TRANSLATIONAL AND HIGH END COMPUTING OF CLINICAL DATA IN INDIA

BY

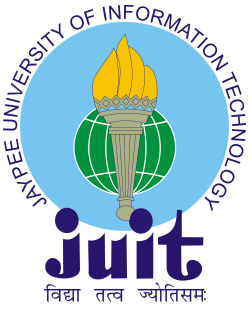
DIPANKAR SENGUPTA
Enrollment No. 106506

SYNOPSIS

FOR THE THESIS TO BE SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
BIOINFORMATICS

UNDER THE GUIDANCE OF
DR. PRADEEP KUMAR NAIK



DEPARTMENT OF BIOTECHNOLOGY & BIOINFORMATICS
JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY
WAKNAGHAT, SOLAN-173234, HP, INDIA

DECEMBER 2013

TRANSLATIONAL AND HIGH END COMPUTING OF CLINICAL DATA IN INDIA

DIPANKAR SENGUPTA

INTRODUCTION

The ability to collect and store data has grown at a dramatic rate in all disciplines over the past decade or so with new techniques being developed for effective storage and analysis. Healthcare has been no exception. Advancement in clinical research and diagnostic processes has lead to a phase corresponding to generation of large amount of data that are heterogeneous in nature. Immense efforts have been made recently for detailed heterogeneous clinical data analyses. As a result a new branch of science has emerged, **clinical bioinformatics**, which combines clinical informatics, bioinformatics, information technology, mathematics, and omics science to a common platform [1]. Among these, **clinical informatics** primarily focuses on various forms of clinical data such as patient complaints, history, clinical symptoms and signs, physician's examinations, biochemical analyses, imaging profiles, pathologies, therapies and other measurements [2] pertaining to clinical diagnostics of a patient. The shift toward evidence-based research presents significant opportunities to extract meaningful information and transform into the knowledge from this clinical data [3]. Interpreting data across multiple systems is challenging and various integration techniques with varying levels of complexity have been proposed to solve the problem of data integration and storage [4]. However, research reveals that the current designs are not efficient for data sets with large numbers of attributes that vary over time [5]. Henceforth, development of novel informatics techniques based on mathematical or statistical models are essential. This development will provide a better understanding of the nature of complex diseases and help in guiding more accurately & improved diagnosis for better therapies.

Path breaking step in the field of clinical informatics was the development of Electronic Health/Medical Records (EHR/EMR) which led to evolution of information technology in the field of clinical sciences[6]. As an effort to facilitate access to this wealth of information, databases were developed that contained clinical data from healthcare organizations [7]. The enormous amount of data collected by EHR/EMR has found additional value when integrated and stored in database. It will be easier to apply data mining techniques like co-occurrence analysis, association mining based study, etc. on this structured data form. As an archetype, National Cancer Institute, USA has developed a medical knowledge information system integrated with data mining applications [8]. Similarly, New York-Presbyterian Hospital, USA is using an electronic health record system for the past couple of years and maintaining a longitudinal record for each of its patients [9]. Data mining

technique such as co-occurrence statistics is a congruous technique to analyze the clinical data as a disease and its associated findings appear together rather than in random combinations [8]. Similarly, technique using association rule mining [10] is a general purpose rule discovery scheme and has been widely used for discovering rules based on the importance of finding disease co-occurrences.

However, in India the research & development in this field is still in nascent stage. Indian policy makers are yet to realize the importance of clinical informatics in delivering healthcare [11]. There is lack of focus on electronic medical related information repository building along with digitalization of all the medical related documents like medical report, diagnostic measurements, etc.[12]. With an aim to encourage the adoption of standards for healthcare information communication within India, HL7 Healthcare Standard Institute (HL7 India) was started, which is an independent & non-profit-distributing organization [13]. The objective of HL7 India is to support the development, promotion and implementation of HL7 standards and specifications in a way which addresses the concerns of healthcare organizations, health professionals and healthcare software suppliers in India. Their current research focuses on XML processing for data interchange between the hospitals [13]. Similarly, Indian Association for Medical Informatics (IAMI) was started up in 1993 with a mission to introduce use of computer & its application all the colleges of medical sciences, dental sciences, nursing and pharmacy in India through Medical Council of India and respective university authorities [14].

Existing Issues concerned to Clinical Informatics & its application

Information technology (IT) is no exception with a good track of success in many fields such as finance, insurance, banking, tourism, etc. However, it is obvious that this is not the case in clinical sciences & healthcare in spite of the vast amount of literature on the proven benefits of IT for tackling the fundamental problems of healthcare [15]. The identified reasons for this paradox are many fold but they mainly focus on added difficulty to information system (IS) development due to the complexity and volatility of clinical data. Another important reason roots from the inherent subjective or non-deterministic nature of multi dimensionality form of this data. Not only is the body of knowledge highly variable but also the practice changes from time to time and place to place. National Institute of Health (NIH), USA in 2011 (17th Aug 2011) has stated about the various aspects which needs to be focused upon pertaining to application of IT in clinical informatics. The proposed roadmap identified a lack of communication between basic and clinical scientists as a major roadblock to the development of translational technologies. The key issues raised which needs to be worked upon were:

1. How to create a patient's health record?

2. How to build a lifelong health history for a patient from information stored in multiple, diverse systems?
3. How to identify a patient uniquely and reliably in each visit?
4. How to join-up different systems in different platforms?
5. How to analyze the health history of patients from a storage source?
6. How to achieve flexibility & agility?
7. How to achieve performance & scalability?

Considering the dynamic and fast track advances in IT, I believe it will be more efficient to address the multidimensionality problem of data along with maintaining health history of a patient, which is the general motivation for this study. The specific problems being addressed in the study are:

1. Storing the data pertaining to a patient temporally in a non-volatile state.
2. Design an approach for temporal mining of the stored clinical data.
3. Decipher patterns and rules from the clinical data set by applying combinatorial mining approaches – predictive models for early identification of diseases.

OBJECTIVES

The objective of this research study is to propose a clinical mining process, that can be used for storage of patients clinical data temporally and use the same for mining hidden/predictive patterns. The data warehouse being proposed for storage of clinical data, should be able to render the data in appropriate structures, provide metadata that adequately records semantics of data and reference pertinent medical knowledge. The data in the warehouse is subject to mining for observing new patterns. Applying mining algorithm 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. The specific goals being focused in this study are:

- To design a clinical dimensional model for development of clinical warehouse.
- Analysis of clinical data stored in the warehouse: association mining based study for identification of clinical parameters akin to occurrence of brain tumor.
- To design an algorithm for mining temporal form of clinical data set.
- Identification of key cognitive and clinical measures for evaluating cognitive performance of human population at high altitude.

➢ PROPOSED ARCHITECTURE & OUTLINE OF WORK

Each of the said objectives has distinct characteristic. At the same time they are related to one another. To clearly and coherently demonstrate the goal, results and conclusion of each piece of work, I have arranged each work as separate chapter in a publishing format. The format will benefit readers to understand the idea of development, conclusion, coherence and full significance as each chapter will be a full manuscript from background to conclusion at publication stage. An overall architecture of the proposed model is being depicted in figure 1, demonstrating a clear understanding for the integration of clinical data and its translation.

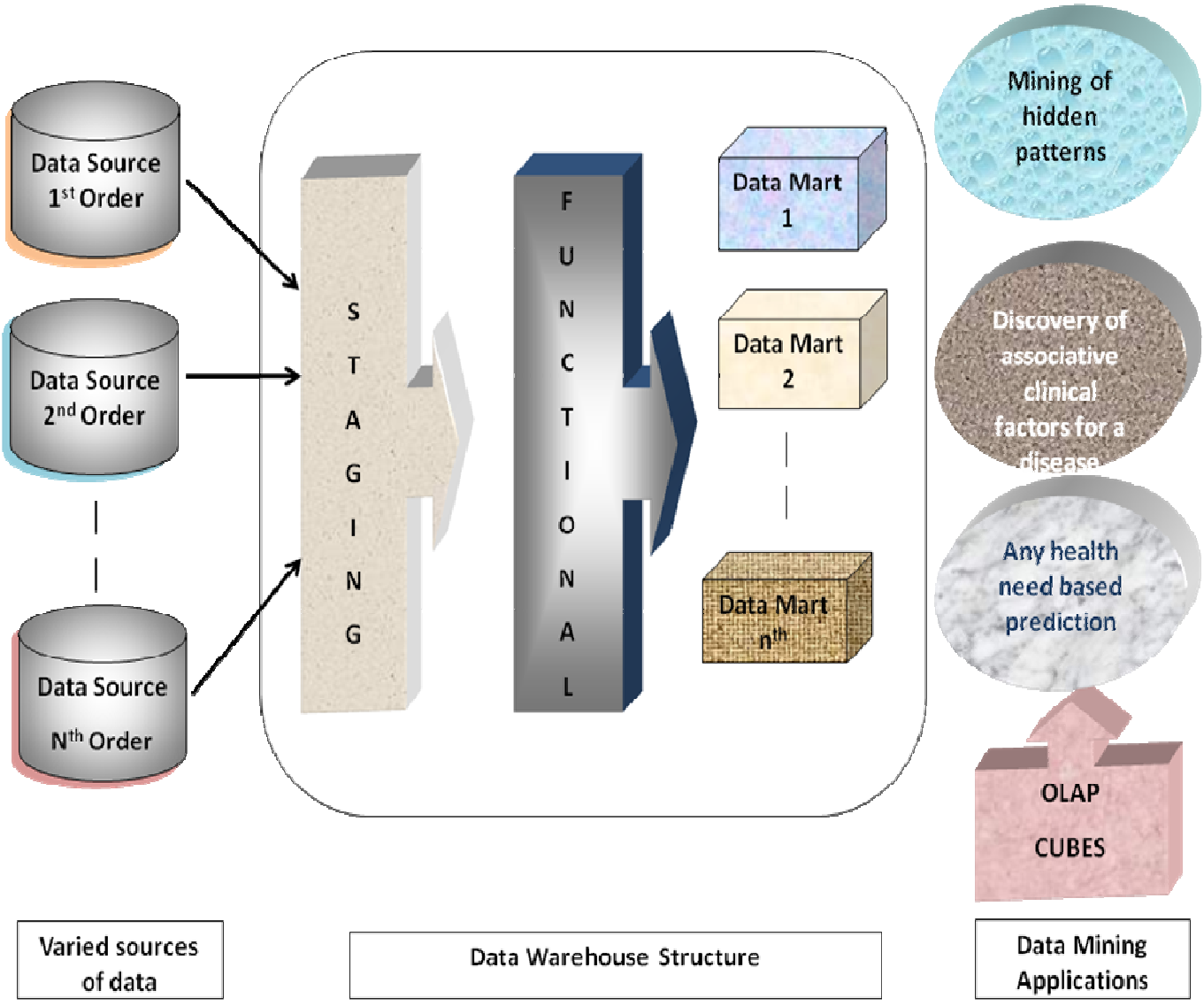


Figure 1 - Architectural design for the proposed Business Intelligence model of clinical data temporal management & analysis.

CHAPTER 1 - Design of Dimensional Model for Clinical Data Storage and Analysis

Current research in the field of life and medical science is generating chunk of data on daily basis. It has thus become a necessity to find solutions for efficient storage of this data, trying to correlate and extract knowledge from it. Clinical data generated in hospitals, clinics & diagnostics centres is falling under a similar paradigm. Patient's records in various hospitals are increasing at an exponential rate, thus adding to the problem of data management and storage. Major problem being faced corresponding to storage is the varied dimensionality of the data, which ranges from images to qualitative/quantitative form. Therefore, there is a need of an efficient data model which can handle this multi-dimensionality data issue and store the data in temporal aspect.

To address the clinical data integration issues and to have a data warehouse based storage structure that can effectively handle the data in temporal form, a clinical dimensional model is being proposed. The proposed model will also address the concerned dimensionality issue. Ralph Kimball [16] addressed about the typical health care cycle, but has discussed the entities in detail concerned with typical billing cycle. However, with respect to challenges highlighted and research being carried out in the fields of genomics, proteomics, etc. along with clinical sciences, it can be associated with personalized medication and therefore would need storing the data at the granular level of a person. The various domains that can be said to associated with effective recording of an individual health data is depicted in figure 2. In nearby future, along with the clinical and drug data, genomics and proteomics data are also going to play a major role for an effective treatment process. Each of the said domains will lead to development of a specific data mart associated in the data warehouse. This research is focusing on one of the said domain, i.e. of clinical data, by proposing a dimensional model design that can be used for development of a clinical data mart.

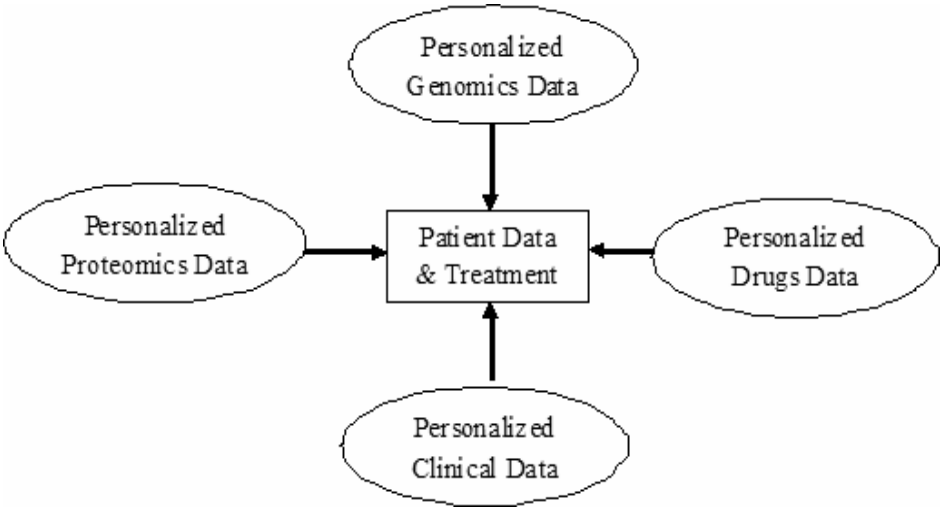


Figure 2 - Domains associated to health care
(Different domains which can be associated with health care in future)

Material & Methods

With the current advents the clinical domain associated with the health cycle needs major attention. The major problem being faced is of varied dimensionality, ranging from images to numerical form of data that needs to be answered (Figure 3).

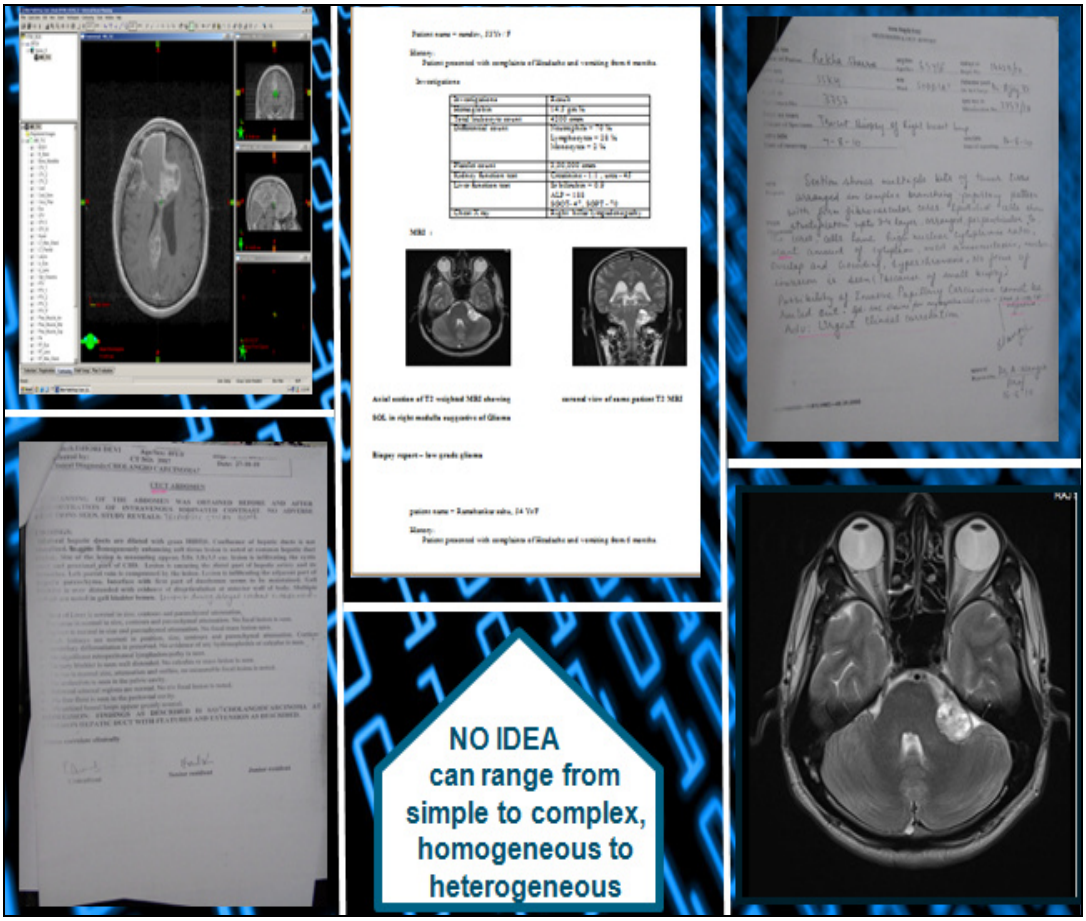


Figure 3 – Various sources of clinical data, ranging from textual, numerical to images.

Based on the heterogeneity observed in the data, we propose an appropriate clinical dimensional model for the structure of a clinical data mart, that will store data at the granular level of an individual, pertaining to time. The given model has been designed using Erwin data modeller (version 8.2) [17] and is represented in star schema representation [18]. While designing the model it has also been taken into consideration that the data in some of the dimensions may change, for which additional attributes are added, so that they can act as slowly changing dimension (SCD) [19-21]. During physical implementation of the data mart an SCD-Type II implementation was made for such dimensions [19]. The data extraction, cleaning & processing process was carried out using ETL (Kettle 3.0) technology (figure 4) and an optimal existing RDBMS package (MySql 5.019) was used to physically create the data warehouse.

Figure 4 – Flow structure of data in the warehouse managed by ETL codes.

The aim of building this warehouse is to lead to a platform for applying data mining technique to find correlation among various attributes, development of decision trees for classification, applying association mining studies to identify associative factors for a disease, etc. which would help in deciphering new translational paradigms that can be used by physicians, health professionals and even by a common man who has knowledge about how to use computer & internet.

Result & Discussion

Proposed design (Figure 5) consists of two fact tables - Fact_Patient and Fact_Patient_Image_Detail, which stores the textual measures and numerical measures obtained from the images respectively. In the given dimensional model, the Fact_Patient table (which would keep track of the numerical measures for diagnostic factors) is referencing to Dim_Date, Dim_Time, Dim_Patient, Dim_Disease, and Dim_Diagnostic_Test and the Fact_Patient_Image_Detail is referencing to Dim_Date, Dim_Time, Dim_Patient and Dim_Image dimension, respectively. Patient_Id serve as the primary key of the Dim_Patient dimension table, which is the unique id generated for each patient and is majorly linking to all other information associated. The dimension further includes other descriptive information associated to a patient like name, age, gender, etc. Keeping in consideration the data in patient dimension may change, Start_date, End_date and Flag attributes have been added so that it can act as SCD. While designing the clinical dimensional model the temporal prospect was taken into consideration, henceforth Date & Time dimension was included. Date_ID is the primary key for Dim_Date dimension, which assigns unique id to each of the date value. The dimension also include various date based attributes like month, week, calendar year, quarter, etc., which can help to make an analysis considering different period. Time_Id is the primary key for Dim_Time which assign unique id corresponding to each second of a minute and hour. Separate inclusion of Time dimension ensures irrespective of number of times a test is conducted for a patient on any given date, each measure would be recorded uniquely in the Fact_Patient table.

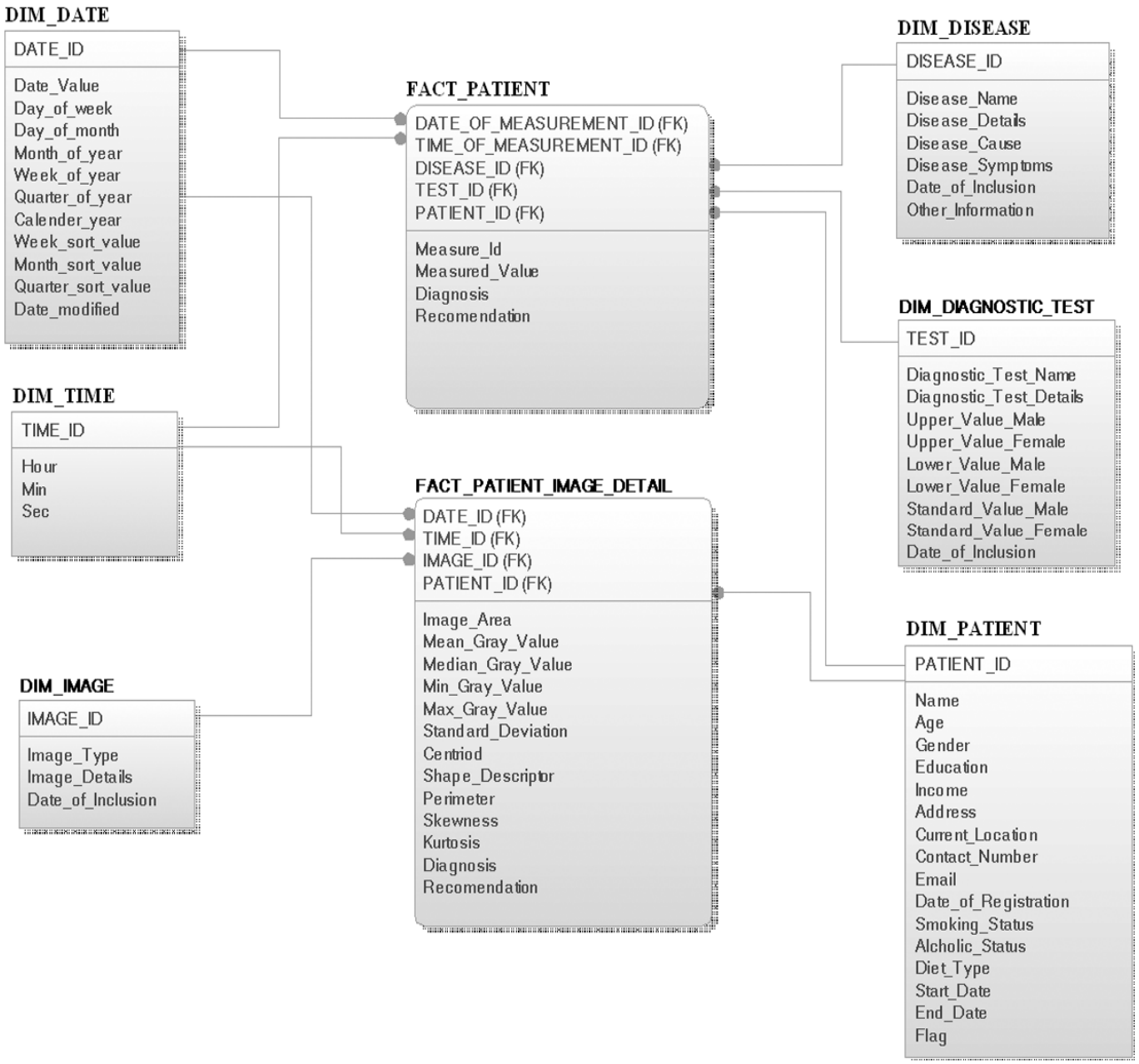


Figure 5 - Logical data model representation for proposed clinical dimensional model.

Disease_Id and Test_ID are the primary keys of Dim_Disease and Dim_Diagnostic_Test dimensions respectively. They include various attributes which would describe diseases and various diagnostic tests respectively. Patient_Id, Disease_Id, Test_Id, Date_of_Measurement_Id and Time_of_Measurement_Id act as composite primary key for Fact_Patient table. It stores with respect to unique key each of the measured values. The Image Dimension (Dim_Image) is linked to Fact_Patient_Image_Details; here Patient_Id and Image_Id (in combination) with Date_Id and Time_Id act as the composite key. The Fact_Patient_Image_Details include attributes which would store measures corresponding to numerical conversion of images like area, skewness, mean gray value, etc.

CHAPTER 2 - Analysis of clinical data stored in the warehouse: Association mining based study for identification of clinical parameters akin to occurrence of brain tumor

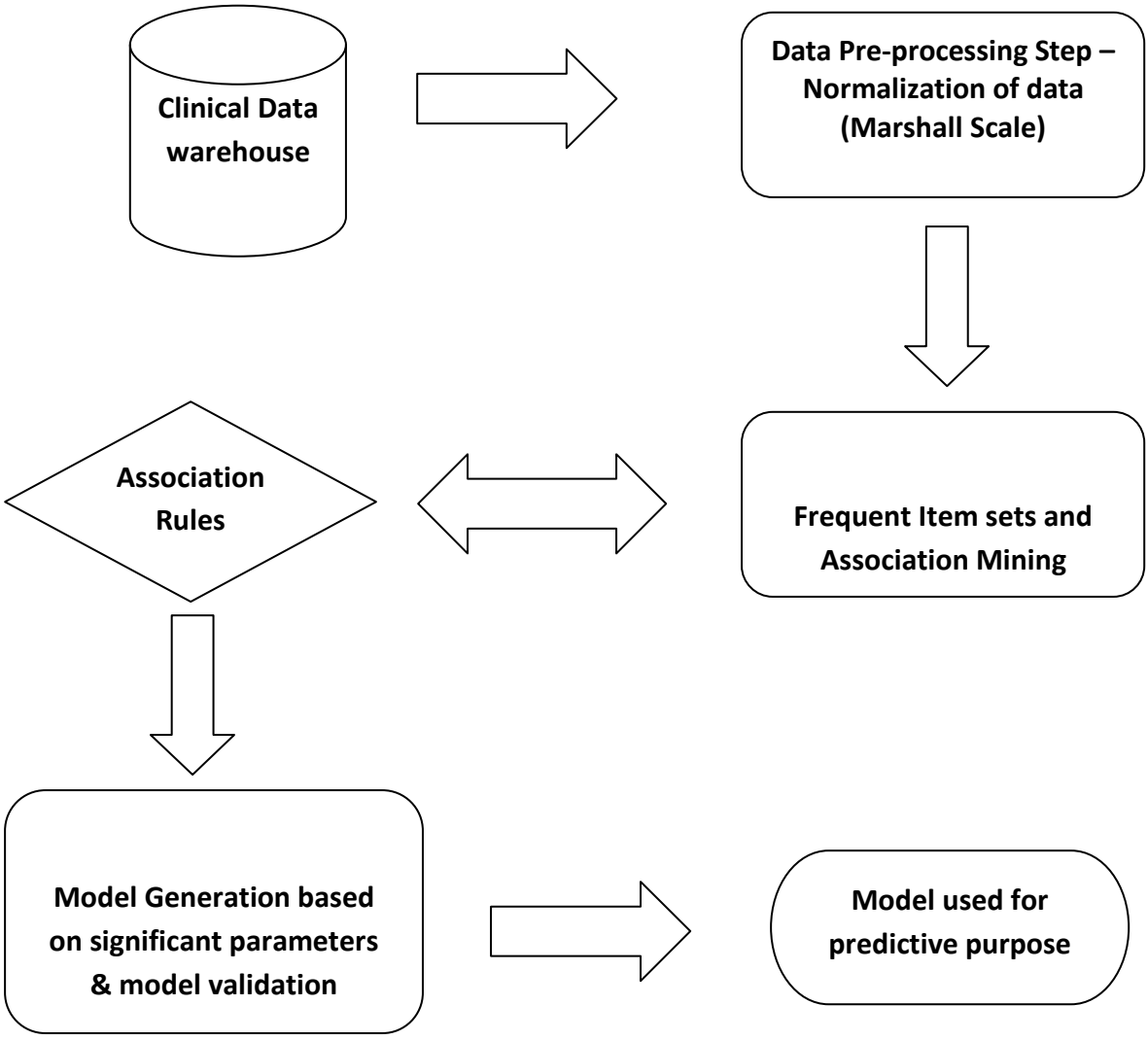
The characteristics of clinical data as it originates during the process of clinical documentation includes issues of data availability and complex representation models that can make data mining process challenging. Therefore, data pre-processing and transformation are required before one can apply data mining algorithms on clinical data. The stored data in the warehouse would provide a basis for the analysis of risk factors of the disease. For example, we can compare tumor with non-tumor patients to find patterns associated with the occurrence of brain tumor. This method has been common practice in evidence-based medicine, which is an approach where clinician is aware of the evidence in support of clinical practice and its associated strength [22]. 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 [22]. Rules are a type of the most human-understandable knowledge and therefore it is most suitable for deciphering new rules corresponding to data associated with medical applications. Association rule mining is a general purpose rule observation scheme that has been widely used for observing rules in medical applications [23]. Association rule mining using objective measures and transitive inference for pruning has also been done in the clinical domain to find associations between medications and clinical problems using electronic health record data [24]. Studies made by Brossette *et al.* [22] & Ordonez *et al.* [25], states about associative rules corresponding to hepatitis & heart.

The objective of this study was to analyze data stored in the clinical warehouse for patients of brain tumor (primary stage) and observe associative rules based on clinical diagnostic parameters. The data in the warehouse is subjected to association mining for observing new rules. 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.

Material & 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 which can be associated with the 'STATE' of brain tumor by applying association rule mining 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).

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. As varied dimensionality was observed in the data, on consultation with oncologists appropriate data forms were selected. 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. Pre-processing of data in the warehouse was done using STATISTICA DATAMINER 9.1 [26], 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 [27] and (4) Feature selection.

STATISTICA DATAMINER 9.1 [26] was used to calculate the frequency of each item set with support % criteria of atleast 30 along with head and body iteration rate of 10. All the frequent item set obtained with atleast 30% support criteria were subjected for the observation of association rules. STATE was declared as the response indicator and the remaining parameters were defined as categorical indicators. The confidence to deduce rule was set to atleast 85% and the process was executed with antecedent and precedent iteration rate of value 10.

The parameters found to be associated with occurrence of tumor were selected to build a predictive model using normalized regression approach. Jackknifing was applied for cross-validation of the predictive model along with accuracy, sensitivity and specificity analysis.

Result & Discussion

Haemoglobin_content, TLC (Total Leucocytes Count), Platelet Count, KFT_Creatinine, KFT_BUN (Blood Urea Nitorgen), LFT_Sr_Bilirubin (Serum Bilirubin), LFT_ALP (Alkaline Phosphatase), LFT_SGOT (Serum Glutamic-oxaloacetic Transminase) and LFT_SGPT (Serum Glutamate Pyruvate Transaminase) are the parameters that showed support of atleast 30%. Item sets satisfying the support % were subjected to observation of association rules within the specified mining criteria that showcased association of high values of Creatinine, BUN, SGOT & SGPT with presence of tumor in patients. Table I (A & B) enlists various Association Rules discovered within the defined criteria.

Association Rule	Support %	Confidence %	Correlation %
KFT_Creatinine = HIGH ==> KFT_BUN = HIGH	56.75	100	77.45
KFT_Creatinine = HIGH ==> STATE = 1	56.75	100	77.77
KFT_BUN = HIGH ==> STATE = 1	78.37	85.29	90.8
KFT_Creatinine = HIGH, KFT_BUN = HIGH ==> STATE = 1	56.75	100	79.77
LFT_SGOT = HIGH ==> STATE = 1	62.16	98.83	81.72
LFT_SGOT = HIGH, LFT_SGPT = HIGH ==> STATE = 1	62.16	95.83	85.71
LFT_SGPT = HIGH ==> STATE = 1	81.08	88.23	89.56
Haemoglobin_content = NORMAL ==> STATE = 1	59.45	100	81.64

Table I (A) - Association Rules deciphered for clinical parameters corresponding to occurrence of brain tumor.

Association Rule	Support %	Confidence %	Correlation %
KFT_Creatinine = HIGH ==> STATE = 0	6.75	100	77.77
KFT_BUN = HIGH ==> STATE = 0	8.7	100	90.8
KFT_Creatinine = NORMAL, KFT_BUN = NORMAL ==> STATE = 0	96.85	100	79.77
LFT_SGOT = HIGH ==> STATE = 0	2.63	98.83	81.72
LFT_SGOT = NORMAL ==> LFT_SGPT = NORMAL, STATE = 0	92.26	100	85.71
LFT_SGPT = HIGH ==> STATE = 0	11.08	98.32	89.56

Table I (B) - Association Rules deciphered for clinical parameters corresponding to non-occurrence of brain tumor.

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 the 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 are 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 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.

Based on the parameters identified among the associative rules with 85% (Creatinine, BUN, SGOT, SGPT) & 75% confidence (Hemoglobin Content, Alkaline Phosphatase and Serum Biliuribin), a predictive model is developed to predict the possible STATE of a individual i.e whether suffering from tumor or not. The model was generated using the normalized regression approach, given by equation (i).

$$\Theta_j := \Theta_j - \alpha \, |\partial/\partial(\Theta_j)| J(\Theta) \quad \text{... equation (i)}$$

In normalized regression approach we try to obtain the minimal set of coefficients (Θ_j) for the independent parameters by varying the learning rate (α). 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 Θ_j. Convergence (steepest decent approach) was observed at $\alpha = 0.04$.

Most significant model obtained is:

STATE = 0.171 + 0.0491 Haemoglobin_content + 0.0652 KFT_Creatinine + 0.0171 KFT_BUN - 0.0504 LFT_Sr_Bilirubin + 0.0304 LFT_ALP - 0.07 LFT_SGPT+ 0.0806 LFT_SGOT

The cross-validation results obtained from Jackknifing: $R^2_{(calculated)}$ = 74.66% and PRESS = 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.

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 other brain related diseases.

CHAPTER 3 - SN algorithm: Analysis of temporal form of clinical data for mining periodic patterns and impending augury

An interesting prospective in mining of heterogeneous clinical data would be an approach that could analyse the temporal form. The discovery of hidden periodic patterns in temporal data, apart from unveiling important information can facilitate data management substantially [28]. However, very limited work has been done so far on data mining of temporal data, which demonstrates generalization of pattern mining in time-series data [29]. For instance, we can model the change of climatic conditions in a spatial region as a sequence of existing or a past set of values. Periodicity has only been studied in the context of temporal analysis of time-series based databases that addressing the following problem: given a long sequence S and a period T, the aim is to discover the most representative trend that repeats itself in S every T time stamps [30]. This uses a tree structure to count the support of multiple patterns at two database points and comparatively studies the problem of finding sets of events that appear together periodically [31]. However, it does not take into consideration the order of occurrence of events. Whereas, in case of temporal clinical data it is necessary to consider specific order of occurrence of events that are associated with the state of a disease.

In this study we have defined the temporal mining problem of clinical data in terms of (a) discovery of clinical parameters that can be associated with a specific disease; and (b) an algorithm for traversing the clinical parameters of temporal points (T_0, T_1 ... T_n) in order of their occurrences, along with mapping the values observed for each point with the previous one. This helps in auguring the state of a specific disease at point T_n whose result is unknown. To predict the state of a disease at point T_n, we propose a new algorithm (SN algorithm) based on Jacobian transformation by considering different temporal points in which Jacobian of selected clinical parameters are associated with the state of that disease. Hence, derivatives (J_0, J_1...) of temporal points (T_0, T_1 ...) along with respective states (S_0, S_1....) are mapped with a future point (T_n) Jacobian (J_n) and finally its determinant (J") is calculated to obtain a possible state (S_n).

Material & Methods

The clinical parameters are identified to be associated with a disease based on the associative rules deciphered. The selected clinical parameters act as base point for SN algorithm, which consists of the following four steps.

i. With an input of set of temporal points (T_0, T_1, T_2,....,T_n), a set of selected clinical parameter values (P_0, P_1, P_2,..., P_n) for a patient along with the state of disease (S_0, S_1, S_2,...,S_n) is chosen for each temporal point, where State 'S_n' is unknown for the point T_n.

ii. Jacobian transformation is applied over the set of selected parameters (P_0, P_1, P_2,..., P_n) for each of the temporal point 'T' to obtain the Jacobian.

iii. Jacobian (J_0, J_1, J_2, ..., J_n) for each temporal point along with state of disease 'S' is then mapped to the values of other temporal point.

iv. Jacobian determinant (J'') is then determined based on the mapping done in step iii for predicting augury of state S_n for point T_n.

Mathematically, Jacobian, mapping of Jacobian in time-space as area and estimation of its determinant for area can be explained as follows [32].

Let T (u, v) be a smooth coordinate transformation with Jacobian J (u,v) and let R be the rectangle spanned by du = (du, 0) and dv = (0, dv). If du and dv are sufficiently close to 0, then T (R) is approximately the same as the parallelogram spanned by equation (ii):

$$dx = J(u, v)\, du = (x_u du, y_u du, 0) \quad \text{... equation (ii)}$$

$$dy = J(u, v)\, dv = (x_v dv, y_v dv, 0) \quad \text{... equation (ii)}$$

Let dA denote the area of the parallelogram spanned by dx and dy parameter, then dA approximates the area of T (R) for du and dv sufficiently close to 0.

The cross product of dx & dy is given as in equation (iii),

$$dx * dy = \langle 0,0 \begin{vmatrix} xu & xv \\ yu & yv \end{vmatrix} \rangle dudv \quad \text{... equation (iii)}$$

from which the differential area *dA* can be obtained as in equation (iv):

$$dA = \left|\frac{\partial(x, y)}{\partial(u, v)}\right| dudv \quad \text{... equation (iv)}$$

Area of a small region in the uv-plane is scaled by Jacobian determinant to approximate areas of small images in the xy-plane (figure 6).

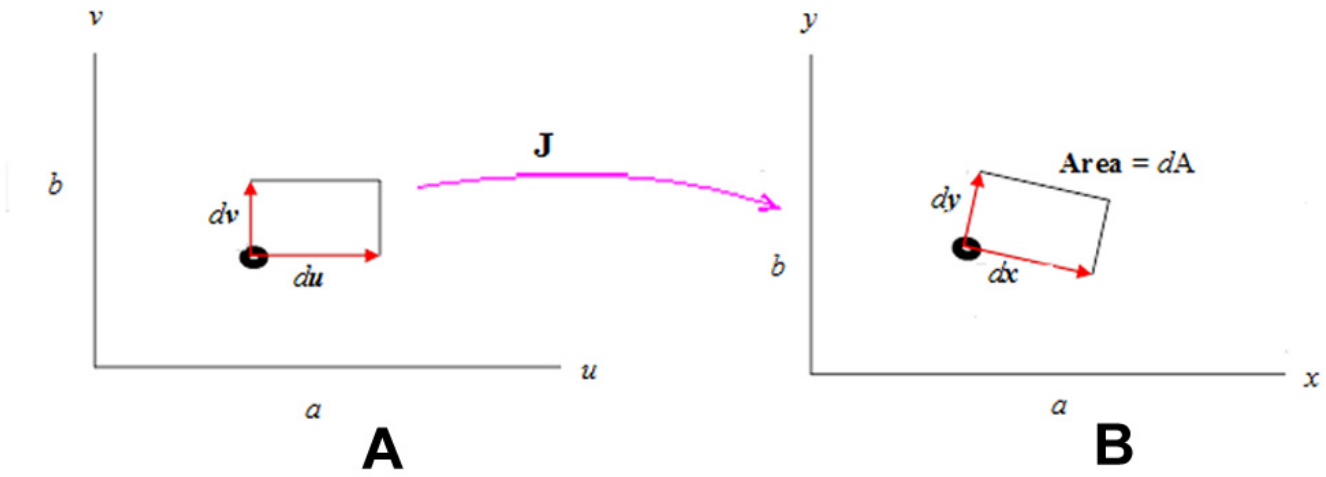
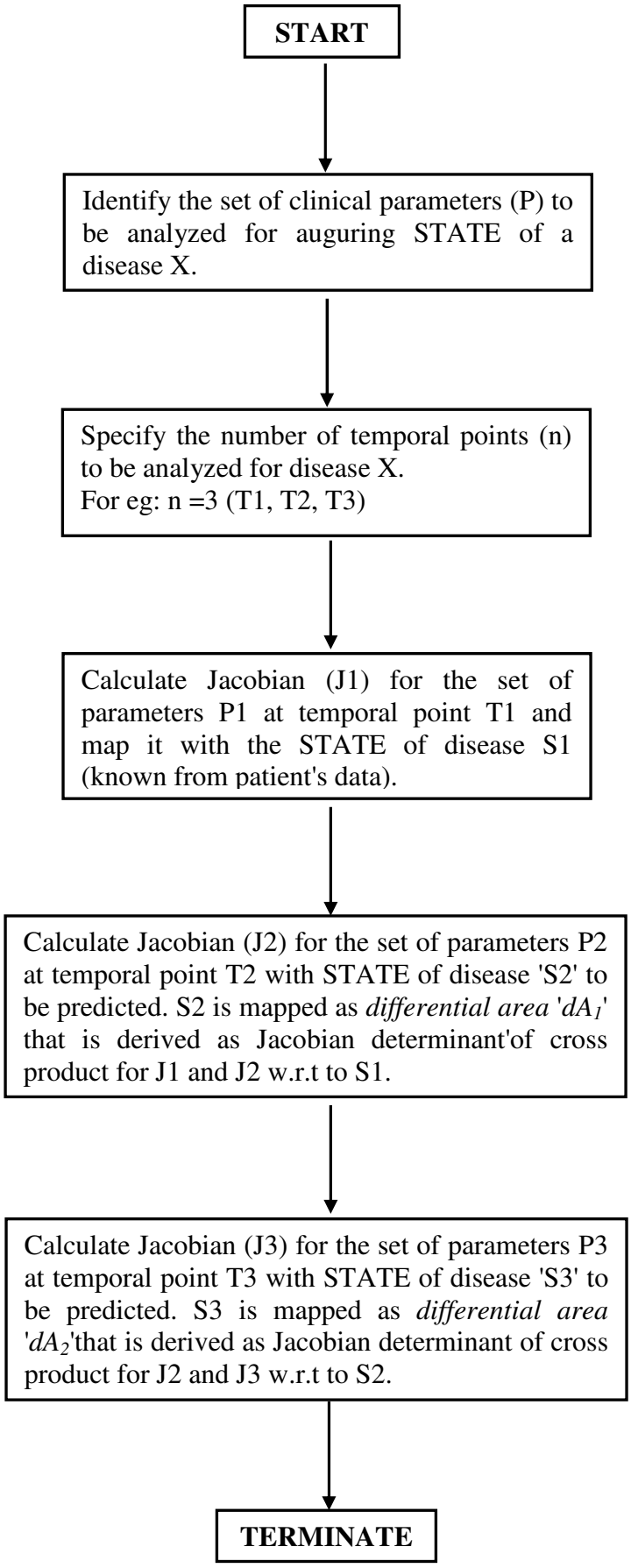


Figure 6 - Area differential approach based on Jacobian Transformation

(A - Temporal point 1; B - Temporal point 2)

The flow diagram 2 depicts the methodology of SN algorithm in a logical representation.



Flow diagram 2 - Demonstrating **v**arious steps in SN algorithm.

Result & Discussion

A case study of 55 patients suffering from brain tumor over a period of 6 months was used to evaluate temporal mining technique proposed in this study. In the first phase of this study, based on set of rules deciphered from association mining study (Chapter 2): creatinine, blood urea nitrogen (BUN), SGOT and SGPT were the parameters (P) selected to be associated with the state of the disease (S).

In the second phase of the study, SN algorithm was applied over 3 temporal state points T_0 (T1, T2, T3) for each patient P_o(P1, P2,...P55) in which state S_o(S1, S2, S3) of the disease at each temporal point was considered along with the values for Creatinine c(c1, c2, c3), BUN b(b1, b2, b3), SGOT s(s1, s2, s3) and SGPT g(g1, g2, g3) parameters as depicted in Table II.

P1, T1, c1, b1, s1, g1	P2, T1, c'1, b'1, s'1, g'1	…	P55, T1, c''1, b''1, s''1, g''1
P1, T2, c1, b1, s1, g1	P2, T2, c'2, b'2, s'2, g'2	…	P55, T2, c''2, b''2, s''2, g''2
P1, T3, c1, b1, s1, g1	P2, T3, c'3, b'3, s'3, g'3	…	P55, T3, c''3, b''3, s''3, g''3

Table II - Temporal points along with various selected clinical parameters corresponding to brain tumor.

L (c, b, s, g) is the transformation with Jacobian J (c, b, s, g) applied for each state S_o (S1, S2, S3). Jacobian is calculated for each of the functional parameter (c,b,s,g) of the first temporal point T1 which is mapped with the state S1 as area curve. Similarly for the second temporal point Jacobian is mapped to S2. Based on the cross product of Jacobian for point T1 and T2, the differential area was mapped as Jacobian determinant to obtain S2 state. For S2 predicted state there was a perfect 100% accuracy (n = 55; No. of True Positives = 55, No. of False Negatives = 0) corresponding to the state of brain tumor. However, for the third temporal point only Jacobian for the parameters was obtained and S3 result was in a hidden state. To obtain the S3 state, differential area was mapped as Jacobian determinant based on cross products of Jacobian for points T2 and T3. Validation of the S3 predicted state values with the hidden state showed 92.7% accuracy (n = 55; No. of True Positives = 51, No. of False Negatives = 04). A very high prediction accuracy of ~97% is achieved for a brain tumor state 'S_n' for any temporal point 'T_n' provided.

Taken together, the algorithm developed in this study hold a great potential in monitoring the state of disease based on regular input values for minimal set of clinical parameters. The effectiveness of the algorithm needs to be further evaluated by analyzing the parameters associated with other diseases and analyzing it over various temporal points for a group of patients.

CHAPTER 4 - Identification of key measures for evaluation of cognitive performance at high altitude

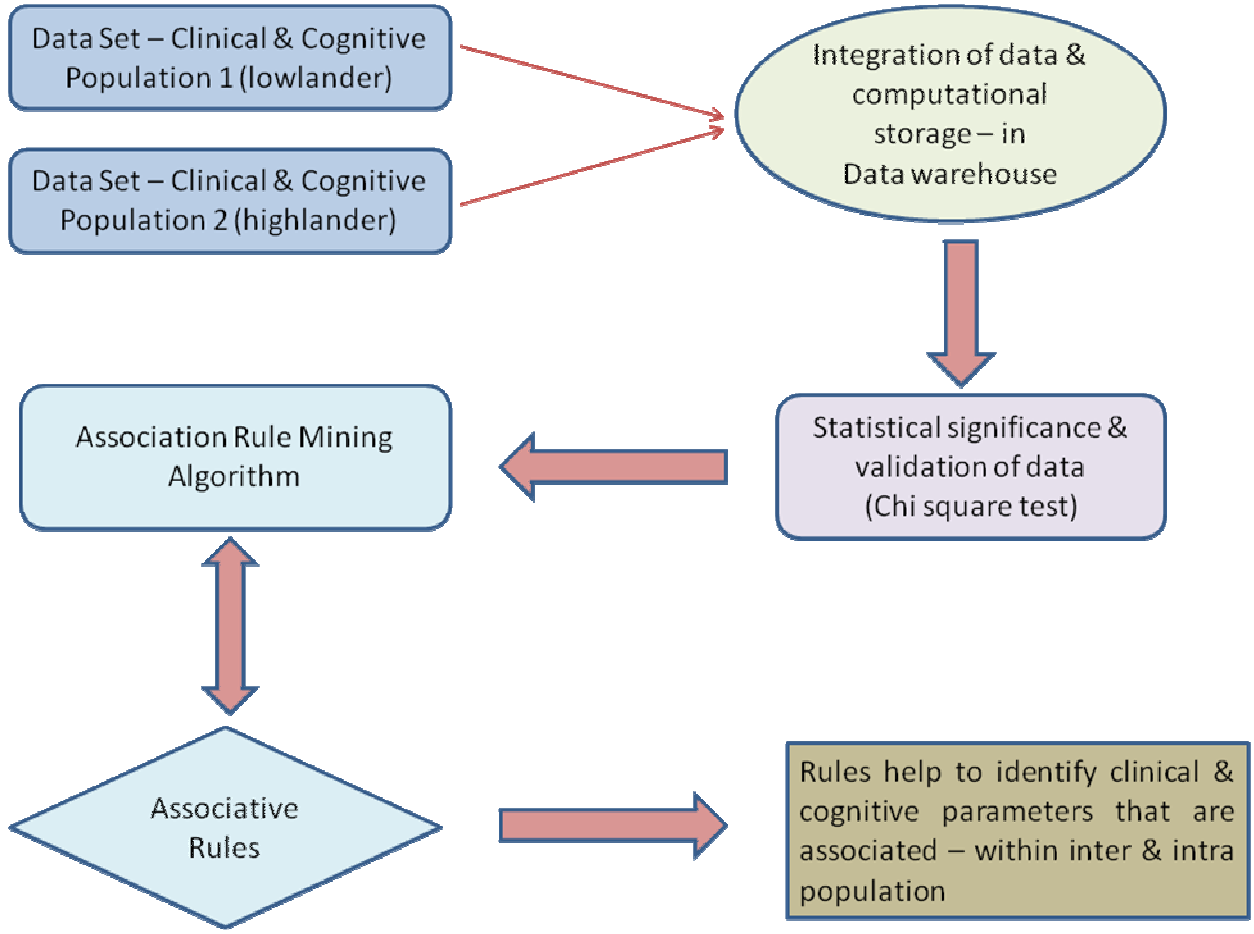
As altitude increase there is scarcity of oxygen which may lead to cognitive impairment and other clinical disorders. There have been reports on sensory discrimination & evaluation impairment in short term memory following ascent to high altitude [33]. Study on group of human volunteers residing at high altitude revealed impairment in verbal working memory which could be associated with chronic hypoxia exposure [34]. There also have been study to investigate percentage prevalence and extent of mild cognitive impairment (MCI) in a young and healthy population staying at altitudes above 4,300 m in the trans-Himalayan regions for durations longer than 12 months based on multi-domain cognitive screening test (MDCST) for demographic studies in remote locations at high altitude [35]. MDCST, is comprehensive in that it covers nine domain assessments (Orientation, Memory Registration, Visuospatial Executive, Object Recognition, Attention, Recall, Coordination & Learning, Language and Procedural Memory) and has a better specificity with sensitivity in comparison to other screening instruments. Thus, it provides an opportunity to increase the scope of cognitive assessment to domains like procedural memory, mind-body co-ordination, attention and learning of complex tasks through improvised and customized psychometric tests. Further, analysis of Beck Depression Inventory (BDI) with insomnia can help in identification of hitherto undetected depression [36]. However, there has not been any demographic study to measure the prevalence of MCI that can be associated along with Beck Depression Inventory (BDI), insomnia, and clinical parameters such as, blood glucose level, blood pressure, blood cholesterol, kidney and liver function. A rule based study can allow us to help in analyzing and identifying risk factors specific to lowlanders (<=350m) as well as highlanders (>=1500m) for their cognitive decline at higher altitude (>=4300m).

In summary, this study attempts to examine cognitive parameters identified by MDCST along with clinical measures associated with cognitive decline, BDI and insomnia among the populations at high altitudes (>=4300m) for prolonged duration. Comparison groups in this study were highlanders (1500m and above) and lowlanders (less then equal to 350m) residing for at least 2 years at altitude >=4300m. We suggest clinical and cognitive parameters that should be evaluated among the lowlander and highlander population for evaluation of their respective cognitive performance.

Material & Methods

This study is based on the research conducted by High altitude Physiology division of Defence Institute of High Altitude Research, Leh, India, during August 2009 - January 2012. The approach used for this study has been demonstrated in flow diagram 3. The tests were performed on group of

volunteers in age group of 25-40 comprising of lowlanders and highlanders located at higher altitude (>4300m) for more than a year. MDCST screening along with measurement of clinical diagnostic parameters were performed for all the volunteers. The set of clinical parameters selected for the study focuses on Lipid profiling, KFT (Kidney Functionality Test) result, LFT (Liver Functionality Test) result, sugar level, triplets of blood pressure and pulse rate observed. Beside, a medical questionnaire comprising questions related to occurrence of chronic diseases, physical and physiological ailments, heart problems, stroke, epilepsy, head injury, drug abuse, psychological disorders and general health status was administered to all the volunteers [35]. Core behavioural measures (CBM) like core alcohol consumption, core tobacco use, core diet and core physical activity [35] were also applied to all subjects in accordance with WHO guidelines [37]. The tests were administered by field investigators under the supervision of a clinical psychiatrist.



Flow Diagram 3 - Representation of knowledge discovery process (identification of key evaluation parameters for MCI)

Subjects with underlying heart disease, chest pain, stroke/infarction/cerebral haemorrhage, renal failure, diabetes, viral hepatitis, chronic disease and gastroesophageal reflux disease (GERD) were excluded. Subjects with previous neurologic/psychiatric symptoms, major surgery and

familial disorders were also excluded. This ensured inclusion of only healthy subjects in this study. Data of 200 volunteers, 100 each for lowlander and highlander qualified for the study.

The MDCST comprises cognitive domains involved in Orientation, Memory Registration, Visuospatial Executive, Object Recognition, Attention, Recall, Coordination & Learning, Language and Procedural Memory [35]. Each of the parameter is assigned weightage of 5. The cumulative score <=34 indicates onset or presence of cognitive impairment. For the subjects who qualified inclusion criterion, the Insomnia and Beck Depression Inventory (BDI) [36] was applied to assess activities of daily living and to investigate the presence of hitherto undetected depression. Sleep is measured in terms of 0/1, with 1 indicating insomnia and BDI value >= 7 indicates abnormality. All the subjects were labelled as normal/abnormal based on the given scale parameters.

Following clinical parameters were measured for all the volunteers under the supervision of registered medical practitioner - Blood Glucose (BS), systolic blood pressure (SBP), diastolic blood pressure (DBP), pulse rate (PR), Blood Urea Nitrogen (BUN), Creatinine - serum, Serum Glutamic Oxaloacetic Transaminase (SGOT), Serum Pyruvic Transaminase (SGPT), Total Cholesterol (TC), Triglycerides (TGL), High-density Lipoprotein (HDL), Low-density Lipoprotein (LDL), Very Low-density Lipoprotein (VLDL), TC/HDL cholesterol ratio, LDL/HDL ratio, Homocysteine, Vitamin B-12 & Folic acid and were categorized as low, normal or high as per the prescribed range [26].

In this study, the dataset is specifically subjected for chi-square test as statistical significance of it implies that the differences are not due to chance alone, but instead may be indicative of other processes at work. The test was performed on the lowlander & highlander population dataset respectively to observe the significance among the cognitive screening (MDCST, BDI & Sleep) and clinical parameters [26].

STATISTICA DATAMINER 9.1 [26] was used to calculate the frequency of each item set with support % criteria of atleast 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. All the frequent item set obtained with atleast 30% support criteria were subjected for the discovery of association rules. Total cognitive score, BDI score & Sleep were declared as the response indicator and the other parameters were defined to be categorical indicators. The confidence to deduce rule was set to atleast 50% and the process was executed with antecedent and precedent iteration rate of value 10.

Result & Discussion

Statistically significant parameters in lowlanders among MDCST associated domains of cognitive impairment are Visuospatial Executive, Attention and Coordination & Learning; whereas

Object recognition is significantly impaired corresponding to BDI. The Orientation impairment parameter was significantly associated with insomnia with an observed p-value <= 0.05. In contrast, the test performed for the dataset of highlander suggest that Procedural Memory, Coordination and Learning, Visuospatial Executive, Recall, Language parameters are statistically significant MDCST associated domains with p-value <= 0.05; whereas no parameter is significant corresponding to BDI and insomnia.

Among the clinical measures, the chi-square test reveals that there is not a significant parameter for lowlanders corresponding to either total cognitive score, BDI or insomnia. The significant parameter for highlanders was vitamin-B12 levels (p-value = 0.0099) that corresponded to total cognitive score, total cholesterol level (p-value = 0.0094) that corresponded to BDI, and folic acid level (p-value = 0.0023) that corresponded to insomnia.

Table III enlists rules discovered within the defined criteria for lowlander population. Item sets satisfying the support-percentage were subjected to discovery of association rules within the specified mining criteria that showed association of high Vitamin B-12, HDL, BUN and Folic Acid levels for MDCST, BDI, and insomnia. Also alcoholic population showed rules for abnormal MDCST, which indicate its association to cognitive impairment.

Association Rule	**Support %**	**Confidence %**	**Correlation %**
alcoholic/non-alcoholic == NA ==> BDI == Normal_BDI	40.91	54.55	68.38
alcoholic/non-alcoholic == A ==> Abnormal_MDCST	31.82	60.87	62.24
BDI == Normal_BDI ==> sleep == Normal_Sleep	47.73	87.50	74.62
Normal_MDCST ==> BDI == Normal_BDI	36.36	48.48	59.38
Vit-B12 == High_VitB12, HDL == High_HDL ==> Abnormal_MDCST	50.00	100	74.16
Blood Urea Nitrogen == High_BUN, Vit-B12 == High_VitB12, Vit-B12 == High_VitB12 ==> Abnormal_Sleep	52.27	100	76.79
Folic Acid == High_FA, Blood Urea Nitrogen == High_BUN ==> Abnormal_BDI	61.36	64.28	78.73
VLDL == Normal_VLDL==>Normal_BDI	61.36	64.28	80.17

Table III - Association Rules deciphered for lowlanders

For highlanders population, item sets that satisfied the support-percentage were subjected to discovery of association rules within specified mining criteria, showcased association of high values of Vitamin B-12, HDL, VLDL, BUN, Cholesterol, Triglycerides and Folic Acid for MDCST, BDI, and insomnia. Also, the alcoholic population showed rule associated to sleep. Table IV enlists the association rules discovered within the defined criteria for highlander population.

Association Rule	**Support %**	**Confidence %**	**Correlation %**
Normal_MDCST ==> BDI == Normal_BDI	41.30	76.00	66.15
Alcoholic/Non alcoholic == A ==> sleep == Normal_Sleep	43.48	58.82	67.27
sleep == Normal_Sleep ==> BDI == Normal_BDI	47.83	88.00	75.46
sleep == Normal_Sleep ==> Normal_MDCST	50.00	69.70	68.66
HDL == High_HDL, VLDL == High_VLDL, Blood Urea Nitrogen == High_BUN, Vit-B12 == High_VitB12 ==> Abnormal_MDCST	50.00	58.97	72.22
Total_Cholesterol == High_TC ==> Abnormal_BDI	52.17	96.00	73.19
Creatinine == High_Creatinine, Triglycerides == High_TGL, HDL == High_HDL, Folic_Acid == High_FA, Vit-B12 == High_VitB12 ==> Abnormal_Sleep	50.00	58.97	72.22

Table IV - Association Rules deciphered for highlanders

Our results show that different set of key MDCST domains and clinical measures need monitoring for the analysis of MCI and undetected depression (BDI and insomnia) for lowlander and highlander populations. The MDCST considers 9 different domains for analyzing cognitive performance of an individual. However, the statistical significance data from our observations suggest that Visuospatial Executive, Attention, Coordination & Learning, Object recognition and Orientation are the major domains among MDCST that needs to be regularly monitored for lowlanders. Whereas Procedural Memory, Coordination and Learning, Visuospatial Executive, Recall, Language are the key domains for highlanders. The lowlanders showed a higher rate of cognitive impairment and insomnia compared to highlanders at an altitude of 4300m or more. There are no significant clinical measures observed from goodness of fit test for lowlanders;

whereas vitamin-B12, total cholesterol and folic acid levels were found to be significantly associated with highlander's cognitive performance.

Rules deduced from association mining suggest that the alcoholic lowlander population showed low cognitive response for cumulative MDCST score with 60.87% of confidence. Also high levels of vitamin-B12 and HDL were associated with cognitive impairment (100% confidence), high vitamin-B12 and BUN were associated with insomnia (100% confidence), and high levels of folic acid were associated with BDI (64.28%). Associative rules deduced for highlander population were significantly different compared to lowlander population. An alcoholic highlander population did not represent any significant rule under the mined criteria corresponding to MDCST, but an important statistical observation suggests that they had normal sleep with 88% confidence. In highlander population, high level of HDL, VLDL, BUN and Vitamin B-12 was found to be associated with cognitive impairment (58.97% confidence), while high level of total cholesterol was associated to BDI (96%) and high level of creatinine, HDL, Vitamin B-12 and Folic Acid to the insomnia (58.97%).

CONCLUSION

- For the data storage problem lying in façade of clinical informatics, a clinical dimensional model is being proposed which can be used for development of a clinical data mart. The model has been designed keeping in consideration temporal storage of patient's data with respect to all possible clinical parameters which can include both textual and image based data. Availability of said data for each patient can be then used for application of data mining techniques for finding the correlation of all the parameters at the level of individual and population.
- Case study of association mining for brain tumor patients from the warehouse suggests - high values of Creatinine, Blood Urea Nitrogen (BUN), SGOT & SGPT to be directly associated with tumor occurrence for patients in the primary stage with atleast 85% confidence and more than 50% support. A normalized regression model is proposed based on these parameters along with Haemoglobin content, Alkaline Phosphatase and Serum Bilirubin for prediction of occurrence of STATE (brain tumor) as 0 (absent) or 1 (present).
- A new "SN" algorithm is being proposed, to map clinical parameters found to be associated with a disease, to its state at various temporal points. It is based on Jacobian's approach, which augurs the state of a disease 'S_n' at a given temporal point 'T_n' by mapping the derivatives with the temporal point 'T_0', whose state of disease 'S_0' is known, for estimating its Jacobian determinant. The proposed algorithm has been applied in a case study on a temporal clinical data set of brain tumor patients. A very high prediction accuracy of ~ 100% is been observed for a brain tumor state 'S_n' for any temporal point 'T_n' provided. However the algorithm needs to be further validated on other clinical data sets.
- Cross-sectional study was performed to identify cognitive analyzers among cognitive screening and clinical parameters for the low (<=350m) & highlander (>=1500m) population staying at higher altitude (>4300m) for prolonged duration, that can be associated with cognitive impairment, beck depression inventory and sleep abnormality. Visuospatial Executive, Attention, Coordination & Learning, Object recognition, Procedural Memory, Recall, Language for lowlander population while Procedural Memory, Coordination and Learning, Visuospatial Executive, Recall, Language for highlander population respectively, are the key MDCST parameters identified for analyzing the cognitive performance with observed p-value <= 0.05. For low & highlanders respectively, different set of cognitive performance based associative rules are observed atleast with 30% support and more than 60% confidence for behavioural and clinical measures.

REFERENCES

1. Wang X, Liotta L. Clinical bioinformatics: a new emerging science. J Clin Bioinforma. 2011;1(1):1.
2. Schwarz E, Leweke FM, Bahn S, Liò P. Clinical bioinformatics for complex disorders: a schizophrenia case study. BMC Bioinforma. 2009;10(12):S6.
3. Berger AM, Berger CR. Data mining as a tool for research and knowledge development in nursing. Comput Inform Nurs. 2004;22(3):123-31
4. Brazhnik O, Jones JF. Anatomy of Data Integration. J Biomed Inform. 2007;40(3):252-69.
5. Dinu V, Nadkarni P. Guidelines for the Effective Use of Entity–Attribute–Value Modeling for Biomedical Databases. Int J Med Inform. 2007;76(11-12):769-79.
6. Atherton J. Development of the Electronic Health Record. Amer Med Asso J of Eth. 2011, 13(3)**:**186–189.
7. Evans RS, Lloyd JF, Pierce LA. Clinical use of an enterprise data warehouse. AMIA Annu Symp Proc. 2012;189-98.
8. Houston AL, Chen H, Hubbard SM, Schatz BR, Tobund NG, et al. Medical Data Mining on the Internet: Research on a Cancer Information System. Artif Intell Rev. 1999;13: 437-466.
9. Holmes AB, Hawson A, Liu F, Friedman C, Khiabanian H, et al. Discovering disease associations by integrating electronic clinical data and medical literature. PLOS One 2011;6(6).
10. Brossette SE, Sprague AP, Hardin JM, Jones KWT, Moser SA. Association rules and data mining in hospital infection control land public health surveillance. J Am Med Inform Assoc. 1998;5:373—81.
11. Sarbadhikari SN. The state of medical informatics in India: a roadmap for optimal organization. J Med Syst. 2005; 29(2):125-141.
12. Paul PK, Kumar A, & Chatterjee D. Health Informatics and its practice: Emerging Domain of Information Science - Indian Scenario. Current Trends in Biotechnology and Chemical Research 2012;2(2).
13. Health Level Seven [Internet]. [Accessed on Jan. 11, 2012]. Available from: http://*www.**hl7india**.org*
14. Indian Association for Medical Informatics [Internet]. [Accessed on Jan. 11, 2012]. Available from: http://*www.**iami**.org.in*
15. Atalag K. Archetype based domain modeling for health information systems. Graduate School of Informatics, The Middle East Technical University, Turkey, 2007.

16. Kimball R, & Ross M. The Data Warehouse Toolkit, 2nd edition 2002. John Wiley & Sons, Inc., New York.
17. CA, Tech. (2012) ERWIN DATA MODELER (data modeling software system), version 8.2 2012. http://erwin.com/products/data-modeler.
18. Mozes A. Mining Star Schemas A Telco Churn Case Study [Internet]. 2011 [Accessed on 2012 Oct 31]. Available from: http://www.oracle.com/technetwork/database/options/advancedanalytics/odm/odmtelcowhitepaper-326595.pdf.
19. Kimball R. Slowly Changing Dimensions, Types 2 and 3. Kimball Group [Internet]. 2009 [Accessed on 2012 Oct 31]. Available from: http://www.kimballgroup.com/2008/09/22/slowly-changing-dimensions-part-2/
20. Browning D, & Mundy J. Data Warehouse Design Considerations [Internet]. 2001 [Accessed on 2012 Oct 31]. Available from: Microsoft http://msdn.microsoft.com/enus/library/aa902672%28v=sql.80%29.aspx .
21. Lunexa. White Paper on Slowly Changing Dimensions [Internet]. 2011 [Accessed on 2012 Oct 31]. Available from: http://www.lunexa.com/documents/10923/12405/SlowlyChangingDimensions.pdf.
22. Li J, Fu AW, Fahey P. Efficient discovery of risk patterns in medical data. Artif Intell Med. 2009;45(1):77-89.
23. Brossette SE, Sprague AP, Hardin JM, Jones KWT, Moser SA. Association rules and data mining in hospital infection control land public health surveillance. J Am Med Inform Assoc. 1998;5:373—81.
24. Wright A, Chen ES, Maloney FL. An automated technique for identifying associations between medications, laboratory results and problems. J Biomed Inform. 2010;43(6):891-901.
25. Ordonez C, Ezquerra NF, Santana CA. Constraining and summarizing association rules in medical data. Knowledge and Information Systems 2006;3: 1-2.
26. StatSoft, Inc. (2010). STATISTICA (data analysis software system), version 9.1; www.statsoft.com.
27. Marshall WJ. Clinical Biochemistry: Metabolic and Clinical Aspects 2008. Edinburgh: Churchill Livingstone/Elsevier.
28. Mamoulis N Cao H, Kollios G, Hadjieleftheriou, Tao Y, et al. Mining, Indexing, and Querying Historical Temporal Data. Knowledge discovery and data mining 2004;236-245.

29. Peng WC, Chen MS. Developing data allocation schemes by incremental mining of user moving patterns in a mobile computing system. IEEE Transactions on Knowledge and Data Engineering 2003;15(1):70–85.
30. Indyk P, Koudas N, Muthukrishnan. Identifying representative trends in massive time series data sets using sketches. In Proc. of Very Large Data Bases. *Cairo*, Egypt: 26th VLDB Conference 2000;363-372.
31. Ma S & Hellerstein JL. Mining partially periodic event patterns with unknown periods. In Proc. of International Conference on Data Engineering. Heidelberg: Data Engineering; 2001;205-214.
32. Kinsley J. Multivariable Calculus Online adapted from Calculus: A Modern Approach [Internet]. 2001 [Accessed 26 January 2013]. Available from: http://math.etsu.edu/multicalc/prealpha/.
33. Singh SB, Thakur L, Anand JP, Panjwani U, Yadav D, Selvamurthy W. Effect of high altitude (HA) on event related brain potentials. Indian J Physiol Pharmacol 2003; 47:52-8.
34. Yan X, Zhang J, Gong Q, Weng X. Prolonged high-altitude residence impacts verbal working memory: an fMRI study. Exp Brain Res 2011; 208:437-45.
35. Hota SK, Sharma VK, Hota K, Das S, Dhar P, Mahapatra BB, *et al.* Multi-domain cognitive screening test for neuropsychological assessment for cognitive decline in acclimatized lowlanders staying at high altitude. Indian J Med Res 2012; 136:411-20.
36. World Health Organization, Switzerland. WHO STEPS Surveillance [Internet]. 2008 [Accessed 15 Jan 2013]. Available from: www.who.int/chp/steps.
37. Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. Arch Gen Psychiatry 1961; 4:561-71.

ABBREVIATIONS

ALP - Alkaline Phosphatase

BDI - Beck Depression Inventory

BS - Blood Glucose

BUN - Blood Urea Nitrogen

CBM - Core Behavioural Measures

CT - Computed Tomography

DBP - Diastolic Blood Pressure

DIM - Dimension

EHR/EMR - Electronic Health/Medical Record

ETL - Extraction, Transformation & Loading

GERD - Gastroesophageal Reflux Disease

HDL - High Density Lipoprotein

HL7 - Health Level 7

IAMI - Indian Association for Medical Informatics

IT - Information Technology

KFT - Kidney Functionality Test

LDL - Low-density Lipoprotein

LFT - Liver Functionality Test

MCI - Mild Cognitive Impairment

MDCST - Multi Domain Cognitive Screening Test

MRI - Magnetic Resonance Imaging

NIH - National Institute of Health

PR - Pulse Rate

PRESS - Predicted Residual Sum of Squares

RDBMS - Relational Database Management System

SBP - Systolic Blood Pressure

SCD - Slowly Changing Dimension

SGOT - Serum Glutamic-oxaloacetic Transminase

SGPT - Serum Glutamate Pyruvate Transaminase

TC - Total Cholesterol

TGL - Triglycerides

TLC - Total Leucocytes Count

VLDL - Very low-density Lipoprotein

RESEARCH GRANT

The Dimensional Model for Clinical Data Storage and Analysis developed in this study is well accepted by **DIHAR, DRDO** in the form of a Research Grant, **Grant No. - DIHAR/01/ASSIGN/12** for the storage of clinical data of army persons.

JOURNAL PUBLICATIONS

1. **Dipankar Sengupta** & Pradeep K Naik. SN algorithm: Analysis of temporal clinical data for mining periodic patterns and impending augury. *BMC: Journal of Clinical Bioinformatics* 3(24). (doi:10.1186/2043-9113-3-24)
2. **Dipankar Sengupta**, Meemansa Sood, Poorvika Vijayvargia, Sunil Hota & Pradeep K Naik. Association rule mining based study for identification of clinical parameters akin to occurrence of brain tumor. *Bioinformation* 9(11): 555-559.
3. **Dipankar Sengupta**, Priyanka Arora, Shradha Pant, Pradeep K. Naik. Design of Dimensional Model for Clinical Data Storage and Analysis. *Applied Medical Informatics* 32(2)/2013: 47:53
4. **Dipankar Sengupta,** Vijay K Sharma, Sunil K Hota, Ravi B Srivastava, Pradeep K Naik. Key measures for evaluating cognitive performance of human population at high altitude. *Frontiers in Cognition* (Revised manuscript submitted, under review).

PAPER PRESENTED IN CONFERENCE

1. **Dipankar Sengupta,** Rajinder S Chauhan & Pradeep K Naik. Biomedical Informatics. *International Conference on Clinical Trials & Data Management* organized by CII and Ministry of Science & Technology, Govt. of India, 19^{th}-20^{th} Nov. 2010, Mumbai, India.