinical diagnostic parameters. Based on the associative clinical parameters deciphered, we propose for a predictive model which can be used for an early prediction of brain tumor in suspected
patients independent of results from MRI, CT scan, arteriogram or small dime craniotomy. Applying association rule mining to a given clinical data set has the potential to confirm existing knowledge regarding disease co-occurrences as well as to discover new disease relationships that could potentially lead to improved clinical health care.

3.2 Materials and Methods

The path of knowledge discovery process is said to be complete when knowledge has been extracted from pool of data. The said path involves collection, cleaning and storage of data followed by mining of knowledge from this pool. Considering the same, this study focuses on deciphering the clinical parameters that can be associated with the 'STATE' of brain tumor by applying association rule discovery algorithm. For a patient not having tumor, 'STATE' is represented as 0 while for diseased as 1. The approach used for this study has been demonstrated in flow diagram 1.

Flow Diagram 1 - Representation of knowledge discovery process (identification of clinical parameters associated to primary brain tumor identification).
3.2.1 OLAP cube definition for data selection

In this study, data for brain tumor patients from the clinical warehouse was extracted by using an OLAP cube. As varied dimensionality was observed in the data, on consultation with oncologists appropriate data forms were defined for the cube. The set of clinical parameters selected for the study focuses on blood analysis result, KFT (Kidney Functionality Test) result, LFT (Liver Functionality Test) result, sugar level, triplets of blood pressure and MRI/CT scan images.

3.2.2 Data preprocessing

Information of 550 patients, out of which 350 patients were tested for presence of brain tumor (positive cases) and 200 patients were diagnosed for absence of brain tumor (negative cases) from hospitals across India, stored in the warehouse was used for this study. Pre-processing of data in the warehouse was done using STATISTICA DATAMINER 9.1 [11], to select the features for mining purpose. We have used systematic tests of (1) missing value identification, (2) selection of integrated forms of data, (3) identification of incorrect values based on prescribed scale [12] and (4) Feature selection. From the feature selection step, the parameters selected for the study were: Haemoglobin_content, Total_Leucocyte_count (TLC), Eosinophils, Neutrophils, Lymphocytes, Monocytes, Platelet count, KFT_Creatinine (Kidney Functionality Test - Creatinine), KFT_BUN (Kidney Functionality Test - Blood Urea Nitrogen), LFT_Sr_Bilirubin (Liver Functionality Test - Serum Bilirubin), LFT_ALP (Liver Functionality Test - Alkaline Phosphatase), LFT_SGOT (Liver Functionality Test - Serum Glutamic Oxaloacetic Transaminase), LFT_SGPT (Liver Functionality Test - Serum Pyruvic Transaminase). Each of the said parameter values was processed into qualitative form & labeled as HIGH, NORMAL or LOW based on prescribed clinical ranges [12].

3.2.3 Association rule discovery algorithm

This study focuses on identification of clinical parameters that can be associated with progressive state of a disease by implementing association mining algorithm. It is a popular data mining technique [13] that tries to find interesting patterns in large databases [14]. The Apriori algorithm exploits the downward closure property, which states that if an item set is infrequent, all of its supersets must be infrequent too. The classic framework for association rule mining uses support and confidence as thresholds for constraining the search space. Each item set has an associated statistical measure called support. For an itemset X ⊆ I, support(X) = s, if the fraction of transactions in the dataset D containing X = s [6]. The confidence of an association rule X → Y in D is the conditional probability of having Y contained in a transaction, given that X is contained in that transaction: confidence (X → Y) = P(Y|X) = support(XY)/support(X) [14]. A confidence value
of 100 for a certain rule means that the possibility of obtaining outcome Y when X is a given condition (X → Y) is 100%; if not, the possibility of A → B is defined as a value (possible rule) between 0 and 100.

It is arduous to predispose appropriate criteria for any two parameters in association rule mining, because information is obtained based on a minimum threshold for support and confidence [14]. As such, in this study, the frequent item sets were discovered based upon selected parameters for pre-processed clinical dataset that were subjected to confidence of at least 85%, when the minimum support was defined to 30%. STATISTICA DATAMINER 9.1 [11] was used to calculate the frequency of each item set with support % criteria of at least 30 along with head and body iteration rate of 10. All frequent item sets obtained were subjected for the discovery of association rules.

3.2.3.1 Calculation of Frequent Clinical Parameters

STATISTICA DATAMINER 9.0 [22] was used to calculate the frequency of each item-set with support % criteria of at least 30 along with head and body iteration rate of 10. Analysis of the item-sets satisfying the criteria indicates further analysis of the following clinical parameters can indicate their significant relationship with STATE of the disease

3.2.3.2 Association Rule Mining

All the frequent item set obtained with at least 30% support criteria were subjected for the discovery of association rules. Association mining was performed using STATISTICA DATA MINER 9 [11]. STATE was declared as the response indicator and the other parameters were defined as categorical indicators. The final confidence to deduce rule was set to at least 85% through a physician’s opinion and the process was executed with antecedent and precedent iteration rate of value 10.

3.2.4 Predictive Model

The parameters found to be associated with occurrence of tumor were selected to build a predictive model using normalized regression approach as given by equation i:

$$\Theta_j := \Theta_j - \alpha \frac{\partial J(\Theta)}{\partial \Theta_j}$$  ... equation (i)

In normalized regression approach we try to obtain the minimal set of coefficients (\(\Theta_j\)) for the independent parameters by varying the learning rate (\(\alpha\)). For example, in case of a simple linear regression (\(y = a+bx\)) we try to get minimal set of coefficients i.e. value for \(a\) & \(b\). The learning rate
was varied from 0.001 to 0.1 to obtain $\Theta_j$. Convergence (steepest decent approach) was observed at $\alpha = 0.04$.

Jackknifing was applied for cross-validation of the predictive model along with accuracy, sensitivity and specificity analysis.

3.3 Results

Haemoglobin_content, TLC, Platelet Count, KFT_Creatinine, KFT_BUN (Blood Urea Nitorgen), LFT_Sr_Bilirubin, LFT_ALP, LFT_SGOT and LFT_SGPT are the parameters that showed support of at least 30%. Item-sets satisfying the support % subjected to discovery of association rules within the specified mining criteria showcased association of high values of Creatinine, BUN, SGOT & SGPT with presence of tumor in patients. Table V (A & B) enlists various association rules that are discovered pertaining to occurrence and non-occurrence of brain tumor.

<table>
<thead>
<tr>
<th>Association Rule</th>
<th>Support %</th>
<th>Confidence %</th>
<th>Correlation %</th>
</tr>
</thead>
<tbody>
<tr>
<td>KFT_Creatinine = HIGH $\Rightarrow$ KFT_BUN = HIGH</td>
<td>56.75</td>
<td>100</td>
<td>77.45</td>
</tr>
<tr>
<td>KFT_Creatinine = HIGH $\Rightarrow$ STATE = 1</td>
<td>56.75</td>
<td>100</td>
<td>77.77</td>
</tr>
<tr>
<td>KFT_BUN = HIGH $\Rightarrow$ STATE = 1</td>
<td>78.37</td>
<td>85.29</td>
<td>90.8</td>
</tr>
<tr>
<td>KFT_Creatinine = HIGH, KFT_BUN = HIGH $\Rightarrow$ STATE = 1</td>
<td>56.75</td>
<td>100</td>
<td>79.77</td>
</tr>
<tr>
<td>LFT_SGOT = HIGH $\Rightarrow$ STATE = 1</td>
<td>62.16</td>
<td>98.83</td>
<td>81.72</td>
</tr>
<tr>
<td>LFT_SGOT = HIGH, LFT_SGPT = HIGH $\Rightarrow$ STATE = 1</td>
<td>62.16</td>
<td>95.83</td>
<td>85.71</td>
</tr>
<tr>
<td>LFT_SGPT = HIGH $\Rightarrow$ STATE = 1</td>
<td>81.08</td>
<td>88.23</td>
<td>89.56</td>
</tr>
<tr>
<td>Haemoglobin_content = NORMAL $\Rightarrow$ STATE = 1</td>
<td>59.45</td>
<td>100</td>
<td>81.64</td>
</tr>
</tbody>
</table>

Table V (A) - Association Rules deciphered for clinical parameters corresponding to occurrence of brain tumor (Min. Support - 50%; Confidence - 85%)
<table>
<thead>
<tr>
<th>Association Rule</th>
<th>Support %</th>
<th>Confidence %</th>
<th>Correlation %</th>
</tr>
</thead>
<tbody>
<tr>
<td>KFT Creatinine = HIGH =&gt; STATE = 0</td>
<td>6.75</td>
<td>100</td>
<td>77.77</td>
</tr>
<tr>
<td>KFT_BUN = HIGH =&gt; STATE = 0</td>
<td>8.7</td>
<td>100</td>
<td>90.8</td>
</tr>
<tr>
<td>KFT Creatinine = NORMAL, KFT_BUN = NORMAL =&gt; STATE = 0</td>
<td>96.85</td>
<td>100</td>
<td>79.77</td>
</tr>
<tr>
<td>LFT_SGOT = HIGH =&gt; STATE = 0</td>
<td>2.63</td>
<td>98.3</td>
<td>81.72</td>
</tr>
<tr>
<td>LFT_SGOT = NORMAL =&gt; LFT_SGPT = NORMAL, STATE = 0</td>
<td>92.26</td>
<td>100</td>
<td>85.71</td>
</tr>
<tr>
<td>LFT_SGPT = HIGH =&gt; STATE = 0</td>
<td>11.08</td>
<td>98.32</td>
<td>89.56</td>
</tr>
</tbody>
</table>

**Table V (B)** - Association Rules deciphered for clinical parameters corresponding to non-occurrence of brain tumor (Min. Support - 02%; Confidence - 85%)

Based on the parameters identified from associative rules with 85% (Creatinine, BUN, SGOT, SGPT) & 75% confidence (Hemoglobin Content, Alkaline Phosphatase and Serum Biliurinbin), a predictive model is proposed to predict the possible STATE of an individual i.e whether suffering from tumor or not. Most significant model was obtained at α = 0.04 & is represented by the equation ii:

\[
\text{STATE} = 0.171 + 0.0491 \text{Hemoglobin content} + 0.0652 \text{KFT Creatinine} + 0.0171 \text{KFT BUN} - 0.0504 \text{LFT Sr Bilirubin} + 0.0304 \text{LFT ALP} - 0.07 \text{LFT SGPT} + 0.0806 \text{LFT SGOT}
\]

... equation (ii)

**3.4 Discussion**

Usually high value of creatinine indicates any renal functional impairment (intrinsic renal lesions, decreased perfusion of the kidney, or obstruction of the lower urinary tract), acromegaly and hyperthyroidism, while that of BUN (Blood Urea Nitrogen) indicates acute & chronic intrinsic renal disease or post renal obstruction of urine because of high protein intake. The SGOT (serum glutamic-oxaloacetic transaminase) test, also known as an AST test, measures the amount of a protein enzyme called glutamic-oxaloacetic transaminase occurring in blood. The SGOT enzyme can be associated with functioning of skeletal muscles, red blood cells, heart muscles, kidney tissue and with the brain as well. An SGPT blood test is a test used to measure the amount of the enzyme glutamate pyruvate transaminase (GPT) in blood and usually associated with occurrence of diseases
like cirrhosis and hepatitis. However the results of this study suggests that the described factors can also be associated in a combined form with occurrence of the disease - brain tumor (primary stage). Diagnostic value of Creatinine & Urea nitrogen (BUN) which are usually tested as part of Kidney Functionality test and; SGOT & SGPT which are usually tested as part of Liver Functionality test were found to be unusually high with no abnormalities reported for Kidney or Liver for patients diagnosed by brain tumor in primary stage. The study suggest Creatinine, Urea Nitrogen, SGOT & SGPT based values can be associated together and used for deterministic analysis for STATE of the disease and its early screening. There have been significant associative rules observed corresponding to the discovered parameters with respect to STATE parameter of brain tumor. There is 100% confidence observed corresponding to Creatinine and Blood Urea Nitrogen association with the disease whereas 95% confidence with SGOT and SGPT. Also the association mining based study suggests that Haemoglobin_content is usually normal along with other blood related parameters in case of patients suffering from brain tumor during the primary stage with 100% confidence.

The cross-validation results obtained from Jackknifing: $R^2_{(calculated)} = 74.66\%$ and PRESS (predicted residual sum of squares) $= 1.67$; along with accuracy observed $= 75\%$, sensitivity $= 83\%$, specificity $= 62\%$ ($n = 326$; TP $= 50.9\%$; FP $= 14.7\%$; FN $= 10.4\%$; TN $= 23.9\%$), indicates the model has reasonably good predictive accuracy.

3.5 Conclusion

This study primarily focuses on discovery of clinical parameters that can be associated with occurrence of brain tumor which are rarely focused upon, by applying association rule mining algorithm. The study highlights four of the clinical factors, usually tested for Kidney & Liver functionality, to be directly associated with occurrence of brain tumor for patients diagnosed in the primary stage. Based on the discoveries made in this study a predictive model is proposed for its early diagnosis. For robustness & higher accuracy, the model proposed in the study needs to be further validated by including data set of patients suffering from other kind of tumors, renal functional impairment, kidney based problems, metastatic brain tumor and brain related other diseases.
References


