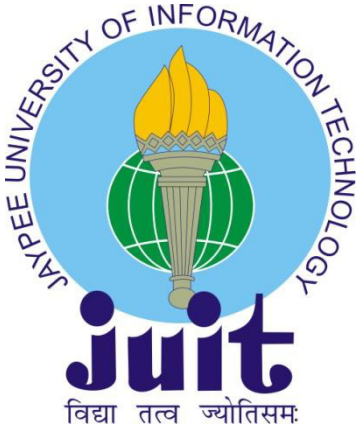
TRANSLATIONAL AND HIGH END COMPUTING OF CLINICAL DATA IN INDIA

BY

DIPANKAR SENGUPTA



JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

WAKNAGHAT, SOLAN-173234, HP, INDIA

2014

TRANSLATIONAL AND HIGH END COMPUTING OF CLINICAL DATA IN INDIA

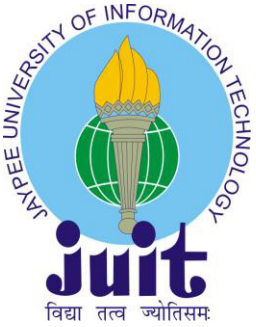
BY

DIPANKAR SENGUPTA
Enrollment No. 106506

A THESIS 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
MARCH 2014

In fond memory of

Dadu (Late Shri. Narendra Nath Sengupta),

Thakuma (Late Smt. Mina Sengupta)

and my best friend my **Kaku** (Late Shri. Manab Sengupta)

DECLARATION

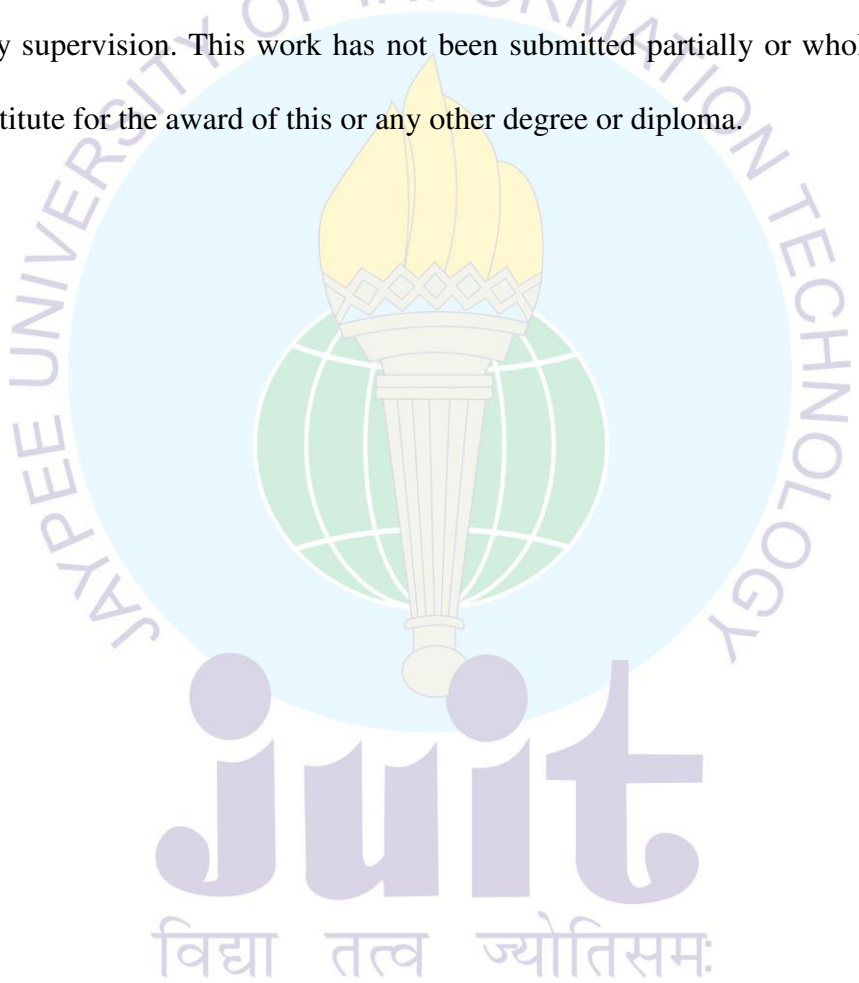
I certify that:

a. the work contained in this thesis is original and has been done by me under the guidance of my supervisor.
b. the work has not been submitted to any other organisation for any degree or diploma.
c. wherever, I have used materials (data, analysis, figures or text), I have given due credit by citing them in the text of the thesis.
d. clinical data for all the human subjects have been analyzed anonymously. Based on NOC (No objection certificate) received from the hospitals in India, all the patient information was received corresponding to IDs (Identification Number). There is no disclosure of the hospital names or patient information in this study.

Dipankar Sengupta **Date:**
(Enrollment No. 106506)
Department of Biotechnology and Bioinformatics
Jaypee University of Information Technology, Waknaghat, India
Email: dipankarsengupta.1982@gmail.com; dipankar.sengupta@juit.ac.in

CERTIFICATE

This is to certify that the thesis entitled, "**Translational and High End Computing of Clinical Data in India**" which is being submitted by **Dipankar Sengupta (Enroll. No. - 106506)** in fulfillment for the award of degree of **Doctor of Philosophy** in **Bioinformatics** at **Jaypee University of Information Technology**, **India** is the record of candidate's own work carried out by him under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.



Dr. Pradeep Kumar Naik **Date:**
Associate Professor,
Dept. of Biotechnology & Bioinformatics
Jaypee University of Information Technology
Waknaghat, Solan - 173234
Himachal Pradesh, India.
Email: pknaik1973@gmail.com, pradeep.naik@juit.ac.in

ACKNOWLEDGMENT

Firstly, I would like to thank **Prof. (Dr.) Y. Medury**, COO (Jaypee Education System) and Vice Chancellor (Acting), JUIT; **Brig. (Retd.) Balbir Singh**, Director JUIT; **Prof. (Dr.) T.S. Lamba, Dean (Academic & Research)** and **Prof. (Dr.) R.S. Chauhan, Dean (Biotechnology) and HoD (Biotechnology & Bioinformatics)** for entrusting and providing me the opportunity to pursue my research work at Jaypee University of Information Technology, Waknaghat, India.

From deep of my heart, I would like to thank my guide and mentor **Dr. Pradeep Kumar Naik, Associate Professor (Bioinformatics)** for his encouragement & guidance in pursuing my research in the emerging and challenging area of clinical informatics. I would like to specially thank him for leading me through all these years and being more than a research guide. His positive attitude and zest for quality research always encouraged me and bought the best out of me. I deem it a privilege to be doing my PhD research under him, who has endeared himself to his students.

I am grateful to DIHAR - Leh, DRDO, Ministry of Defence, Government of India, as part of this research work was kindly supported by **DIHAR, DRDO Grant - DIHAR/01/ASSIGN/12**.

Would take this opportunity to thank all the hospital and diagnostic centres across India, who readily agreed to share the clinical data (as per the NOC, won't be disclosing the names). I also owe a big thanks to four of my B.Tech (Bioinformatics) students - **Ms. Priyanka Arora** (2007-11), **Ms. Shradha Pant** (2007-11), **Ms. Meemnasa Sood** (2008-12) and **Ms. Poorvika Vijayvargia** (2008-12); who helped me in collection of clinical data, by visiting many of these hospitals and diagnostic centres for collection of hard copied reports. Also would like to thank **Prof. (Dr.) M. C. Pant** from Ram Manohar Lohia Hospital, Lucknow, India and **Dr. Ankur** from King's George Medical College, Lucknow, India who provided the technical guidance corresponding to aspects of brain tumor and other clinical parameters during the course of this study.

A big thanks to **Dr. Pradeep Kumar Pandey** from Dept. of Mathematics, Jaypee University of Information Technology, for his kind inputs and help provided to understand the concept of Jacobian and coordinate transformation. Also, my sincere thanks go to **Dr. Harvinder Singh** (Assistant Professor) and **Dr. Anil Kant** (Assistant Professor) for their continuous encouragement and support. Also would like to thank all the **faculty members, support staff** and **students** of Jaypee University of Information Technology, for their helping attitude and moral support. Beside, I can never forget my fellow researchers **Ms. Seneha Santoshi** and **Ms. Charu Suri** who have helped me in numerous ways. I will always cherish the years spent in the Department of Biotechnology & Bioinformatics, Jaypee University of Information Technology.

Also, would like to thank all my ex-colleagues and mentors from **Sapient Corp. Pvt. Ltd., India** who introduced me to the world of business intelligence technology.

Lastly, my **parents (Shri. Tapas Kumar Sengupta and Smt. Bithika Sengupta)** whose blessings have helped me in this journey. I cannot forget the pain that my parents have taken throughout my studies. It was only because of their support, constant encouragement, prayers and blessings that I could overcome all frustrations and failures. Also my **dear friends** and **teachers** from **Delhi Public School, Korba** (Std. II -XII) and **Jaypee University of Information Technology** (B.Tech. 2003-07) who were always with me during hard times.

I would like to express my heartfelt gratitude to all those who have contributed directly or indirectly towards obtaining my doctorate degree and apologize if have missed out anyone.

(Dipankar Sengupta)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF FLOW DIAGRAMS

LIST OF TABLES

LIST OF ABBREVIATIONS

ALP - Alkaline Phosphatase

BDI - Beck Depression Inventory

BS - Blood Glucose

BUN - Blood Urea Nitrogen

CANS - Computer Administered Neuropsychological Score

CBM - Core Behavioural Measures

CDMSs - Clinical Study Data Management Systems

cDNA - Complimentary Deoxyribonucleic acid

CPRSs - Clinical Patient Record Systems

CT - Computed Tomography

DBP - Diastolic Blood Pressure

DIM - Dimension

DM - Data mart

DSS - Decision Support System

DW - Data warehouse

EAV - Entity Attribute Value

EHR/EMR - Electronic Health/Medical Record

ER - Entity Relationship

ETL - Extraction, Transformation & Loading

GERD - Gastroesophageal Reflux Disease

HDL - High Density Lipoprotein

HL7 - Health Level Seven

HM - Hierarchal Modeling

IAMI - Indian Association for Medical Informatics

ID - Identification number

IDSs - Integrated Data Systems

IS - Information System

IT - Information Technology

KETTLE - Kettle Extraction, Transport, Transformation & Loading Environment

KFT - Kidney Functionality Test

LDL - Low-density Lipoprotein

LFT - Liver Functionality Test

MCI - Mild Cognitive Impairment

MDCST - Multi Domain Cognitive Screening Test

MMSE - Mini Mental State Examination

MoCA - Montreal Cognitive Assessment

MRI - Magnetic Resonance Imaging

NIH - National Institute of Health

NM - Network Modeling

OLAP - Online Analytical Processing

PR - Pulse Rate

PRESS - Predicted Residual Sum of Squares

PROCOG - Patient Reported Outcomes in Cognitive Impairment

RDBMS - Relational Database Management System

SBP - Systolic Blood Pressure

SCD - Slowly Changing Dimension

SGOT - Serum Glutamic-oxaloacetic Transaminase

SGPT - Serum Glutamate Pyruvate Transaminase

SQL - Structured Query Language

TC - Total Cholesterol

TGL - Triglycerides

TLC - Total leucocytes count

USA - United States of America

VLDL - Very Low Density Lipoprotein

WHO - World Health Organization

XML - Extensible Markup Language

ABSTRACT OF THE DISSERTATION

Healthcare sector is generating large amount of data pertaining to diagnosis, disease identification and treatment of an individual. Mining knowledge and providing scientific decision-making for the diagnosis & treatment from the clinical dataset is therefore increasingly becoming necessary. Aim of this research study was to assess the applicability of knowledge discovery on a clinical warehouse. A major contribution of the study consists of significant extensions to the data modelling for the structure of clinical warehouse. The data stored in the warehouse was subjected to data mining in form of case studies. Also, a novel temporal mining algorithm is being proposed in the study to augur state of disease for a particular patient. Clinical data used in this study was collected during the period of Oct. 2010 - Apr. 2012 from various hospitals and diagnostic centres across India. Data for all the human subjects have been analyzed anonymously. Based on NOC (No objection certificate) received from the hospitals in India, all the patient information was received corresponding to IDs (Identification Number). Utmost care has been taken for non-disclosure of the hospital names or patient information in this study.

Patient's records are increasing at an exponential rate, thus adding to the problem of data management and storage. Major problem being faced corresponding to temporal storage of this clinical data, is the varied dimensionality, ranging from images to qualitative and quantitative form. Therefore there is a need for development of efficient data model which can handle this multi-dimensionality data issue and store the data in temporal aspect. For the stated problem lying in façade of clinical informatics, a clinical dimensional model design is being proposed in the study, that can be used for development of clinical data mart. The model has been designed in this study, keeping in consideration temporal storage of patient's data corresponding to all possible clinical parameters in a non-volatile, subject-oriented and integrated state. Availability of said data for each patient can be then used for application of data mining techniques for finding the correlation & association among parameters at the level of individual and population.

Section of data from this warehouse, comprising of 550 patients diagnosed with brain tumor (primary stage), was subjected to mining of associative rules. Apriori association rule algorithm was applied to discover associative rules among the clinical parameters. The rules discovered in the study 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 predictive 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). The results indicate that the methodology followed may be of good value for the diagnostic procedure of brain tumor, especially when large data volumes are involved and screening based on discovered parameters would allow clinicians to detect tumors at an early stage of development.

It has been a big challenge for mining clinical data considering varied temporal points. Therefore, a conjoined solution to analyze the clinical parameters akin to a disease is also being proposed in this study in form of a new algorithm, SN algorithm, to map clinical parameters along with a disease state at various temporal points. SN algorithm is based on Jacobian approach, which augurs the state of a disease 'Sn' at a given temporal point 'Tn' by mapping the derivatives with the temporal point 'T0', whose state of disease 'S0' is known. The predictive ability of the proposed algorithm was evaluated on a temporal clinical data set of brain tumor patients. A very high prediction accuracy of ~97% for a brain tumor state 'Sn' for any temporal point 'Tn' was obtained. The results indicate that the methodology followed may be of good value to the diagnostic procedure, especially for analyzing temporal form of clinical data.

Another case study was performed in this study for identification of screening parameters to be associated with Mild Cognitive Impairment (MCI) for human population staying at high altitude (>4300m). As altitude increases there is scarcity of oxygen which may lead to cognitive impairment and other clinical disorders. There are various screening tests for analyzing the onset of cognitive impairment but one of the key tests developed in recent years is multi-domain cognitive screening test (MDCST). Aim of this particular study was to identify cognitive analyzers among MDCST 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. Chi-square significance test was applied on the lowlander & highlander population data set respectively to identify the significant MDCST and clinical parameters. Apriori algorithm was applied to discover association rules among the clinical, behavioural and cognitive screening parameters. 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 with atleast 30% support and more than 60% confidence for behavioural and clinical measures. This study was first of an attempt to identify

key cognitive & clinical measures specific to lowlander and highlander population staying at high altitude for prolonged duration which can be associated with cognitive impairment.

In conclusion, this thesis's results highlights appropriate storage structure of clinical data, and subjecting it to data mining for discovery of knowledge pertaining to a patient or disease. It makes suggestions for developing or extending evaluation methods that can be applied to this area with a multi-actor perspective in order to understand the effects, consequences, and prerequisites that have to be achieved for the successful implementation and use of IT in healthcare.

Keywords: Apriori algorithm; Association rule mining; Beck depression inventory (BDI); Blood urea nitrogen; Creatinine; Cholesterol; Clinical Informatics; Datamart; Data mining; Datawarehouse; Dimensional model; Folic Acid; HDL, Jacobian; Jacobian determinant; Mild cognitive impairment (MCI); Multi-domain cognitive screening test (MDCST); SGOT; SGPT; Slowly changing dimension; Temporal mining; Vitamin B-12.

CHAPTER - 1

INTRODUCTION TO CLINICAL INFORMATICS

1.1 What is Clinical Informatics & its relation with Clinical Bioinformatics?

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].

In the early 1990's, clinical physicians were open to advances in omics technology despite the barriers which existed for physicians applying genetic tests, for example the low tolerance for uncertainty, negative attitudes about their responsibility for genetic counselling and testing, as well as unfamiliarity with ethical issues raised by testing [2]. By mid 1990's, bioinformatics was being suggested to be applied for clinical toxicology [3] and cancer [4]. Whereas by the beginning of the 21st century, gene expression profiles in human cancer cell lines were evaluated by cDNA microarrays and corrected with drug activity patterns by combining bioinformatics and chemoinformatics [5]. Thus, clinical bioinformatics came into existence with a purpose of providing biological and medical information for individualized healthcare; enable researchers to search online biological databases and use bioinformatics in medical practice; selection of appropriate software to analyze the data for medical decision-making; optimize the development of disease-specific biomarkers; and supervise drug target identification and clinical validation [6]. It was said to play an important role in a number of clinical applications, including omics technology, metabolic and signalling pathways, biomarker discovery and development, computational biology, genomics, proteomics, metabolomics, transcriptomics, high-throughput analysis, human molecular genetics, human tissue bank, mathematical medicine and biology, protein expression, molecular profiling and systems biology.

But it’s not only about discovering a drug to cure an epidemic. Technology allows clinical research and patient care to become more integrated and interactive. In so-called translational research, it’s a need that basic science and clinical researchers work together on interpretation and application of research data in clinical settings [7]. Data sharing is necessary to improve the quality of healthcare and to accelerate progress in biomedical sciences from bench to bedside to community. To go from clinical research to community practice, integrated data systems (IDSs) must be created to allow community researchers to easily access secure and confidential research

data [7]. Henceforth, **clinical bioinformatics** evolved and is being primarily suggested to be associated with the analysis and visualization of complex clinical datasets [1]. It majorly focuses on **clinical informatics**, and deals with 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 [8] 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 [9]. Analyzing the 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 [10-13]. However, research reveals that this design is not efficient for data sets with large numbers of attributes that vary over time [14]. Therefore, 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 and improved diagnosis for better therapies.

1.2 Current Scenario of clinical informatics

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 [15]. As an effort to facilitate access to this wealth of information, databases were developed that contained clinical data from healthcare organizations [16]. Nagarajan *et al.* introduced data-warehousing-based solutions utilizing relational database management systems (RDBMSs) for assembling and integrating data [13]. A relational database model is composed of classes of data, with each class characterized by a set of attributes. Continuously changing the number and type of attributes necessitates frequent modification of the database structure. Recent research has focused on the conceptual development of IDs using ontology-based systems for the design and integration of clinical trial data [11]. Wang *et al.* developed a BioMediator system to provide a theoretical and practical foundation for data integration across diverse biomedical domains via a "knowledge-base-driven centralized federated database" model [12]. However, the efficiency of query processing time and the need to filter out unnecessary query results still are concerns. The data architecture required for clinical data warehousing has been researched in applications such as clinical study data management systems (CDMSs) and clinical patient record systems (CPRSs). They both use an entity-attribute-value (EAV) system (i.e., row modeling) as

opposed to conventional database design [17]. The EAV system has the advantage of remaining stable as the number of parameters increases when knowledge expands, a common situation in the basic sciences and in clinical trials [14,18].

The enormous amount of data collected by EHR/EMR has found additional value when integrated and stored in database. It is 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 [19]. 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 [20]. 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 [19]. Similarly, technique using association rule mining [21] 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 and development in this field is still in nascent stage. Indian policy makers are yet to realize the importance of clinical informatics in delivering healthcare [22]. 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.[23]. 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 and non-profit-distributing organization [24]. 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 [24]. Similarly, Indian Association for Medical Informatics (IAMI) was started up in 1993 with a mission to introduce use of computer and its application all the colleges of medical sciences, dental sciences, nursing and pharmacy in India through Medical Council of India and respective university authorities [25]. IAMI is currently focusing on development of clinical intelligence model for hospitals across Andhra Pradesh, India [20].

1.3 Existing Challenges in Clinical Informatics & Specific Problems Addressed

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

healthcare in spite of the vast amount of literature on the proven benefits of IT for tackling the fundamental problems of healthcare [26]. 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), US in 2011 (17th Aug 2011) has stated about the various aspects which needs to be focused pertaining to application of IT in clinical science. 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 are:

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 and agility?
7. How to achieve performance and scalability?

As Clinical informatics is the study of information systems (computers and programs), used in the clinical practice of medicine, an attempt might be to contribute in the following aspects:

- **Data Storage** - data for patients can be stored temporally for long period of time.
- **Telemedicine** - the clinical reviewer in X location can observe the findings for a patient in rural Y location without the physical slides - easy accessibility
- **Data Display** - vital signs can be highlighted when abnormal; mean or median values can be graphed with the raw numbers over time to simplify clinician review.
- **Decision Support** - immediate feedback at the time of order entry about any correlation among various diagnostic parameters of that patient can be shown to reduce both patient morbidity and healthcare costs.

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 spatiotemporal mining of the stored clinical data.*
3. *Decipher patterns and rules from the clinical data set by applying existing data mining approaches.*

1.4 Objectives and Goals of this study

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.*

1.5 Proposed architectural model

Each of the said objective 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, we have arranged each work as separate chapter in a publishing format (Chapter 2, 3, 4 and 5 respectively). 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.

The research study to attain the said goals has been organized in following steps:

1. Clinical data and use of IT for its management as the research domain.
2. Investigation of domain knowledge and associated terminology.
3. Collection of data from varied sources

4. Data modelling and proposal of a clinical dimensional model.
5. Develop a clinical warehouse based on proposed model.
6. Apply existing data mining algorithm to decipher hidden patterns and rules from the clinical warehouse.
7. Design an algorithm for temporal mining of clinical data.

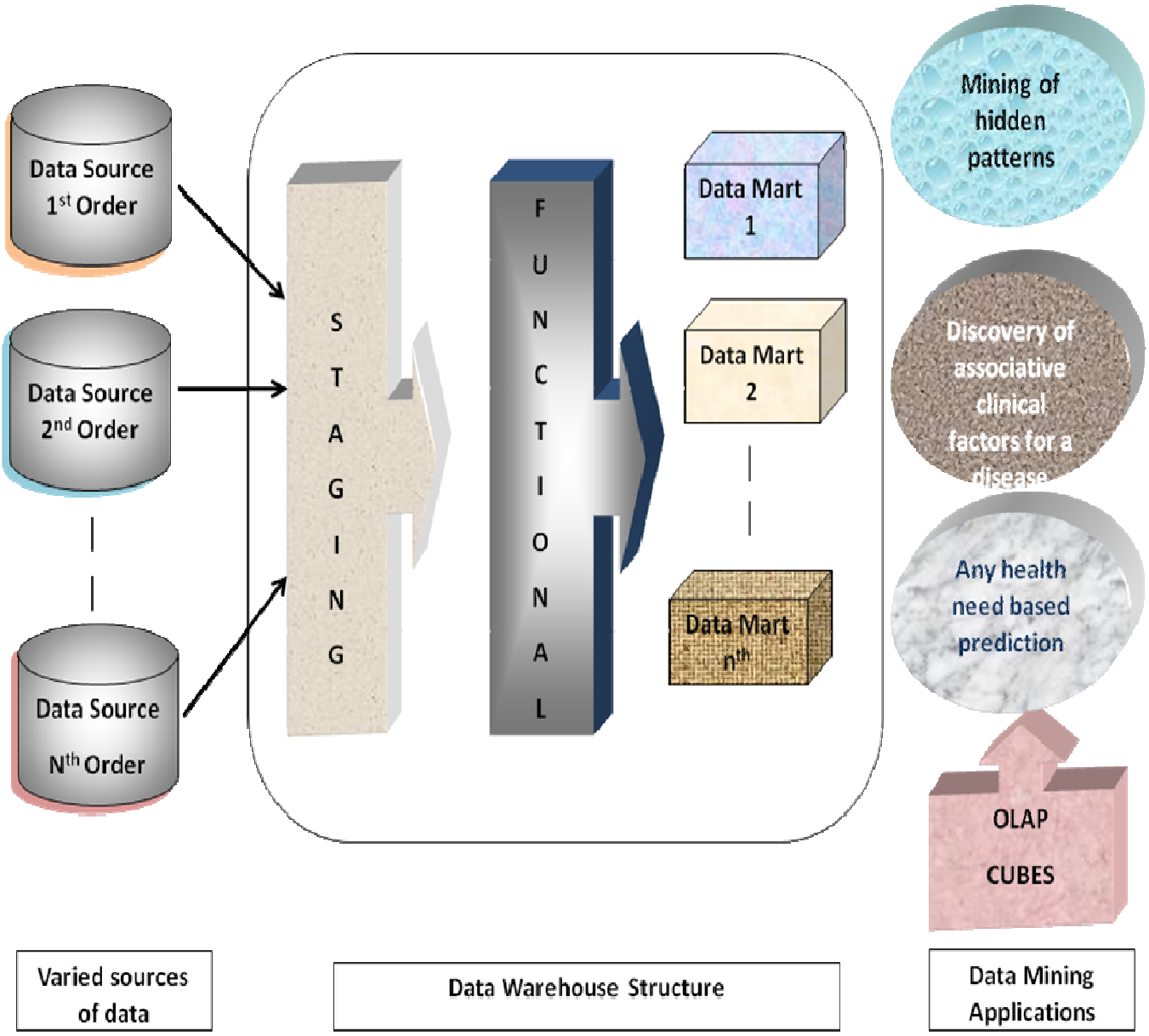


Figure 1 - Architectural design for the proposed business intelligence model of clinical data temporal management and analysis.

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. The architecture has been encompassed into 3 sections: Varied sources of data; Data warehouse structure comprising of

staging and functional schema; and Data mining applications comprising of OLAP cube along with application of existing data mining algorithms on the warehouse data and development of new algorithm for spatiotemporal mining. Each of the said component is being discussed in brief:

1.5.1 Data Sources

The study was initiated by defining clinical informatics as the research domain. Clinical aspects and terminologies associated to domain was analyzed by the help of Doctors (clinical physicians), Pharmacist and Radiologists. Various forms in which clinical data is being generated were analyzed. The major problem observed is of varied dimensionality, ranging from images to numerical form of data which needs to be answered (Figure 2).

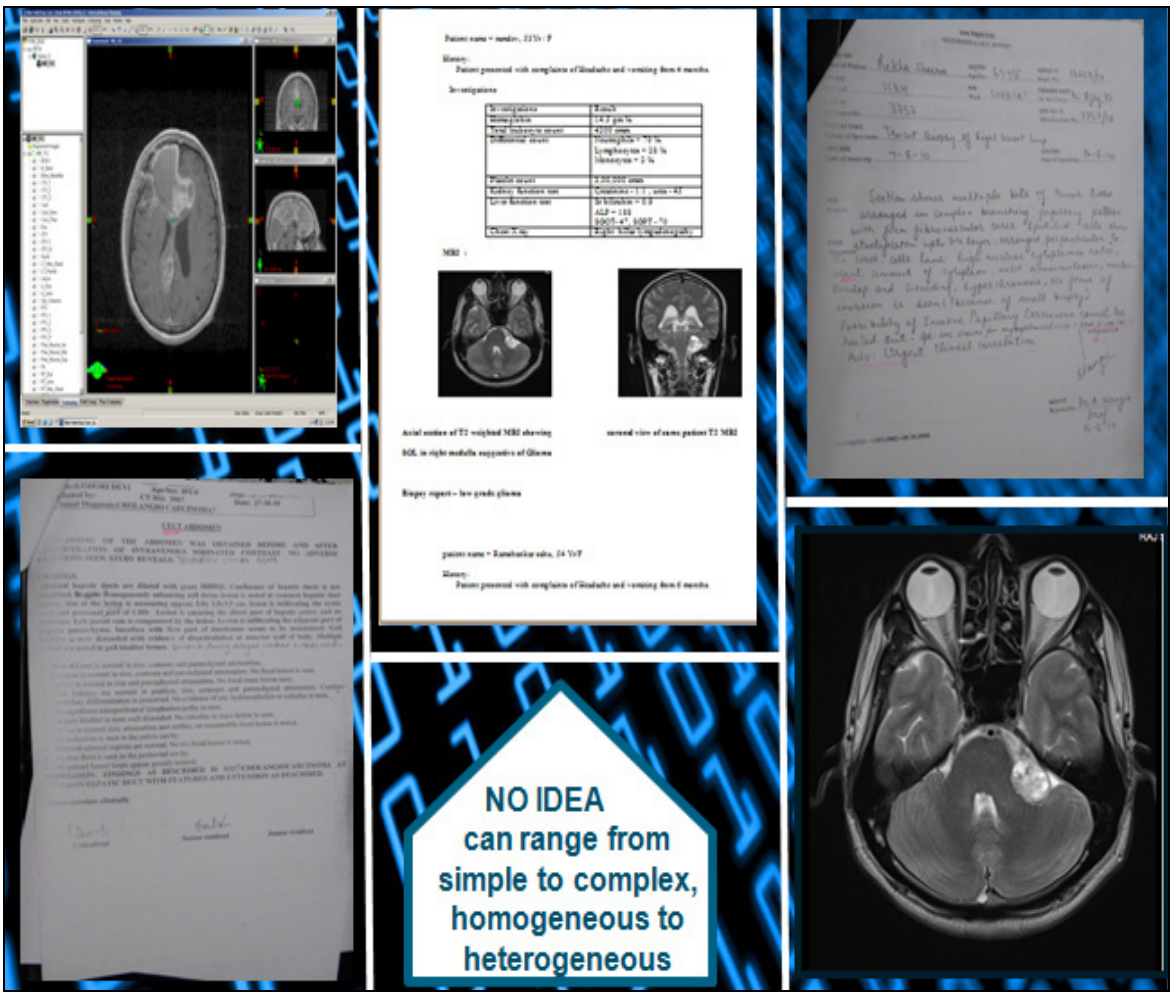


Figure 2 – Various sources of clinical data, ranging from textual, numerical to images.

During the phase of data collection in this study we got acquainted with the realities of the problems associated with data storage in hospitals across India, which are:

- Awareness about usage of informatics is very less among the faculty members and the medical practitioners in the country.

- The hospitals do not keep a record of all the data related to the patient, like usually they delete the MRI/CT scan images because it takes lot of memory, so it's practically not possible to store them for long term.
- Data is not available in integrated form i.e different IDs (identification number) were given to a particular patient in different departments of a hospital. For example, if a patient is prescribed for biopsy then an ID is assigned corresponding to this department, now if the same patient is prescribed for MRI then they assigned a different ID respective to MRI.
- The data is being stored across hospitals in different formats, like electronic records, handwritten and printed reports, images in different formats (for example DICOM, jpeg, png etc.).
- Hospitals do not owe any responsibility for maintaining data of patients in temporal form, rather patients are required to maintain and carry all their reports during the visit to hospital.
- Out of all, the major problem still remains the confidential policy of the health care institutes in the process of data collection.

1.5.2 Data Model

A data model is a conceptual representation of the data structures that is required for building a database. The data structures include the data objects, the associations between data objects, and the rules which govern operations on the objects. It focuses on what kind of data is required and how it should be organized rather than what operations will be performed on the data. To use a common analogy, the data model is equivalent to an architect's building plans. A data model is independent of hardware or software constraints. Rather than try to represent the data as a database would see it, the data model focuses on representing the data as the user sees it in the "real world". It serves as a bridge between the concepts that make up real-world events and processes and the physical representation of those concepts in a database. The data model gets its inputs from the planning and analysis stage. Here the modeler, along with analysts, collects information about the requirements of the database by reviewing existing documentation and interviewing end-users [27].

1.5.2.1 Importance of data modeling

Data modeling is probably the most intensive and time consuming part of the development process. A common response by practitioners who write on the subject is that you should no more build a database without a model than you should build a house without blueprints [27]. The goal of

the data model is to make sure that the all data objects required by the database are completely and accurately represented. Because the data model uses easily understood notations and natural language, it can be reviewed and verified as correct by the end-users. The data model is also detailed enough to be used by the database developers to use as a "blueprint" for building the physical database. The information contained in the data model will be used to define the relational tables, primary and foreign keys, stored procedures, and triggers. A poorly designed database will require more time in the long term. Without careful planning a database may be created that omits data required to create critical reports, produces results that are incorrect or inconsistent, and is unable to accommodate changes in the user's requirements.

There are different data modeling concepts like ER Modeling (Entity Relationship modeling), DM (Dimensional modeling), Hierarchal Modeling (HM), Network Modeling (NM), etc. [27]. But most popular are ER and DM.

1.5.2.2 Dimensional Model

Dimensional modeling is a technique for conceptualizing and visualizing data models as a set of measures that are described by common aspects of the domain [28]. Dimensional modeling has two basic concepts:

Fact:

- A fact is a collection of related data items, consisting of measures.
- A fact is a focus of interest for the decision making process.
- Measures are continuously valued attributes that describe facts.
- A fact is the domain measure.

Dimension:

- The parameter over which we want to perform analysis of facts.
- The parameter that gives meaning to a measure number of customers is a fact, perform analysis over time.

Since a dimensional model is visually represented as a fact table surrounded by dimension tables, it is frequently called *star schema*. Another kind of schema include *snowflake schema*, which is a logical arrangement of tables in a multidimensional database such that the entity relationship diagram resembles a snowflake in shape. The snowflake schema is represented by centralized fact tables which are connected to multiple dimensions.

Clinical dimensional model has been proposed considering the clinical data structure and how the temporal storage can be managed. The detail of the model has been discussed in *chapter 2*.

1.5.3 Data warehouse

Based on the dimensional model designed the data warehouse structure is physically constructed using a Relational Database Management System (RDBMS).

To understand a data warehouse, it is important for us to realize that it is not a single object. It is more of a strategy or a process, an integration of various support systems and programs that are knowledge based. The goal of using a data warehouse is to have the historical data availability to make strategic decisions. The data warehouse concept originated in an effort to solve data synchronization problems and resolve data inconsistencies that resulted when analysts acquired data from multiple operational or production systems. One of the most important functions of a data warehouse is to serve as a collection point for consolidating and further distributing data extracts from an organization's production systems.

A data warehouse may be defined as structured repository of subject-oriented, integrated, time-variant, and non-volatile collection of data in support of decision-making process. The meanings of the key terms associated are as defined below:

- **Subject-Oriented**: Organization of data in a warehouse is around the key subjects (or high-level entities) of the enterprise. For instance: patients, disease and diagnostic test.
- **Integrated**: The data is assumed to be using consistent naming conventions, formats, encoding structures, and related characteristics for sharing and usability.
- **Time-variant:** Data contain a time dimension so that they can be used for historical purposes.
- **Nonvolatile**: Data are refreshed from operational data, and cannot be updated by users.

Considering the above key terms data warehousing could be defined as the process by which an organization extract meaningful information from historical data. There may be different needs for the data; therefore smaller data units may be constructed that are tailored towards certain subjects. These small data warehouses are referred to as *data marts*. A *data mart* (DM) is the access layer of the data warehouse (DW) environment that is used to get data out to the users. A data mart is a focused subset of a data warehouse that deals with a single area (like clinical data, genomic data, etc.) of data and is organized for quick analysis.

Data warehouses provide access to data for complex analysis, knowledge discovery, and decision-making to build a *decision support system* (DSS). **The quality of decisions that are facilitated by a data warehouse is only as good as the quality of the data.**

A data warehouse structure usually consists of 2 schemas, which are:

Staging Schema - A staging schema is being created to dump the data from varied sources in its original form.

Functional Schema - The designed dimensional model is created physically as the functional schema. In case a warehouse comprises of *data marts,* then it may have more than 1 functional schema and each one is specific towards a mart.

The detailed description of the data warehouse developed in this study is discussed in *chapter 2*.

1.5.4 OLAP cube

OLAP stands for on-line analytical processing. Codd made first of an attempt in 1993 to define it by proposing 12 rules of OLAP [29]. One of the applications of OLAP is defining an OLAP cube, which is an array of data that is described in terms of 1 or more dimensions. It's an technique being used for fetching information from the warehouse for the purpose of data mining. Cube is a shortcut for fetching multidimensional dataset, given data has arbitrary number of dimensions. In database theory, an OLAP cube is an abstract representation of a projection of an RDBMS relation. Given a relation of order N, consider a projection that subtends X, Y and Z as the key and W as the residual attribute. Characterizing this as a function,

$$f: (X, Y, Z) \text{ -----> } W,$$

the attributes X, Y and Z correspond to the axes of the cube, while the W value into which each (X, Y, Z) triple maps corresponds to the data element that populates each cell of the cube.

1.5.5 Data mining

Data mining, *the extraction of hidden predictive information from large databases and data warehouses*, is a powerful technology with great potential to help organizations focus on their most important information [30]. Data mining helps to predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Most of the organizations collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of these existing information resources, and can be integrated with new products and systems as they are brought on-line.

It is the extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data [31]. Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. It can include statistical models, mathematical algorithms, and machine learning methods (algorithms that improve their performance automatically through experience, such as neural networks or decision trees). Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction.

Application of data mining algorithms on the clinical warehouse is being discussed as a case study on *brain tumor in chapter 3* and on *mild cognitive impairment (MCI) in chapter 5* respectively. A new *algorithm for spatiotemporal mining of clinical data is being proposed and discussed in chapter 4*.

References

1. Wang, X. & Liotta L. *Clinical bioinformatics: a new emerging science*. Journal of Clinical Bioinformatics, 1(1) pp 1. 2011.
2. Geller, G. & Holtzman, N.A. *Implications of the human genome initiative for the primary care physician*. Bioethics, 5 pp 318-25. 1991.
3. Breckenridge, A. *A clinical pharmacologist's view of drug toxicity*. British Journal of Clinical Pharmacology, 42 pp 53-58. 1996.
4. Hainaut, P., Soussi, T., Shomer, B., et al. *Database of p53 gene somatic mutations in human tumors and cell lines: updated compilation and future prospects*. Nucleic Acids Research, 25 pp 151-157. 1997.
5. Scherf, U., Ross, D.T., Waltham, M., et al. *A gene expression database for the molecular pharmacology of cancer*. Nature Genetics, 24 pp 236-244. 2000.
6. Chang, P.L. *Clinical bioinformatics*. Chang Gung Medical Journal, 28 pp 201-211. 2005.
7. Viangteeravat, T., Brooks, I.M., Smith, E.J. et al. *Slim-prim: a biomedical informatics database to promote translational research*. Perspectives in Health Information Management, 6(6). 2009.
8. Schwarz, E., Leweke, F.M., Bahn, S. & Lio, P. *Clinical bioinformatics for complex disorders: a schizophrenia case study*. BMC Bioinformatics, 10 S6. 2009.
9. Berger, A.M. & Berger C.R. *Data mining as a tool for research and knowledge development in nursing*. Computers Informatics Nursing, 22(3) pp 123-131. 2004.
10. Brazhnik, O & Jones J.F. *Anatomy of Data Integration*. Journal of Biomedical Informatics, 40(3) pp 252-269. 2007.
11. Geisler, S., Brauers, A., Quix, C. & Schneink, A. *Ontology-based System for Clinical Trial Data Management*. Proceedings: Annual Symposium of the IEEE/EMBS Benelux Chapter Heeze, the Netherlands, pp 53-55. 2007.
12. Wang, K. *et al. BioMediator Data Integration: Beyond Genomics to Neuroscience Data*. AMIA Annual Symposium Proceedings, pp 779–783. 2005.
13. Nagarajan, R., Ahmed, M. & Phatak, A. *Database Challenges in the Integration of Biomedical Data Sets*. Proceedings of the Thirtieth International Conference on Very Large Data Bases, Toronto: VLDB Endowment, 2004.
14. Dinu, V. & Nadkarni, P. *Guidelines for the Effective Use of Entity–Attribute–Value Modeling for Biomedical Databases*. International Journal of Medical Informatics, 76(11-12) pp 769-79. 2007.

15. Atherton, J. *Development of the Electronic Health Record.* American Medical Association Journal of Ethics*,* 13(3) pp 186–189. 2011.
16. Evans, R.S., Lloyd, J.F. & Pierce, L.A. *Clinical use of an enterprise data warehouse*. AMIA Annual Symposium Proceedings, pp 189-98. 2012.
17. Deshpande, A.M., Brandt, C. & Nadkarni, P.M. *Metadata-driven Ad Hoc Query of Patient Data*. Journal of the American Medical Informatics Association, 9 pp 369–382. 2009.
18. Anhøj, J. *Generic Design of Web-Based Clinical Databases*. Journal of Medical Internet Research, 5(4) e27. 2003.
19. Houston, A.L., Chen, H., Hubbard, S.M., et al. *Medical Data Mining on the Internet: Research on a Cancer Information System*. Artificial Intelligence Review, 13 pp 437-466. 1999.
20. Holmes, A.B., Hawson, A., Liu, F., et al. *Discovering disease associations by integrating electronic clinical data and medical literature*. PLOS One, 6(6). 2011.
21. Brossette, S.E., Sprague, A.P., Hardin, J.M., et al. *Association rules and data mining in hospital infection control land public health surveillance*. Journal of the American Medical Informatics Association, 5 pp 373-381. 1998.
22. Sarbadhikari, S.N. *The state of medical informatics in India: a roadmap for optimal organization*. Journal of Medical Systems, 29(2) pp 125-141. 2005.
23. Paul, P.K., Kumar, A. & Chatterjee D. *Health Informatics and its practice: Emerging Domain of Information Science - Indian Scenario*. Current Trends in Biotechnology and Chemical Research. 2012.
24. Health Level Seven. Internet: http://*www.**hl7india**.org,* [Jan. 11, 2012].
25. Indian Association for Medical Informatics. Internet: http://*www.**iami**.org.in*, [Jan. 11, 2012].
26. Atalag, K. *Archetype based domain modeling for health information systems*. Graduate School of Informatics, The Middle East Technical University, Turkey. 2007.
27. Data Modeling. Internet: http://www.liberty.edu/media/1414/%5B6330%5DERDDataModeling.pdf, [Aug 24,2010].
28. Kimball, R. & Ross, M. *The Data Warehouse Toolkit, 2nd edition*. John Wiley & Sons, Inc., New York. 2002.
29. Codd, E.F., Codd, S.B., & Salley C.T. *Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate*. Codd & Date, Inc. 1993. Internet: http://olap.com/w/index.php/Codd%27s_Paper, [Jan 30, 2012].

30. An Introduction to Data Mining. Internet: www.thearling.com/text/dmwhite/dmwhite.htm, [Mar 30, 2011].
31. Seifert, J.W. *Data Mining: An Overview*. Internet: www.fas.org/irp/crs/RL31798.pdf, [Mar 30, 2011].

CHAPTER - 2

DESIGN OF DIMENSIONAL MODEL FOR CLINICAL DATA STORAGE AND ANALYSIS

2.1 Introduction

A major problem being faced by most of the organizations and industries around the world is efficient storage of large amount of data and its maintenance. Most of the financial services, telecom giants and other service providers hence forth have tried to take help from information technology for getting storage solutions. Beside storage, they are also interested in using the available data for future predictions for growth of the business. Ralph Kimball suggested [1] about operational systems and data warehouse that can be associated with a organization corresponding to their data storage needs. If operational system is meant for turning the wheel of the organization then a data warehouse, on the other hand, watch the wheels of the organization getting turned [2]. It is widely recognized that the data warehouse has profoundly different needs, clients, structures, and rhythms than the operational systems of record. A Datawarehouse (DW) is a specialized form of relational database that stores information oriented to satisfy decision-making requests [3]. A very frequent problem in enterprises is the impossibility for accessing to corporate, complete and integrated information of the enterprise that can satisfy decision-making requests. In general, a DW is constructed with the goal of storing and providing all the relevant information that is generated along the different databases of an enterprise.

A similar kind of scenario is being faced now in the field of *clinical science*, where large amount of data is generated on daily basis. Every day in a different country in a different state in a different city in a different hospital lands a new patient, a new case, a new heap of data but an old problem still persists, i.e. of data storage for nearly every hospital or research institute. It's time now for change and advancement. The extraordinary explosion of medical knowledge, technologies, and ground-breaking drugs may vastly improve healthcare delivery for the welfare of its consumers, but the key is to implement these technologies, to extract as much as we can. Since clinical informatics is a multidisciplinary field, it combines data representation, cognitive science, policies, telemedicine and data discovery. The ability to quickly and efficiently retrieve information makes the creation of an organized database indispensable, and thus clinical informatics makes the representation and interpretation of complex medical terms quite simple for a spcialized form of clinicial database. Cognitive science comes into play to help those in the medical community, understand process and perceive artificial intelligence and computing.

Once diagnosis process of an individual is being completed, what hospitals usually consider to be the junk data might be as important and meaningful as a medicine given to a patient. It’s all about fetching information from this raw data which can form a base for knowledge discovery. The information may be of help to a patient corresponding to temporal analysis of clinical data as and

when studied. Also on a larger scale this information can help in prevention, proactive treatments and early detection of certain life threatening diseases at population level.

Clinical informatics, deals majorly with the clinical data concerned with a patient or a group of patients, which may include a patient's health records, history with the disease, and treatment description etc. Technology allows clinical research and patient care to become more integrated and interactive. These data can then be used to answer questions relevant to specific communities and can be extrapolated to a national level, a classical example of which is the Slim Prim Biomedical Database [4]. Furthermore, information can be assimilated for community education to help improve healthcare.

Datawarehouse is usually developed using a specific blue print design, said to be the dimensional model. A dimensional model is a "specific discipline for modeling data that is an alternative to entity relationship modeling" [2]. Like an entity relationship model, a dimensional model reflects a data structure and is specifically designed to model data in a way that emphasizes user understandability, enhances query performance, and tracks change [2,5,6]. To achieve these design characteristics, a dimensional model is typically being kept in a de-normalized state. There are two kinds of tables in a dimensional model - dimension and fact. Dimensional tables consist of descriptive attributes which can help in describing business entity whereas a fact table consists of measure corresponding to each of the feature. Dimension tables contain primary keys which associate the dimension attributes to the fact table, and textual descriptions. Fact tables contain foreign keys and measurements. An effective data warehouse can be built and maintained only when, it has an effective design and well defined grain of its dimensional model.

To address the clinical data integration issues and to have a data warehouse based storage structure which can effectively handle the clinical data temporally, a clinical dimensional model is being proposed, that will address the concerned dimensionality issue. Ralph Kimball [2] 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 various 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 which can be said to associated with effective recording of an individual health data is being depicted in Figure 3, which depicts in near future along with clinical and drug related data, in addition, the genomics and proteomics data are also going to play a major role with respect to an effective treatment process. Each of the said domains can lead to development of a specific data mart associated with the data warehouse. The current study is focusing on one of the

said domains of clinical data by providing a dimensional model design which can be used for development of a clinical data mart.

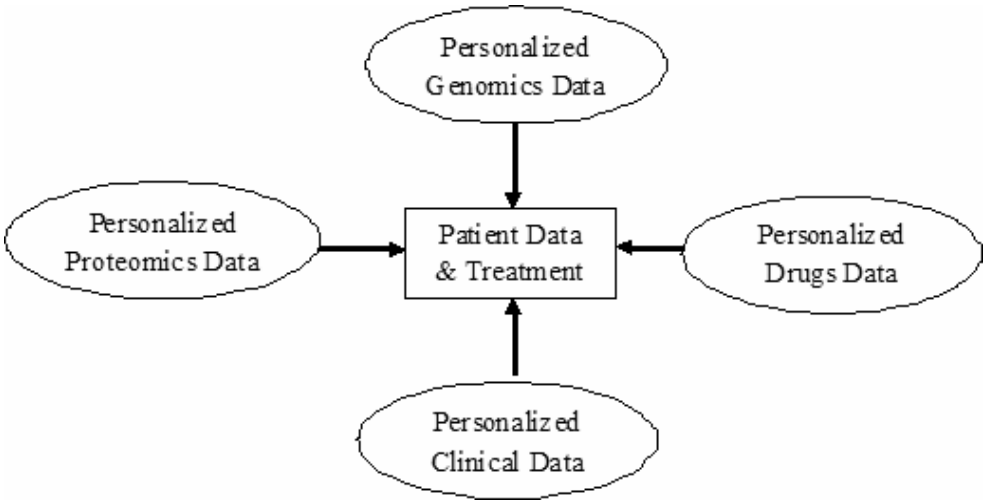


Figure 3 - Domains associated to health care.
(different domains that can be associated with health care in future)

2.2 Materials and Methods

2.2.1 Data processing

The data obtained from different hospitals could not be straight away used for analysis, "*data in the real -world is dirty*" i.e. have errors, unusual values, and inconsistencies. Data quality can be assessed in terms of accuracy, completeness and consistency. The data that we obtained was:

- *Incomplete*: Some of the records were lacking attribute values. Eg: few of the patient's record did not have 'basophil' count while few missed 'serum bilirubin', that leads to incompleteness in the data.
- *Noisy:* Means that the data contains errors or outliers Eg: A record had value of 10 in platelet count which was an error.
- *Inconsistent:* Containing discrepancies in codes or names or format Eg: The date in few records were in mm/dd/yy format while in others as dd/mm/yyyy format.

There are many reasons that lead to such kind of data. Incomplete, noisy, and inconsistent data are commonplace properties of large, real-world databases and data warehouses. Attributes of interest may not always be available, such as kidney function test may not be performed for a patient every time visiting the hospital. Other data may not be included simply because it was not considered important at the time of entry. Relevant data may not be recorded due to

misunderstanding, or because of equipment malfunctions. Data that were inconsistent with other recorded data may have been deleted. Furthermore, the recording of the history or modifications to the data may have been overlooked. Missing data, particularly for tuples with missing values for some attributes, may need to be inferred. Data can be noisy, having incorrect attribute values, owing to the following:

- The data collection instruments used may be faulty.
- There may have been human or computer errors occurring at data entry.
- Errors in data transmission can also occur.
- There may be technology limitations, such as limited buffer size for coordinating synchronized data transfer and consumption.
- Incorrect data may also result from inconsistencies in naming conventions or data codes used.

2.2.1.1 Level of Redundancy

It is another important factor in data processing. It is useful to know how much of the data is repeated from the various sources. Redundant data can slow down or confuse the knowledge discovery process. Data reduction and cleaning methods, carefully employed, can aid in removing duplicated data prior to its usage.

2.2.1.2 Major Tasks in Data Preprocessing

- Data cleaning: Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.
- Data integration: Integration of multiple databases, data cubes, or files.
- Data transformation: Normalization and aggregation.
- Data reduction: Obtains reduced representation in volume but produces the same or similar analytical results.
- Data discretization: Part of data reduction but with particular importance, especially for numerical data.

Accuracy of data is an important criteria to be considered during development of a clinical warehouse especially when there are no Electronic Health/Medical Records (EHR/EMR) implemented [7]. Data incorrectness usually exists because of design or operational deficiency and can be identified where the mapping between the information system state and the real world state break down [7]. Henceforth, with utmost care the dimensional model (data model) of the clinical warehouse was designed based on the descriptive and measurable features of the clinical data. Further, it consists of date and time dimension that ensures temporal storage of data for a patient.

Also, to check the operational deficiencies, the quality assurance of data was ensured by implementing appropriate data processing codes for range and data validation checks [8], re-entering samples of data to assess for accuracy, checks for data completeness and attention for data consistency [9].

The dimensional model had being designed using Erwin data modeller 8.2 [10]; Kettle 3.1 [11] was used to encode ETL mappings for processing of data from source files into the data warehouse; and MySQL 5.019 [12] relational database management system was used for the physcial creation of warehouse.

2.2.2 Proposed Design for the Clinical Dimensional Model

The advanced form of clinical data warehouse would be complex and time consuming to review a series of patient records. However, its going to be a efficient data repository existing to deliver quality patient care. Data integration tasks of medical data store are challenging scenarios when designing clinical data warehouse architecture.

A few decades ago, physicians knew pretty much everything that is to be known about medicines; most doctors could recollect the names of their patients. However, today, no doctor can keep up with the explosion of medical and health information. While health care organizations have recognized the use of computers, but in comparison to other industries its application in healthcare have not been encouraging. This is because, among other factors, it takes too long to get information in many cases; there is no easy accessibility to data, and no uniform standard among various vendors. But once the data warehouse is ready, it's worth spending the time and money in it.

With its 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 which needs to be answered. Based on the same we propose for an appropriate clinical dimensional model for the structure of a clinical data mart that will store data at granularity of an individual corresponding to time. The given model has been designed using Erwin data modeller (version 8.2) [10].

2.2.3 Creation of clinical warehouse

MySql 5.019 RDBMS package was used to physically create the data warehouse. The data extraction, cleaning and processing process was done using Kettle 3.1 [Extraction, Transformation, Loading (ETL) technology].

The aim of building this warehouse is to lead to a platform for applying data mining technique to find correlation among various attributes, applying association mining studies, etc. which would help us in deciphering new translational paradigms which could be used by doctors, physicians, other health professionals and even by a common man who has got knowledge about how to use computer and internet.

2.2.3.1 Staging Schema

The **staging database** is a separate data cache (storage area) that helps users in continuous access to application data. Its access continues even when data is being imported from the various external sources and prepared for loading. This minimizes the downtime that user experiences during data loading or data refreshing. Here the data is dumped as it is, without any changes being made to it i.e. the data here is in its original form.

The data obtained from varied sources were formated into a common input form. All the files were conversed into 'csv' and images into 'jpeg' form. The files were categorized into date, time, patient, disease, diagnostic test and image categories respectively. A staging database named as *Clinical_Staging_Data* was created for storage of data from source files. Each category file types were processed to a particular table in the staging database as defined in Table I. For uniquely identifing data from a particular hospital an additional field was added as "hospital name and location", that was picked from the name of file being processed. SQL queries for creation of tables for staging database can be accessed from *Appendix I - A1*. ETL mapping were designed using Kettle 3.1 platform to process the files from diferrent source files into respective tables of staging database.

TABLE NAME (STAGING SCHEMA)	TYPE OF DATA STORED
Date	This table stores the date records. Here each date has been given a unique id and other details like week of the year, in which quarter of the year is the dates falling are given.
Time	This table stores the time records. Here each time value (in terms of hour, minute & second) has been given a unique id.
Patient	It stores the data collected from hospitals with details of patients records.
Disease	It stores the data corresponding to all the disease for which a patient may be diagnosed.

Diagnostic_test	It stores the data corresponding to all diagnostic tests which may be subjected to a patient.
Image	It stores numerical information of all the images along with the path details of their physical location in the system.

Table I - Details of Clinical_Staging_Data (Staging Schema).

2.2.3.2 Functional Schema (Clinical Warehouse)

A functional schema was created by the name *Clinical_Warehouse* in which tables were created based on the proposed dimensional model (Figure 4). SQL queries for creation of tables for staging database can be accessed from *Appendix I - A2.* The data from the Clinica_Staging_Data database, was subjected to cleaning and conversion process so that its approriate form can be stored in the functional schema which can be used further for analysis and to find correlation. Appropriate mappings designs were developed to process and store the data based on the dimensional model designed. For example values of a particular diagnostic test coming from different sources should be in a common format and stored in Dim_Diagnostics_Test table. The tables like patient dimension have only selected attributes so we accordingly use appropriate function to load the required attributes in the table.

TABLE NAME (FUNCTIONAL SCHEMA)	TYPE OF DATA STORED
Date dimension (DIM_DATE)	This tables stores the date information which have been processed from the 'date' staging table.
Time dimension (DIM_TIME)	This tables stores the date information which have been processed from the 'time' staging table.
Disease dimesnion (DIM_DISEASE)	This table contains description for all the diseases associated to humans. Each of the disease have been assigned a unique id. Data processed from 'disease' staging table.
Diagnostic test dimension (DIM_DIAGNOSTIC_TEST)	This table contains details for all the diagnostic tests that a patient gets done e.g. platelet count, TCL, KFT, LFT etc. All the

	tests have been given a unique id. Data processed from 'diagnostic test' staging table.
Patient dimension (DIM_Patient)	Patient's personal details such as name, patient_id , age, sex etc.. are stored in this table. For each patient an unique id is being autogenerated (given as patient id) whenever information is entered for the first time. Data processed from 'patient' staging table.
Patient fact table (FACT_PATIENT)	This table stores the patient id and corresponding measuements corresponding to a particular disgnostic test with a disease conducted on a particule date & time. It helps to store all the historical information corresponding to any number of tests being conducted for a patient.
Image dimension (DIM_IMAGE)	Stores descriptive information of various types of medical images.
Patient Image Table (FACT_PATIENT_IMAGE_DETAIL)	All the numerical parameters associated with a patient's image are stored in this table.

Table II - Details of Clinical_Datawarehouse (Functional Schema).

2.2.2.3 ETL mappings (using Kettle 3.1)

The process of cleaning and transforming data is known as ETL, or Extraction, Transformation, and Loading. Proper care of the data is an important part of maintaining a successful data warehouse. Kettle is an open source ETL (Extraction, Transformation and Loading) tool. The product name is spelled as K.E.T.T.L.E, which is a recursive acronym for "Kettle Extraction, Transport, Transformation and Loading Environment". It's a platform-independent ETL tool by Matt Casters [11]. Being an ETL tool, Kettle is an environment that's designed to:

- Collect data from a variety of sources (extraction)
- Move and modify data (transport and transform) while cleansing, denormalizing, aggregating and enriching it in the process.
- Frequently store data (loading) in the final target destination, which is usually a large, dimensionally modeled database called a data warehouse.

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

Figure 4 depecits how flow of data from various source files into the clinical data mart is managed using ETL codes in Kettle. Table III enlists description of ETL mappings for processing of data from source files into respectiv tables of staging schema (Cancer_Staging_Data). Mappings can be accessed from *Appendix I - B1*.

ETL Mapping Name	Source	Target	Operators Used
Staging_date	Date.csv	Date	CSV file input, Table Output
Staging_time	Time.csv	Time	CSV file input, Table Output
Staging_patient	CSV files based on data of patients obtained from different hospitals	Patient	CSV file input, Get File Name, Target
Staging_disease	Disease.csv	Disease	CSV file input, Table Output
Staging_test	Test.csv	Diagnostic_test	CSV file input, Table Output
Staging_image	Image files alongwith numerical measures in csv files	Image	CSV file input, Table Output

Table III - ETL mapping description for processing data from source files into Cancer_Staging_Data.

Table IV enlists description of ETL mappings for processing of data from tables of Clinical_Staging_Data staging schema into respective tables of Clinical_Warehouse. Mappings can be accessed from *Appendix I - B2*.

ETL Mapping Name	Source	Target	Operators Used
DW_dim_date	Date	DIM_DATE	Table Input, Add Sequence, Table Output
DW_dim_time	Time	DIM_TIME	Table Input, Add Sequence, Table Output
DW_dim_disease	Disease	DIM_DISEASE	Table Input, Add Sequence, Data Validator, Table Output
DW_dim_test	Diagnostic_test	DIM_DIAGNOSTIC_TEST	Table Input, Add Sequence, Data Validator, Table Output
DW_dim_patient	Patient	DIM_PATIENT	Table Input, Add Sequence, Select Values, Table Output
DW_fact_patient	Patient	FACT_PATIENT	Table Input, Select Values, DB Lookup, Data Validator, Table Output
DW_dim_image	Image	DIM_IMAGE	Table Input, Add Sequence, Select Values, Table Output
DW_fact_image	Image	FACT_PATIENT_IMAGE_ DETAIL	Table Input, Select Values, DB Lookup, Data Validator, Table Output

Table IV - ETL mapping description for processing data from Cancer_Staging_Data (Staging Schema) into Clinical_Warehouse (Functional Schema)

2.3 Results

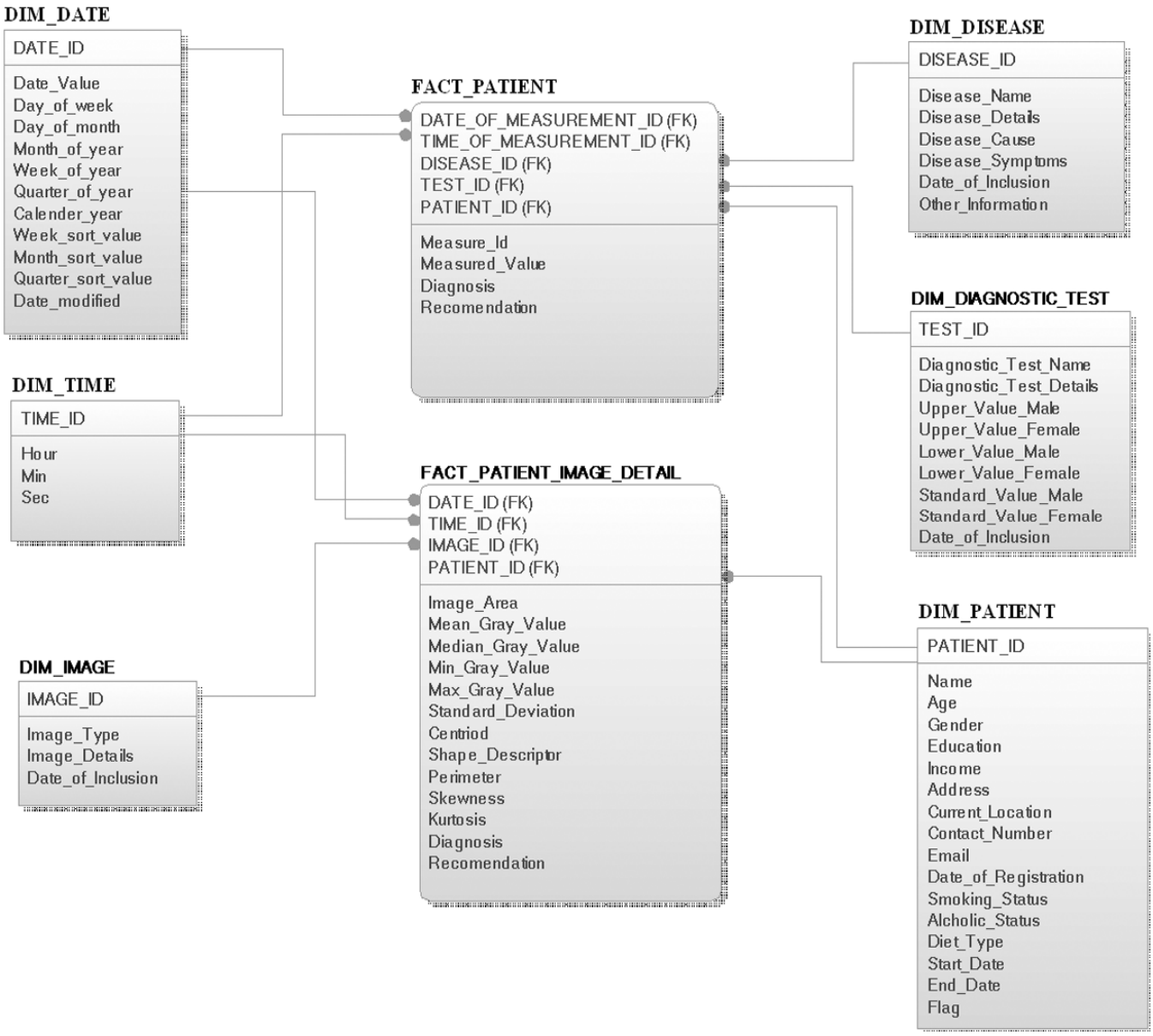


Figure 5 - Logical representation of clinical dimensional model
(logical data model for the clinical data mart)

The logical design form of the dimensional model (figure 5) is in a star schema representation [13]. Proposed design 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 that is provided to each patient and this is the id which is majorly linking all other information related to that patient. 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 and so it can act as slowly changing dimension (SCD) [14-16]. During its physical implementation for a data mart an SCD-Type II implementation can be made for Dim_Patient dimension [14]. While designing the clinical dimensional model the temporal based storage prospect was taken into consideration, henceforth Date and Time dimension are 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. 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 respective 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.

2.4 Discussion

Interpreting data across multiple systems has been always challenging, and various integration techniques, with varying levels of complexity, have been proposed in the past to solve the problem of data integration and storage [17-20]. Nagarajan *et al.* [20] identified the potential utilization of solutions using relational database management systems (RDBMSs) for assembling and integrating the data for data-warehousing-based solutions. A relational database model is composed of classes of data, with each class characterized by a set of attributes. This conventional design is ideal for data sets composed of classes with a limited and fixed number of attributes. When each instance has values for all attributes (or columns) within a class (or table), then the database is not filled with numerous null entries and memory is used efficiently. However, research

has revealed that this design is not effective for data sets with large numbers of attributes that vary taking into consideration the time dimensionality [21]. Some of the researches propose use of knowledge-based terminology for identifying data dimensions in clinical informatics [22] and on the conceptual development of IDs using ontology-based systems for the design and integration of clinical data [23]. The inherent variation between databases due to the demands on each system means that there is no consensus on ontology and metadata descriptions. It might therefore be necessary to define a new ontology for each database. Although this approach gives the database designer freedom at the outset, inexperienced designers can spend excess time in researching previous knowledge, seeking an optimum design. Where possible, designers should use pre-existing ontologies. These can be modified as necessary to improve accessibility. The Bio-mediator system provide a theoretical and practical foundation for data integration across diverse biomedical domains via a "knowledge-base-driven centralized federated database" model [21]. However, the efficiency of query processing time and the need to filter out unnecessary query results still are concerns. The data architecture required for clinical data warehousing has been researched in applications such as clinical study data management systems (CDMSs) and clinical patient record systems (CPRSs). They both use an entity-attribute-value (EAV) system (i.e., row modeling) as opposed to conventional database design [22]. The EAV system has the advantage of remaining stable as the number of parameters increases when knowledge expands, a common situation in the basic sciences and in clinical trials [23]. The characteristics of clinical data as it originates during the process of clinical documentation, including issues of data availability and complex representation models, can make data mining applications challenging. Data preprocessing and transformation are required before one can apply data mining to clinical data.

The lacunae's reported can be addressed to an extent by the proposed clinical dimensional model. Further the data storage structure formed would acts as a data collector, data integrator and data provider in the data mining process that could be used by doctors, physicians and other health professionals. The application of classical data warehousing process should be thus able to answer the queries being raised and also be able to mitigate issues like appropriate storage structure of clinical data, able to handle varied sources of data, reduce the dimensionality constraint, and handling of multiple data variables. The data mart for clinical data should be able to render the data in appropriate structures, provide metadata that adequately records syntax/semantics of data and reference pertinent medical knowledge.

2.5 Conclusion

Clinical Informatics is one of the most versed fields and new IT solutions are being designed for its effective management. However, there still a gap in the effective storage solution along with techniques for correlation of the data. We dream of an era in which all the genetic (genomic and proteomic) information of an individual along with drugs data will be correlated with his/her clinical and information aspect.

References

1. Kimball, R. *The Data Warehouse Lifecycle Toolkit*. John Wiley & Sons, Inc., New York. 1998.
2. Kimball, R. & Ross, M. *The Data Warehouse Toolkit, 2nd edition*. John Wiley & Sons, Inc., New York. 2002.
3. Gutierrez, A. & Marotta, A. *An Overview of Data Warehouse Design Approaches and Techniques*. Instituto de Computación, Facultad de Ingenieria, Universidad de la República, Montevideo, Uruguay. 2000.
4. Viangteeravat, T., Brooks, I.M., Smith, E.J. et al. *Slim-prim: a biomedical informatics database to promote translational research*. Perspectives in Health Information Management, 6:6. 2009.
5. Corey, J.M., Abbey, M., Abramson, I. & Taub, B. *Oracle 8 Data Warehousing – A Practical Guide to Successful Data Warehouse Analysis, Build, and Roll-Out*. Osborne/McGraw-Hill, Berkeley. 1998.
6. Ross, M. *The 10 Essential Rules of Dimensional Modeling - Kimball Group*. Internet: http://www.kimballgroup.com/2009/05/29/the-10-essential-rules-of-dimensional-modeling/. 2009.
7. Leitheiser, R.L. *Data quality in health care data warehouse environments. In* Proc. of the 34th Hawaii International Conference on System Sciences. Hawaii: System Sciences; 2001.
8. Marshall, W.J. *Clinical Biochemistry: Metabolic and Clinical Aspects.* (2nd ed.). Churchill Livingstone. 2008.
9. Gliklich, R.E. & Dreyer, N.A. *Data Collection and Quality Assurance, Registries for Evaluating Patient Outcomes: A User's Guide*. 2nd edition. Rockville: AHRQ Publication No.10-EHC049. 2010.
10. CA, Tech. *ERWIN DATA MODELER (data modeling software system), version 8.2*. http://erwin.com/products/data-modeler. 2012.
11. Pentaho Corporation. *Pentaho Data Integration, Kettle 3.1*. http://www.kettle.pentaho.com. 2011.
12. MySQL. *MySQL 5.019.* http://dev.mysql.com/downloads/. 2010.
13. Mozes, A. *Mining Star Schemas A Telco Churn Case Study*. Internet: http://www.oracle.com/technetwork/database/options/advanced-analytics/odm/odmtelcowhitepaper-326595.pdf. 2011.
14. Kimball, R. *Slowly Changing Dimensions, Types 2 and 3*. Kimball Group. Internet: http://www.kimballgroup.com/2008/09/22/slowly-changing-dimensions-part-2/. 2009.
15. Browning, D. & Mundy, J. *Data Warehouse Design Considerations.* Internet: http://msdn.microsoft.com/en-us/library/aa902672%28v=sql.80%29.aspx. 2001

16. Lunexa. *White Paper on Slowly Changing Dimensions.* Internet: http://www.lunexa.com/documents/10923/12405/SlowlyChangingDimensions.pdf. 2011.
17. Brazhnik, O. & Jones, J. *Anatomy of Data Integration*. Journal of Biomedical Informatics, 40(3) pp 252–269. 2007.
18. Geisler, S., Brauers, A., Quix, C. & Schneink, A. *Ontology-based System for Clinical Trial Data Management*. Proceedings: Annual Symposium of the IEEE/EMBS Benelux Chapter Heeze, the Netherlands, pp 53-55. 2007.
19. Wang K *et al* . *BioMediator Data Integration: Beyond Genomics to Neuroscience Data*. AMIA Annual Symposium Proceedings, pp 779–783. 2005.
20. Nagarajan, R., Ahmed, M. & Phatak, A. *Database Challenges in the Integration of Biomedical Data Sets*. Proceedings of the Thirtieth International Conference on Very Large Data Bases, Toronto: VLDB Endowment, 2004.
21. Dinu, V. & Nadkarni, P. *Guidelines for the Effective Use of Entity–Attribute–Value Modeling for Biomedical Databases*. International Journal of Medical Informatics, 76(11) pp 769–779. 2007.
22. Deshpande, A.M., Brandt, C. & Nadkarni, P.M. *Metadata-driven Ad Hoc Query of Patient Data*. Journal of the American Medical Informatics Association, 9 pp 369–382. 2009.
23. Anhøj, J. *Generic Design of Web-Based Clinical Databases*. Journal of Medical Internet Research, 5(4) e27. 2003.

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.

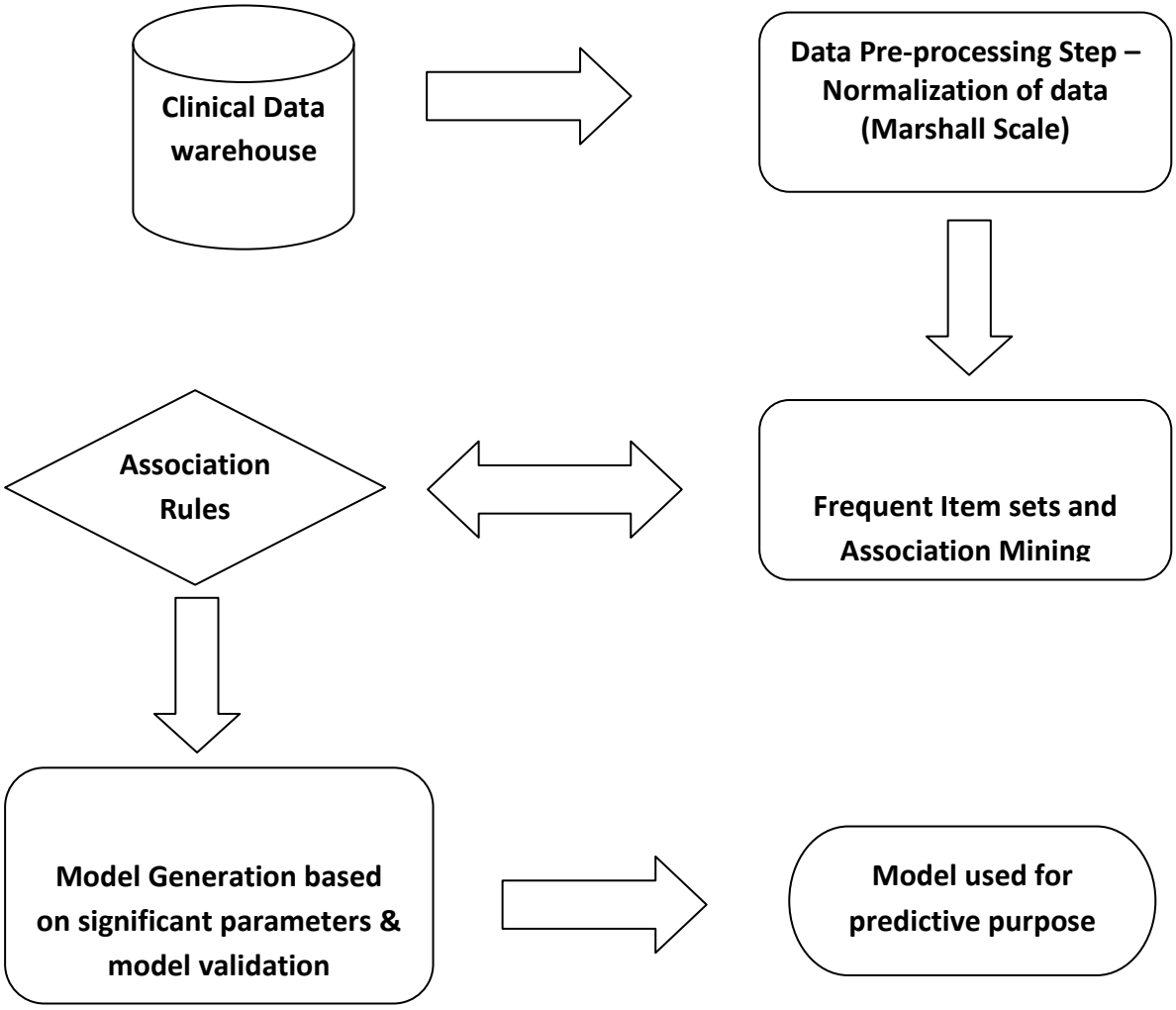
The stored data in the warehouse would provide a basis for the analysis of risk factors for the disease. For example, we can compare tumor with non-tumor patients to find patterns associated with occurrence of brain tumor. This method has been common practice in evidence-based medicine, which is an approach 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 $X \subset I$, 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 STATSTICA 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:

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

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 $\alpha = 0.04$.

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.

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 V (A) - Association Rules deciphered for clinical parameters corresponding to occurrence of brain tumor (Min. Support - 50%; Confidence - 85%)

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 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 Biliuribin), a predictive model is proposed to predict the possible STATE of a individual i.e whether suffering from tumor or not. Most significant model was obtained at $\alpha = 0.04$ & is represented by the equation ii:

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

... 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

1. Li, J., Fu, A.W. & Fahey P. *Efficient discovery of risk patterns in medical data*. Artificial Intelligence in Medicine, 45(1) pp 77-89. 2009.
2. Agrawal, R., Imielinski, T. & Swami A. *Mining association rules between sets of items in large databases*. Proceedings of ACM SIGMOD international conference on management of data. New York, pp 207-216. 1993.
3. Brossette, S.E., Sprague, A.P., Hardin, J.M., et al. *Association rules and data mining in hospital infection control land public health surveillance*. Journal of the American Medical Informatics Association, 5 373-381. 1998.
4. Ohsaki, M., Sato, Y., Yokoi, H. & Yamaguchi, T. *A rule discovery support system for sequential medical data in the case study of a chronic hepatitis dataset*. Proceedings of the ECML/PKDD-2003 discovery challenge. 2003.
5. Paetz, J. & Brause R.W. *A frequent patterns tree approach for rule generation with categorical septic shock patient data.* Proceedings of the second international symposium on medical data analysis; London: Springer-Verlag, pp 207-12. 2001.
6. Agrawal, R. & Srikant, R. *Fast algorithms for mining association rules*. Proceedings of 20th international conference on very large data bases. San Mateo, Santiago, Chile: 20th VLDB Conference, pp 487-499. 1994.
7. Chen, E.S., Hripcsak, G., Xu, H., et al. *Automated acquisition of disease drug knowledge from biomedical and clinical documents: An initial study*. Journal of the American Medical Informatics Association, 15(1) pp 87-98. 2008.
8. Hanauer, D.A., Rhodes, D.R. & Chinnaiyan, A.M. *Exploring clinical associations using '-omics' based enrichment analyses*. PLoS One, 4(4) e5203. 2009.
9. Wright, A., Chen, E.S. & Maloney, F.L. *An automated technique for identifying associations between medications, laboratory results and problems*. Journal of Biomedical Informatics, 43(6) pp 891-901. 2010.
10. Ordonez, C., Ezquerra, N.F. & Santana, C.A. *Constraining and summarizing association rules in medical data*. Knowledge and Information Systems, 3 pp 1-2. 2006.
11. StatSoft, Inc. *STATISTICA (data analysis software system), version 9.1*; www.statsoft.com. 2010.
12. Marshall, W.J. *Clinical Biochemistry: Metabolic and Clinical Aspects.* (2nd ed.). Churchill Livingstone. 2008.

13. Borgelt, C. *Simple algorithms for frequent item set mining advances in machine learning* Springer Berlin / Heidelberg, pp 351-369. 2010.
14. Goethals, B. *Survey on frequent pattern mining*. University of Helsinki. 2003. Internet: http://adrem.ua.ac.be/~goethals/software/survey.pdf

CHAPTER - 4

SN ALGORITHM:

ANALYSIS OF TEMPORAL CLINICAL DATA FOR MINING PERIODIC PATTERNS AND IMPENDING AUGURY

4.1 Introduction

Crucial to mining in clinical informatics is to use background knowledge to discover interesting interpretable and non-trivial relationships, to construct rule-based and other symbolic-type models that can be reviewed and scrutinized by experts, to discover models that offer an explanation when used for prediction and, also to bridge model discovery and decision support to deploy predictive models in daily clinical practice [1]. Among the various mining approaches, predictive data mining approach is gaining impulse among the researchers and clinical practitioners as it utilizes the knowledge available in the clinical domain and explains proposed decision for the proposed model [1]. The goal of predictive data mining in clinical medicine is to derive models that can use patient specific information to predict the outcome of interest and thereby support clinical decision-making [1]. Among the various approaches, Naive Bayesian classifier is one of the earliest designed approach that is based on probability. It is one of the simplest yet a useful and often a fairly accurate predictive data mining method. However, since it is dependent on the type of data subjected to mining, it may be inclined in case of biased clinical data set [2]. Another popular data mining technique is decision tree which is based on recursive data partitioning, where in each iteration the data is split according to the value of a selected clinical attribute. However, its performance is impacted because of clinical data segmentation [3]. Logistic regression is another powerful and well-established statistical method used in predictive clinical mining. It is an extension of normal regression method that models a two-valued outcome for occurrence or non-occurrence of some event. It is based on multiplicative probability model that utilizes maximum likelihood estimation to determine the coefficients in its probability formula. Handling of the missing values usually causes problem in this approach [4]. For a long period artificial neural network models were the most popular artificial intelligence-based predictive algorithm used in clinical medicine. Albeit they have a number of deficiencies that include high sensitivity to the parameters of the method - including those that determine the architecture of the network and induction of the model that may be hard to interpret by domain experts [5]. Support vector machines (SVM) are perhaps today's most powerful classification algorithm in terms of predictive accuracy and most popular in clinical informatics. However, the exception are linear kernels, where the structure of the model can be easily revealed through the coefficients that define a linear hyperplane, and it use a formalism that is often unsuitable for interpretation by human experts [6].

An interesting prospective in these predictive mining of heterogeneous clinical data would be an approach that could analyze the temporal form. The discovery of hidden periodic patterns in temporal data, apart from unveiling important information, can facilitate data management

substantially [7]. 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 [8]. 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 timestamps [9]. 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 [10]. 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. Considering the given scenario, SN algorithm proposed in this study, is a novel predictive data mining algorithm based on Jacobian approach. It will traverse selective clinical parameters at different temporal points to augur the possible "STATE" of the disease. The advantage of this algorithm over existing predictive techniques like logistic regression or ANN or SVM is that it is independent of coefficients for prediction. Moreover, it keeps a track of previous versus new information i.e. for a given patient it predicts the corresponding state of the disease based on the value of input clinical parameters along with the state of the disease at previous temporal point.

In this study we have defined the temporal mining problem of clinical data in terms of (a) discovery of associative rules for clinical parameters, which can be associated with a specific disease (clinical parameters are discovered by apriori association mining); and (b) an algorithm for traversing the clinical parameters of temporal points 'T' (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 (we termed it as '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' (J_0, J_1…) of temporal points 'T' (T_0, T_1 …) along with respective states 'S' (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).

4.2 SN (Sengupta and Naik) Algorithm

The proposed SN algorithm is being designed for traversing across the clinical measures of a patient pertaining to particular disease at varied temporal points and augur the possible "STATE"

of that disease. The state of temporal point 'T_n' is obtained as Jacobian determinant for cross product of derivatives of selected clinical parameters for 'T_n' and its immediate predecessor point. The clinical parameters are selected for the disease based on the existing knowledge or associative rules deciphered from mining process. The selected clinical parameters acts as base point for SN algorithm to extrapolate the progression of disease at given time point 'T_n'.

In detail the algorithm consists of 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 [11]. 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 iii & iv:

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

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

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 equation v,

$$dx * dy = < 0,0 \begin{vmatrix} xu & xv \\ yu & yv \end{vmatrix} > dudv \quad \text{... equation v}$$

from which the differential area *dA* can be obtained as equation vi:

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

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).

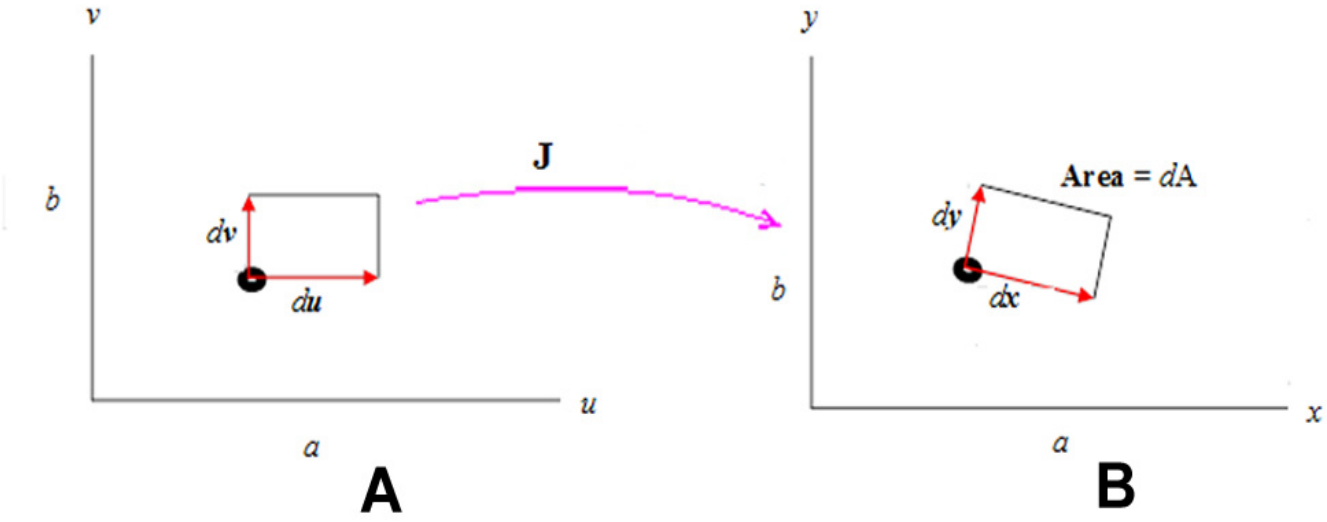
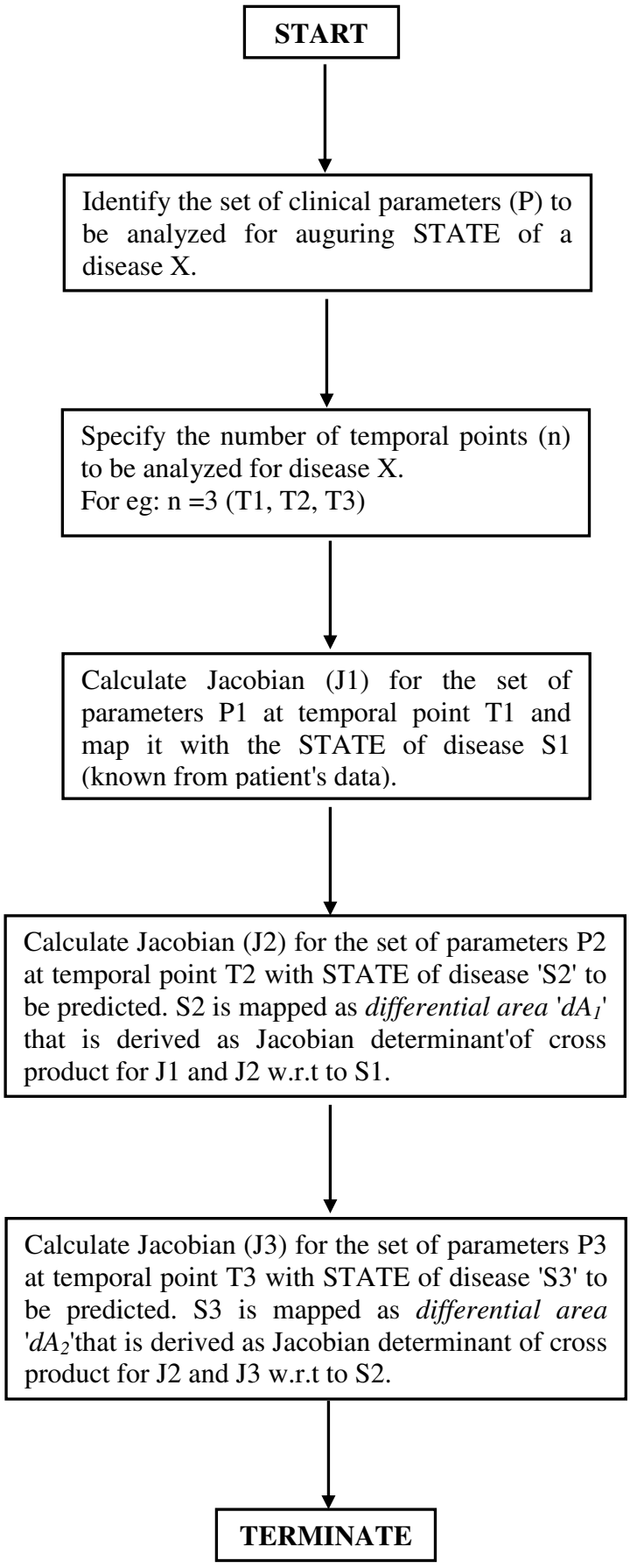


Figure 6 - Area differential approach based on Jacobian Transformation
(A - Temporal point 1; B - Temporal point 2)

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



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

4.3 Analysis of SN algorithm

To test the predictability of SN algorithm we have taken a temporal case study of 55 patients suffering from brain tumor over a period of 6 months. The clinical data was collected from various Government Hospitals in India and stored in the in-house developed data warehouse. The data mining process involves two stages. In the first phase, brain tumor was treated as a state and was analyzed corresponding to investigating parameters of blood analysis, KFT (Kidney Functionality Test), LFT (Liver Functionality Test), sugar level, triplets of blood pressure and MRI/CT scan images. Association rule mining was applied to this dataset using STATSTICA DATAMINER. The set of rules deciphered from association mining (Table VA, Chapter 3) with 85% confidence and atleast 50% support criteria suggests that creatinine, blood urea nitrogen (BUN), SGOT and SGPT are the clinical diagnostic parameters which can be associated with occurrence of brain tumor in patients.

In the second phase of the study, SN algorithm was applied over 3 temporal state points T(T1, T2, T3) for each patient P(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 VI. Observed State S'_o(S'1, S'2, S'3) pertaining to each temporal point T(T1, T2, T3) for each patient were determined based on CT/MRI results and diagnosis/recommendation of oncologist. Certainty of the algorithm has being analyzed by the accuracy factor that is based on the observed state "S'_o" and Predicted State "S_o".

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 VI - 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 predicted 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 (S'1 is selected to map the initial state of disease at first temporal point i.e. S1 = S'1) as area curve. J1 (c1,b1,s1,g1) is the Jacobian for patient 'P1′ at time 'T1' that is mapped to the state of the disease 'S1′. Similarly for the second

temporal point 'T2′, Jacobian J2 (c1′, b1′, s1′, g1′) is to be mapped with S2 (represented by area *dA*) for patient 'P1′. Based on the cross product of Jacobian for point T1 and T2, the differential area '*dA'* is mapped as Jacobian determinant to obtain S2 state. The accuracy of predicted state S2 based on SN algorithm was 100% when compared to observed S'2 state (Table VII).

Patient_ID	**State T1** (Known from Patient Data)	**Actual State for T2** (Known from Patient Data)	**Predicted State for T2** (Calculated via SN algorithm)	**Prediction Status**
1	1.0	1.3	1.26	Correct
2	1.0	1.3	1.26	Correct
3	1.0	1.4	1.43	Correct
4	1.0	0.9	0.91	Correct
5	1.0	1.2	1.22	Correct
6	1.0	1.3	1.29	Correct
7	1.0	1.5	1.48	Correct
8	1.0	0.8	0.78	Correct
9	1.0	0.7	0.69	Correct
10	1.0	0.9	0.92	Correct
11	1.0	0.7	0.69	Correct
12	1.0	0.8	0.81	Correct
13	1.0	0.9	0.92	Correct
14	1.0	1.3	1.29	Correct
15	1.0	1.6	1.64	Correct
16	1.0	1.6	1.62	Correct
17	1.0	1.5	1.54	Correct
18	1.0	1.3	1.29	Correct
19	1.0	1.1	1.07	Correct
20	1.0	1.0	1.03	Correct
21	1.0	1.2	1.20	Correct
22	1.0	1.0	1.02	Correct
23	1.0	1.0	1.04	Correct
24	1.0	1.4	1.39	Correct
25	1.0	1.3	1.29	Correct
26	1.0	0.9	0.88	Correct

27	1.0	0.7	0.72	Correct
28	1.0	0.7	0.69	Correct
29	1.0	0.7	0.69	Correct
30	1.0	1.3	1.29	Correct
31	1.0	1.2	1.23	Correct
32	1.0	1.5	1.51	Correct
33	1.0	1.4	1.42	Correct
34	1.0	1.3	1.31	Correct
35	1.0	1.1	1.09	Correct
36	1.0	1.1	1.11	Correct
37	1.0	1.1	1.09	Correct
38	1.0	0.7	0.69	Correct
39	1.0	0.8	0.81	Correct
40	1.0	0.8	0.81	Correct
41	1.0	1.3	1.32	Correct
42	1.0	0.8	0.78	Correct
43	1.0	0.9	0.93	Correct
44	1.0	1.4	1.38	Correct
45	1.0	1.4	1.39	Correct
46	1.0	1.4	1.39	Correct
47	1.0	1.4	1.41	Correct
48	1.0	1.5	1.51	Correct
49	1.0	1.2	1.23	Correct
50	1.0	0.7	0.73	Correct
51	1.0	0.8	0.81	Correct
52	1.0	0.8	0.81	Correct
53	1.0	0.9	0.89	Correct
54	1.0	1.0	0.99	Correct
55	1.0	1.0	0.99	Correct

Table VII - Prediction at Temporal Point T2

Predicted State is being represented till 2nd decimal point. A **5% margin for error rate** is being considered in assessing the prediction status by comparing Actual and Predicted T2 state.

No. of True Positives (Correct Predictions) for T2 state = 55

No. of False Negatives (Incorrect Predictions) for T2 state = 0

Accuracy for T2 state = 100%

However, for the third temporal point only Jacobian J3 (c1", b1", s1", g1") for the parameters was obtained and S'3 result was in a hidden state. To obtain the S3 predicted state, differential area was mapped as Jacobian determinant based on cross products of Jacobian for points T2 and T3. Predictability of the S3 state with the hidden S'3 state was 92.7% accurate (Table VII).

Patient_ID	**State T2** (From T2 prediction)	**Actual State for T3** (Known from Patient Data)	**Predicted State for T3** (Calculated via SN algorithm)	**Prediction Status**
1	1.26	1.4	1.43	Correct
2	1.26	1.5	1.47	Correct
3	1.43	1.4	1.4	Correct
4	0.91	0.8	0.77	Correct
5	1.22	1.1	1.09	Correct
6	1.29	1.0	1.04	Correct
7	1.48	1.4	1.39	Correct
8	0.78	0.8	0.78	Correct
9	0.69	0.7	0.68	Correct
10	0.92	0.9	0.91	Correct
11	0.69	0.6	0.61	Correct
12	0.81	0.7	0.72	Correct
13	0.92	0.8	**0.87**	**Incorrect**
14	1.29	1.2	1.19	Correct
15	1.64	1.6	1.57	Correct
16	1.62	1.5	1.51	Correct
17	1.54	1.4	1.39	Correct
18	1.29	1.3	1.28	Correct
19	1.07	1.0	**0.92**	**Incorrect**
20	1.03	1.0	1.02	Correct
21	1.20	1.1	1.09	Correct
22	1.02	0.8	0.83	Correct
23	1.04	0.8	0.84	Correct

24	1.39	1.3	1.29	Correct
25	1.29	1.3	1.29	Correct
26	0.88	0.8	0.81	Correct
27	0.72	0.7	0.71	Correct
28	0.69	0.7	0.7	Correct
29	0.69	0.6	0.58	Correct
30	1.29	1.2	1.19	Correct
31	1.23	1.2	1.21	Correct
32	1.51	1.4	1.42	Correct
33	1.42	1.3	1.31	Correct
34	1.31	1.1	1.08	Correct
35	1.09	0.8	0.83	Correct
36	1.11	1.0	1.02	Correct
37	1.09	1.3	**1.36**	**Incorrect**
38	0.69	0.6	0.61	Correct
39	0.81	0.7	0.69	Correct
40	0.81	0.7	**0.76**	**Incorrect**
41	1.32	1.3	1.29	Correct
42	0.78	0.9	0.87	Correct
43	0.93	0.9	0.94	Correct
44	1.38	1.2	1.17	Correct
45	1.39	1.2	1.24	Correct
46	1.39	1.3	1.27	Correct
47	1.41	1.1	1.13	Correct
48	1.51	1.4	1.39	Correct
49	1.23	1.0	0.98	Correct
50	0.73	0.6	0.57	Correct
51	0.81	0.7	0.69	Correct
52	0.81	0.7	0.69	Correct
53	0.89	0.9	0.87	Correct
54	0.99	1.0	1.01	Correct
55	0.99	1.0	1.01	Correct

Table VIII - Prediction at Temporal Point T3

Predicted State is being represented till 2nd decimal point. A **5% margin for error rate** is being considered in assessing the prediction status by comparing Actual and Predicted T3 state.

No. of True Positives (Correct Predictions) for T3 state = 51

No. of False Negatives (Incorrect Predictions) for T3 state = 04

Accuracy for T3 state = 92.7%

Overall Accuracy observed (T2 and T3) = 96.35% ~ 97%

Thus, the proposed algorithm is helping in auguring the state of disease for brain tumor patients, independent of results from MRI, CT scan, arteriogram or small dime craniotomy based on temporal values for Creatinine, BUN, SGOT & SGPT clinical parameters.

Analyzing the time complexity of the proposed SN algorithm will be essential to evaluate its robustness. Master method is been applied to estimate the time complexity which can be associated with proposed SN algorithm. The time complexity has been calculated in terms of Big O notation given as:

$$T(n) = O(f(n)) \qquad \text{... equation vii}$$

The expected running time (d) for this algorithm is directly dependent to number of sub-problems (a) to be analyzed which is the number of temporal points for a particular case, considering the shrinkage factor (b) to be greater than 1. Henceforth, for the given algorithm, as observed: $a = b^d$, the time complexity associated with the algorithm can be estimated as:

$$O = n \log n \qquad \text{... equation viii}$$

Where, n -> number of temporal points analyzed for given set of parameters.

The estimated time complexity of the proposed SN algorithm suggests minimal execution time for auguring the STATE of disease at a particular temporal point. However, increasing the number of temporal point will directly proportionate the execution time.

4. 4 Conclusion

In this study, temporal mining problem associated with clinical data was raised as a research problem, corresponding to which SN algorithm has been proposed. The algorithm is based on Jacobian and mapping of its derivative as area. The accuracy of the algorithm was evaluated using a data set of 55 patients suffering from brain tumor. Using this algorithm we have achieved 100% accuracy in predicting the progression of brain tumor at 2nd temporal point by mapping with the Jacobian derivative of 1st temporal point. In contrast, we have predicted the disease progression

with an accuracy of 92.7% at 3rd temporal point based on 2nd temporal point. 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. However, 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.

References

1. Bellazzi, R. & Zupanb, B. *Predictive data mining in clinical medicine: current issues and guidelines.* International Journal of Medical Informatics, 77 pp 81-97. 2008.
2. Kononenko, I. *Inductive and Bayesian learning in medical diagnosis*. Applied Artificial Intelligence, 7 pp 317-337. 1993.
3. Breiman, L. *Classification and Regression Trees*. Boca Raton: Chapman & Hall. 1993.
4. Hosmer, D.W. & Lemeshow, S. *Applied Logistic Regression, (2nd ed.)*. New York: Wiley. 2000.
5. Schwarzer, G., Vach, W. & Schumacher, M. *On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology.* Statistics in Medicine, 19 pp 541-561. 2000.
6. Cristianini N, Taylor JS: *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge: Cambridge University Press. 2000.
7. Mamoulis, N., Cao, H., Kollios, G., et al. *Mining, Indexing, and Querying Historical Spatiotemporal Data*. Knowledge discovery and data mining, pp 236-245. 2004.
8. Peng, W.C. & Chen, M.S. *Developing data allocation schemes by incremental mining of user moving patterns in a mobile computing system*. IEEE Transactions on Knowledge and Data Engineering, 15(1) pp 70–85. 2003.
9. Indyk, P., Koudas, N. & Muthukrishnan. *Identifying representative trends in massive time series data sets using sketches*. In Proc. of Very Large Data Bases, pp 363-372. 2000.
10. Ma, S. & Hellerstein, J.L. *Mining partially periodic event patterns with unknown periods*. In Proc. of International Conference on Data Engineering 205-214. 2001.
11. Kinsley, J. *Multivariable Calculus Online adapted from Calculus: A Modern Approach*. Internet: http://math.etsu.edu/multicalc/prealpha/. 2012.

CHAPTER - 5

IDENTIFICATION OF KEY MEASURES FOR EVALUATION OF COGNITIVE PERFORMANCE AT HIGH ALTITUDE

5.1 Introduction

Hypobaric hypoxia at high altitude can cause the loss of memory, recall and learning [1] resulting to cognitive impairment in addition to the acute mountain sickness (AMS), high altitude pulmonary edema (HAPE), high altitude cerebral edema (HACE) and neurophysiological disturbances with insomnia and dizziness [2]. The neurological symptoms are probably due to lack of proper oxygen supply to brain that can alter neurotransmitter synthesis, uptake and release [3,4] and free radical generation and related excitotoxic neuronal damage [5,6]. Changes in gene expression and protein functions are also associated with hypoxia and ischemia [7-10]. Ascent to high altitude also increases sensory discrimination (latency of P3 component), delay in evaluation process and impairment of short term memory [11]. Chronic hypoxia exposure of volunteers residing at high altitude also revealed impairment in verbal working memory [12].

Mild cognitive impairment (MCI) is an early stage of cognitive impairment and is characterized by cognitive domain tests. Such impairment (MCI) is observed in young and healthy population residing at altitudes above 4,300 m in the trans-Himalayan regions for longer than 12 months and screened by multi-domain cognitive screening test (MDCST) [13], Mini Mental State Examination (MMSE) [14], Montreal cognitive Assessment (MoCA) [15], Mini-Cog [16], Computer Administered Neuropsychological Score (CANS-MCI) [17] and Patient Reported Outcomes in Cognitive Impairment (PROCOG) [18]. A recent study showed MDCST to be a more effective cognitive measure for MCI assessment in demographic studies compared to the traditional measures because it exhibited excellent psychometric properties in terms of sensitivity and test-retest reliability [13]. Other findings also showed increased prevalence of MCI in acclimatized lowlander population during prolonged stay at high altitude which is neurophysiologically distinct from MCI leading to Alzheimer's disease. There is a number of screening instruments to identify mild cognitive impairment of an individual, but they lack the required specificity and sensitivity for Indian population. Some specifically screen for deficit of domain while others are bulky and time consuming. In addition to these limitations these screening instruments have complex scoring system that needs expertise of a trained person. MDCST, on the other hand, 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. Keeping in mind the limitations of MMSE and MoCA, MDCST was designed to establish an easy to administer and a more reliable test for detection of MCI at early stages [13]. Moreover, it is based on findings that suggest involvement of several brain regions in cognitive function. 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, a analysis of Beck Depression Inventory (BDI) with insomnia can help in identification of hitherto undetected depression.

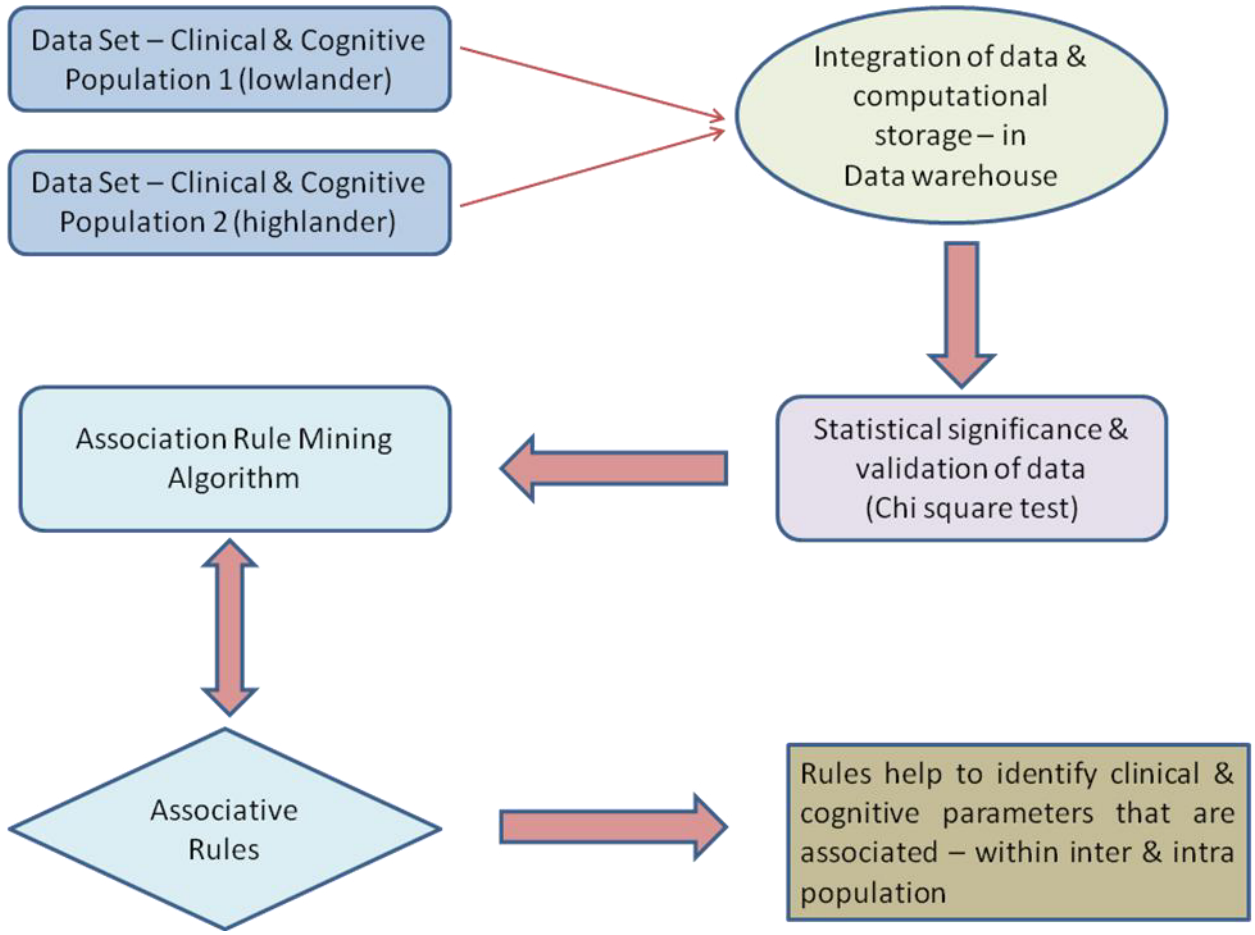
However, there has not been any demographic study to measure the prevalence of MCI that can be associated along with Beck Depression Inventory (BDI), insomnia, and clinical parameters such as, blood glucose level, blood pressure, blood cholesterol, kidney and liver function. A rule based study can allow 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). It has been a common practice in clinical practice to ignore the rigor and sophistication of a data mining method, but rather focus on results however they were obtained [19]. "Rules" are of the most human-understandable knowledge systems, and therefore most suitable for deciphering new rules corresponding to data associated with medical applications. Association rule mining [20,21] is a general purpose rule discovery scheme of finding disease co-occurrences in electronic health record data that has been widely used to find disease-disease, disease-finding, and disease-drug co-occurrences [22,23].

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.

5.2 Materials & Methods

5.2.1 Collection of data

This study is based on the field research of High altitude physiology division, Defence Institute of High Altitude Research, DRDO, Leh, India that was conducted during August 2009 - January 2011, after obtaining ethical clearance from the institutional ethics committee of Defence Institute of High Altitude Research [13]. The approach used for this study has been demonstrated in flow diagram 3.



Flow Diagram 3 - Representation of knowledge discovery process

(identification of key evaluation parameters for MCI)

The clinical tests along with MDCST were performed on group of volunteers in age group of 25-40 comprising lowlanders and highlanders located at high altitude (≥ 4300m) after informing them about the study purpose, protocol, and expected outcomes [13]. It is a longitudinal follow-up of a cohort that ascended to high altitude and lived there for the duration of 18 months. Baseline recordings was done in the beginning of study at low altitude which as a control group of its own [13]. Table IX enlists the inclusion criteria of human subjects for this study.

Parameter	**Value**
Age (in yrs) (Mean ± SEM)	36.3 ± 6.84
Education (in yrs) (Mean ± SEM)	12 ± 2
Geographical Location	Jammu & Kashmir, India
Ethnic origin of all participants	India

Table IX - Inclusion criteria of human subjects for cognitive screening study

The set of clinical parameters include lipid profiling, kidney function test (KFT), liver function test (LFT), sugar level, blood pressure and pulse rate. The MDCST was administered to all the volunteers independently and the scores were compared to determine the subjective reliability of the scoring. In addition, a medical survey of all the volunteers was done via a questionnaire with 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 [13]. Core behavioral measures (CBM) such as the alcohol consumption, tobacco use, diet and physical activity [13] were also applied to all subjects in accordance with WHO (2008) guidelines [24]. All the tests were administered by field investigators under the supervision of a clinical psychiatrist.

5.2.2 Data processing & feature selection

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 from 200 volunteers, 100 each for lowlander and highlander were randomly sampled for the study.

Following clinical parameters were measured 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), serum creatinine, 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. The value for each parameter was normalized based upon the prescribed range [25].

The MDCST comprises screening of 9 cognitive domains: Orientation, Memory Registration, Visuospatial Executive, Object Recognition, Attention, Recall, Coordination & Learning, Language and Procedural Memory [Refer to *Appendix II*, for screening details of parameters]. Each parameter is scored on a scale of 5 and cumulated for MDCST score of 45. The cumulative MDCST score <= 34 indicated onset or presence of mild cognitive impairment. For the subjects who qualified inclusion criterion, Insomnia and Beck Depression Inventory (BDI) [26] is applied to assess activities of daily living and to investigate the presence of hitherto undetected depression. Sleep was measured in terms of 0/1, with 1 indicating abnormality reported; while for BDI, value ≥ 7 indicates abnormality.

5.2.3 Test of significance among the selected cognitive & clinical parameters

We used chi-square test for categorical variables in order to compare differences between the study groups and to determine statistical significance of differences between the expected frequencies and the observed frequencies in one or more categories. It enables comparison of observed and expected frequencies objectively, since it is not always possible to tell just by looking at the data whether they are different enough to be considered statistically significant [27]. Multivariate statistical analysis was conducted on the significant factors in order to identify the independent clinical and cognitive risk factors for onset of cognitive impairment. The level of statistical significance was defined as $p < 0.05$. All statistical analysis were carried out using the STATISTICA DATAMINER 9.1 [28].

5.2.4 Association Rule Mining

Association mining, the task of finding associative rules between items in a dataset, has received considerable attention, particularly since the publication of the AIS and Apriori algorithms. The present study uses an apriori algorithm [20] to obtain the information related to clinical and cognitive screening parameters to be associated with onset of cognitive impairment. It is a popular data mining technique [29] that attempts to find interesting patterns in large databases [30].

As discussed in chapter 3, 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. As such, in this study, the minimum confidence level was subjected to atleast 30%, when the minimum support was defined to 10%; human cognition and physiology specialists were consulted for the association rules generated here, while the final confidence level was determined through a physician's opinion. Furthermore, the association mining component of the commercial data mining program, STATISTICA DATAMINER 9.1 [28], was used for the experimental parameters. Total cognitive score, BDI score & Sleep were declared as the response indicator and the remaining parameters were defined to be as categorical indicators. The process was executed with antecedent and precedent iteration rate of value 10.

3 Results

3.1 Observed significance among the selected cognitive & clinical parameters

Table X shows statistically significant impairment parameters in lowlanders among MDCST associated domains of cognitive impairment i.e., Visuospatial Executive, Attention and Coordination & Learning; whereas Object recognition is significantly impaired corresponding to

BDI as shown in Table XI. The Orientation impairment parameter was significantly associated with insomnia (Table XII) 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 (Table X) with p-value ≤ 0.05; whereas no parameter is significant corresponding to BDI (Table XI) and insomnia (Table XII).

	Lowlanders		**Highlanders**	
	Chi-square	**p-value**	**Chi-square**	**p-value**
Visuospatial Executive	22.5195	0.0002	15.1725	0.0044
Attention	15.2778	0.0016	2.0888	0.5542
Coordination and Learning	14.2017	0.0067	16.4056	0.0058
Language	6.6546	0.1553	10.7607	0.0563
Recall	5.9658	0.3096	12.2461	0.0156
Object recognition	3.4211	0.3311	3.7738	0.1515
Memory Registration	3.0049	0.2226	0.8685	0.6477
Orientation	2.2338	0.5253	3.9102	0.4183
Procedural Memory	2.0308	0.1541	20.4540	0.0000

Table X - Relevance of individual cognitive screening parameters pertaining to MDCST (cumulative cognitive score of 9 domains)

	Lowlanders		**Highlanders**	
	Chi-square	**p-value**	**Chi-square**	**p-value**
Object recognition	9.5439	0.0229	4.7122	0.0948
Coordination and Learning	6.0168	0.1979	6.4409	0.2657
Recall	4.7749	0.4440	2.2947	0.6817
Visuospatial Executive	3.9827	0.4084	5.1914	0.2682
Memory Registration	3.1067	0.2115	1.0602	0.5885
Attention	2.3175	0.5092	2.2973	0.5130

Orientation	1.9394	0.5851	2.6340	0.6208
Language	0.7826	0.9408	2.9747	0.7039
Procedural Memory	0.6769	0.4106	0.0002	0.9880

Table XI - Relevance of individual cognitive screening parameters pertaining to BDI

	Lowlanders		**Highlanders**	
	Chi-square	**p-value**	**Chi-square**	**p-value**
Orientation	9.1667	0.0272	6.1027	0.1916
Attention	4.0908	0.2518	3.0841	0.3788
Object recognition	3.0829	0.3790	1.9799	0.3716
Language	2.4253	0.6581	6.4835	0.2620
Coordination and Learning	1.9974	0.7362	1.8216	0.8732
Visuospatial Executive	1.7417	0.7831	1.3404	0.8545
Recall	1.6155	0.8994	1.2922	0.8627
Memory Registration	1.5109	0.4698	1.3980	0.4971
Procedural Memory	0.0677	0.7947	0.4686	0.4937

Table XII - Relevance of individual cognitive screening parameters pertaining to Insomnia

Among clinical observations, the chi-square test doesn't indicate any significant parameter for lowlanders corresponding to either total cognitive score, BDI or insomnia. However for highlanders significance is reported for Vitamin-B12 (p-value = 0.0099) corresponding to total cognitive score, Total Cholesterol (p-value = 0.0094) corresponding to BDI and Folic Acid (p-value = 0.0023) corresponding to insomnia.

3.2 Discovered associative rules

Table XIII 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 XIII - Associative rules discovered 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 XIV 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 XIV - Associative rules discovered for highlanders

5.4 Discussion

Our results reveals that different set of key MDCST domains and clinical measures need monitoring for the analysis of MCI and undetected depression (via 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%).

5.5 Conclusion

This study is the first of an attempt to investigate the effects of prolonged stay at high altitudes (≥ 4300m) for lowlanders as well as highlanders on their cognitive performance based on clinical parameters and cognitive domains based on MDCST. The MDCST is coupled with clinical measures to analyze high altitude induced cognitive impairment and insomnia. Healthy individuals with no clinical antecedents of depression were recruited to negate the influence of these factors on the cognitive performance. The subjects were recruited randomly for both the high altitude and low altitude location population.

Our data and the analyses identify the MDCST domains and clinical parameters that need to be analyzed for the identification of early onset cognitive-impairment at high altitudes (≥ 4300m) amongst lowlander and highlander populations. In principle, these parameters can be used for screening human subjects from their geographical altitudes for relocalization at higher altitudes.

References

1. Lieberman, P., Protopapas, A., Reed, E.,Youngs, J.W. & Kanki B.G. *Cognitive deficits at high altitude*. Nature, 372, pp 325-26. 1994.
2. Bharke, M. & Hale, S.B. *Effect of altitude on mood, behavior and cognitive functioning*. Sports Medicine, 16, pp 97–125. 1993.
3. Benveniste, H., Drejer, J., Schousboe, A. & Diemer N.H. *Elevation of the extra cellular concentration of glutamate and aspartate in the rat hippocampus during transient cerebral ischemia monitored by intracerebral microdialysis*. Journal of Neurochemistry, 43, pp 1369–1376. 1984.
4. Rossi, D. J., Oshima, T. & Attwell, D. *Glutamate release in severe brain ischaemia is mainly by reversed uptake*. Nature, 403, pp 316–321. 2000.
5. Askew, E.W. *Work at high altitude and oxidative stress: antioxidant nutrients*. Toxicology, 180, pp 107–119. 2002.
6. Hota, S.K., Hota, K.B., Prasad, D., Ilavazhagan, G. & Singh, S.B. *Oxidative-stress-induced alterations in Sp factors mediate transcriptional regulation of the NR1 subunit in hippocampus during hypoxia*. Free Radical Biology & Medicine, 49, pp 178–191. 2010.
7. Hartman, R.E., Lee, J.M., Zipfel, G.J. and Wozniak, D.F. *Characterizing learning deficits and hippocampal neuron loss following transient global cerebral ischemia in rats*. Brain and Research, 10 (1043), pp 48–56. 2005.
8. Pellegrini-Giampietro, D.E., Zukin, R.S., Bennett, M.V., Cho, S. and W. A. Pulsinelli. *Switch in glutamate receptor subunit gene expression in CA1 subfield of hippocampus following global ischemia in rats.* Proceedings of the National Academy of Sciences, 89, pp 10499–503. 1992.
9. Gorter, J.A., Petrozzino, J. J., Aronica, E. M., *et al. Global ischemia induces downregulation of GluR2 mRNA and increases AMPA receptor mediated Ca2+ influx in hippocampal CA1 neuron of gerbil*. Journal of Neuroscience, 17, pp 6179–88. 1997.
10. Chandel, N.S., Maltepe, E., Goldwasser, E. *et al. Mitochondrial reactive oxygen species trigger hypoxia-induced transcription*. Proceedings of the National Academy of Sciences, 95(20), pp 11715–720. 1998.
11. Singh, S.B., Thakur, L., Anand, J.P., *et al. Effect of high altitude (HA) on event related brain potentials*. Indian Journal of Physiology and Pharmacology, 47, pp 52-58. 2003.
12. Yan, X., Zhang, J., Gong, Q., Weng, X. *Prolonged high-altitude residence impacts verbal working memory: an fMRI study*. Experimental Brain Research, 208, pp 437-445, 2011.

13. Hota, S.K., Sharma, V.K., Hota, K., *et al. Multi-domain cognitive screening test for neuropsychological assessment for cognitive decline in acclimatized lowlanders staying at high altitude.* Indian Journal of *Medical Research*, 136, pp 411-420. 2012.
14. Folstein, M.F., Folstein, S.E. & McHugh, P.R. *A practical method for grading cognitive state of patients for the clinicians*. Journal of Psychiatric Research, 12, pp 189-198, 1975.
15. Nasreddine, Z.S., Phillips, N.A., Bédirian, V., *et al.. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment.* Journal of the American Geriatrics Society, 53, pp 695-699. 2005.
16. Borson, S. *The mini-cog: a cognitive "vitals signs" measure for dementia screening in multi-lingual elderly*. International Journal of Geriatric Psychiatry, 15, pp 1021-1027. 2000.
17. Jane, B.T., Emory, H., Laboff, J.A. & Mary E.M. *Self-Administered Screening for Mild Cognitive Impairment: Initial validation of a computerized test battery*. Journal *of* Neuropsychiatry *&* Clinical Neurosciences, 17, pp 98-105. 2005.
18. Lori, F., Jennifer, A.F., Leah, K., *et al*. *Validation of a new symptom impact questionnaire for mild to moderate cognitive impairment*. International Psychogeriatrics, 18, pp 135-56. 2006.
19. Anhøj, J. *Generic Design of Web-Based Clinical Databases*. Journal *of* Medical Internet Research, 5(4), e27. 2003.
20. Agrawal, R., Imielinski, T. & Swami A. *Mining association rules between sets of items in large databases*. Proceedings of ACM SIGMOD international conference on management of data, New York, pp. 207-216, 1993.
21. Brossette, S.E., Sprague, A.P., Hardin, J.M., *et al. Association rules and data mining in hospital infection control land public health surveillance*. Journal of the American Medical Informatics Association, 5, pp 373-381. 1998.
22. Chen, E.S., Hripcsak, G., Xu, H., *et al. Automated acquisition of disease drug knowledge from biomedical and clinical documents: An initial study*. Journal of the American Medical Informatics Association, 15, pp 87-98. 2008.
23. Hanauer, D.A., Rhodes, D.R. & Chinnaiyan, A.M. *Exploring clinical associations using '-omics' based enrichment analyses*. PLoS One, 4(4), e5203, 2009.
24. World Health Organization, Switzerland. *WHO STEPS Surveillance*. Internet: www.who.int/chp/steps. 2008.
25. Marshall, W.J. *Clinical Biochemistry: Metabolic and Clinical Aspects.* (2nd ed.). Churchill Livingstone. 2008.
26. Beck, A.T., Ward, C.H., Mendelson, M., Mock, J., Erbaugh, J. (1961). *An inventory for measuring depression*. Archives of General Psychiatry, 4 pp 561-571. 1961

27. Sokal, R.R. & Rohlf, F.J. *Biometry: the principles and practice of statistics in biological research*. (4th ed.). W. H. Freeman and Co.: New York. 2012.
28. StatSoft, Inc. *STATISTICA (data analysis software system), version 9.1*; www.statsoft.com. 2010.
29. Borgelt, C. *Simple algorithms for frequent item set mining advances in machine learning* Springer Berlin / Heidelberg, pp 351-369. 2010.
30. Goethals, B. *Survey on frequent pattern mining*. University of Helsinki. 2003. Internet: http://adrem.ua.ac.be/~goethals/software/survey.pdf

CONCLUSION AND FUTURE DIRECTION

CONCLUSION AND FUTURE DIRECTION

Clinical Informatics is an upcoming and challenging domain that holds potential for development of new informatics based techniques in the field of clinical science, especially in India. The field entails development of new Information systems, storage solutions and mining techniques for effective health care pertaining to a patient at the level of individual and population. In this research work, I have tried to address the applicability of knowledge discovery process in clinical informatics and plan for devising new strategies for future.

For the data storage problem lying in façade of clinical informatics, I have proposed a clinical dimensional model that can be used for development of 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 includes qualitative/quantitative and also 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. A future direction of the proposed solution can be the large scale implementation of this prototype in form of a patient centric management system. It will provide a platform for longitudinal storage of patient's data in non-volatile, subject-oriented and integrated form, corresponding to each visit or any diagnostic process undergone into one of the networked hospital.

As an application of data mining, I performed a case study on dataset of brain tumor patients (primary stage). The association mining performed in this study 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. Also, based on the parameters identified, I propose a normalized regression model along with Haemoglobin content, Alkaline Phosphatase and Serum Bilirubin for prediction of occurrence of STATE (brain tumor) as 0 (absent) or 1 (present). Parameter identified in this study along with the proposed predictive model is currently under observatory and validation of practicing physicians and oncologist across hospitals in India.

Temporal mining of clinical data is one of the less explored domain in this field. Pertaining to this challenging domain, I have proposed a novel "SN" algorithm, to map clinical parameters found to be associated with a disease, and to its state at various temporal points. The proposed algorithm 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 as

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. Also the future need is to develop a generalized computational tool that can be used for auguring future state associated to any disease based on the clinical parameters found to be associated with it. The future state will be predicted based on clinical values observed at historical temporal points.

Another data mining study that I have performed in this work, is a cross-sectional study 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 insomnia. 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 have been observed atleast with 30% support and more than 60% confidence for behavioural and clinical measures. This particular study was performed by the research support and data received from DIHAR, DRDO, India. The results observed in the study have been confirmed by the human cognition and physiology experts from DIHAR, DRDO, India. An interesting direction to work upon in this area would be to analyze the effect of parameters identified in this study longitudinally at the level of individual and population (lowlander and highlander) with respect to cognitive impairment, beck depression inventory and insomnia. Also it will be interesting to discover the key parameters associated to cognitive impairment, beck depression inventory and insomnia for the native population at an altitude of >4300m.

The future holds lots of challenges and potential new developments in this field. Draft bill of EHR regulation in India is being proposed in parliament during 2013 winter session. Once the bill is passed as a regulation and there are definite policies associated for EHR implementation in India, would like to work in its development. Also would like to work on the possible scenario of integration of EHRs with Aadhar (unique identification system for Indian population). EHRs along with database based storage of clinical data further expands the possibilities regarding data mining thereby opening the door to a vast source of clinical data analysis, that would involve usage of existing as well as new algorithms.

APPENDICES

Appendix I

A1- SQL queries for creation of staging schema & its tables

/*Create Staging Schema*/

CREATE DATABASE CLINICAL_STAGING_DATA;

Staging db sql queries for creation of table:

- USE CLINICAL_STAGING_DATA;

- **/*Create staging table for Date information */**

  DROP TABLE IF EXISTS date;
  CREATE TABLE date (
  DATE_VALUE date default NULL,
  Day_of_month int(11) default NULL,
  Month_sort_value char(20) default NULL,
  Week_sort_value char(20) default NULL,
  Quarter_sort_value char(20) default NULL,
  Day_of_week char(18) default NULL,
  Calender_year int(11) default NULL,
  Month_of_year char(20) default NULL,
  Quarter_of_year int(11) default NULL,
  Week_of_year int(11) default NULL);

- **/*Create staging table for storing Time information*/**

  DROP TABLE IF EXISTS time;
  CREATE TABLE time (
  Hour int(11) default NULL,
  Min int(11) default NULL,
  Sec int(11) default NULL);

- **/*Create staging table for storing Patient information obtained from varied source*/**

  DROP TABLE IF EXISTS Patient
  CREATE TABLE Patient (

PATIENT_NAME varchar(120) default NULL,

AGE int(11) default NULL,

GENDER varchar(6) default NULL,

DATE_OF_REG datetime default NULL,

PATIENT_HISTORY text,

CLINICAL_DIAGNOSIS varchar(57) default NULL,

TREATMENT_PLANNED tinytext,

HAEMOGLOBIN_CONTENT double default NULL,

Total Leucocyte count(TLC)(cmm) int(11) default NULL,

Basophils tinytext,

Eosinophils tinytext,

Neutrophils(%) tinytext,

Lymphocytes(%) tinytext,

Monocytes(%) tinytext,

Platelet count(cmm) int(11) default NULL,

KFT_Creatinine double default NULL,

KFT_BUN int(11) default NULL,

LFT_Sr_Bilirubin double default NULL,

LFT_ALP int(11) default NULL,

LFT_SGOT int(11) default NULL,

LFT_SGPT int(11) default NULL,

Total Protein (g/dl) double default NULL,

Albumin(Alb)(mg/dl) double default NULL,

Serum Alkaline Phosphate(SAP)(IU/L) tinytext,

Chromium(Cr) tinytext,

Sodium(Na) int(11) default NULL,

Potassium(K) double default NULL,

Biopsy Report varchar(207) default NULL,

X-Ray varchar(37) default NULL,

CT Scan text,

MRI tinytext,

ULTRASOUND varchar(255) default NULL,

MAMMOGRAPHY varchar(182) default NULL,

RESULT varchar(45) NOT NULL,

SOURCE varhcar(45) NOT NULL);

- **/*Create staging table for storing Disease information*/**

```
DROP TABLE IF EXISTS Disease
CREATE TABLE Disease (
Disease_name text not null,
Disease_detail text not null,
Disease_cause text not null,
Disease_symptoms text not null,
Date_of_inclusion text not null,
Other text not null);
```

- **/*Create staging table for storing Diagnostic Test information*/**

```
DROP TABLE IF EXISTS  Diagnostic_test;
CREATE TABLE  Diagnostic_test (
  Diagnostic_Test_Name text not null,
  Diagnostic_Test_Details text not null,
  Upper_Value_Male text not null,
  Lower_Value_Male text not null,
  Gender text not null,
  Standard_Value_Male text not null,
  Date_of_Inclusion datetime default NULL);
```

- **/*Create staging table for storing Patient Image information*/**

```
DROP TABLE IF EXISTS  Image;
CREATE TABLE  Image (
  Image_Name text not null,
  Image_Type text not null,
  Patient_Name text not null,
  SOURCE varhcar(45) NOT NULL);
```

A2- SQL queries for creation of Clinical warehouse & its tables

/*Create Warehouse or Functional Schema*/

CREATE CLINICAL_WAREHOUSE;

Clinical warehouse SQL queries for creation of table:

- USE CLINICAL_WAREHOUSE;

- **/*Create Dimension table for storing Date information */**

```
DROP TABLE IF EXISTS DIM_DATE;
CREATE TABLE DIM_DATE (
DATE_ID int(11) NOT NULL auto_increment,
Date_value date NOT NULL,
Day_of_month int(11) default NULL,
Month_sort_value varchar(20) default NULL,
Week_sort_value varchar (20) default NULL,
Quarter_sort_value varchar (20) default NULL,
Date_modified date default NULL,
Day_of_week varchar (18) default NULL,
Calender_year int(11) default NULL,
Month_of_year varchar (20) default NULL,
Quarter_of_year int(11) default NULL,
Week_of_year int(11) default NULL,
PRIMARY KEY (DATE_ID));
```

- **/*Create Dimension table for storing Time information*/**

```
DROP TABLE IF EXISTS DIM_TIME;
CREATE TABLE DIM_TIME (
TIME_ID int(11) NOT NULL auto_increment,
Hour int(11) default NULL,
Min int(11) default NULL,
Sec int(11) default NULL,
PRIMARY KEY (TIME_ID));
```

- **/*Create Dimension table for storing information pertaining to diseases*/**

```
DROP TABLE IF EXISTS DIM_DISEASE;
CREATE TABLE DIM_DISEASE (
DISEASE_ID int(11) not null auto_increment,
Disease_name varchar(200) not null,
Disease_detail text not null,
Disease_cause text null,
Disease_symptoms text not null,
Date_of_inclusion date not null,
Other_information varchar(100) null,
PRIMARY KEY(DISEASE_ID));
```

- **/*Create Dimension table for storing information pertaining to diagnostic test*/**

```
DROP TABLE IF EXISTS  DIM_DIAGNOSTIC_TEST;
CREATE TABLE  DIM_DIAGNOSTIC_TEST (
  TEST_ID int(10) unsigned NOT NULL auto_increment,
  Diagnostic_Test_Name varchar(45) NOT NULL default '',
  Diagnostic_Test_Details varchar(1000) default NULL,
  Upper_Value_Male decimal(10,2) default NULL,
  Upper_Value_Female decimal(10,2) default NULL,
  Lower_Value_Male decimal(10,2) default NULL,
  Lower_Value_Female decimal(10,2) default NULL,
  Standard_Value_Male decimal(10,2) default NULL,
  Standard_Value_Female decimal(10,2) default NULL,
  Date_of_Inclusion datetime default NULL,
  PRIMARY KEY  (TEST_ID));
```

- **/*Create Dimension table for storing image information*/**

```
DROP TABLE IF EXISTS DIM_IMAGE;
CREATE TABLE  DIM_IMAGE (
  Image_ID int(10) unsigned NOT NULL auto_increment,
  Image_Type varchar(100) default NULL,
  Image_Details varchar(1000) default NULL,
  Date_of_Inclusion datetime default NULL,
```

```
PRIMARY KEY  (Image_ID));
```

- **/*Create Dimension table for storing personal information of patients*/**

```
DROP TABLE IF EXISTS DIM_PATIENT;
CREATE TABLE  DIM_PATIENT (
 PATIENT_ID int(11) NOT NULL auto_increment,
 Name varchar(50) default NULL,
 Age int(11) default NULL,
 Gender varchar(20) default NULL,
 Education varchar(18) default NULL,
 Income varchar(18) default NULL,
 Address varchar(400) default NULL,
 Current_Location varchar(40) default NULL,
 Contact_number int(10) default NULL,
 Email varchar(50) default NULL,
 Date_of_Registration date default NULL,
 Smoking_Status varchar(18) default NULL,
 Alcholic_Status varchar(18) default NULL,
 Diet_Type varchar(18) default NULL,
 Start_date date NOT NULL,
 End_date date default NULL,
 Flag varchar(20) NOT NULL,
 PRIMARY KEY  (PATIENT_ID));
```

- **/*Create Fact table for storing measures corresponding to patients for each diagnostic test value on given date & time*/**

```
DROP TABLE IF EXISTS FACT_PATIENT;
CREATE TABLE  FACT_PATIENT (
 Patient_id int(10) unsigned NOT NULL,
 Date_of_measurment_id int(10) unsigned NOT NULL,
 Time_of_measurement int(10) unsigned NOT NULL,
 Disease_id int(10) unsigned NOT NULL,
 Test_id varchar(45) NOT NULL default',
 Measure_ID decimal(10,02 default NULL,
```

```
  Measured_Value decimal(10,2) default NULL,
  Diagnosis varchar(500) default NULL,
  Recomendation varchar(4000) default NULL,
  PRIMARY KEY (DATE_OF_MEASURMENT_ID, TIME_OF_MEASUREMENT,
                    DISEASE_ID, TEST_ID, PATIENT_ID),
  KEY Index_Patient (PATIENT_ID),
  KEY FK_fact_patient_3 (TIME_OF_MEASUREMENT),
  KEY FK_fact_patient_4 (DISEASE_ID),
  KEY FK_fact_patient_5 (TEST_ID)
) ENGINE=MyISAM DEFAULT CHARSET=latin1;
```

- **/*Create Fact table for storing measures corresponding to medical image for patients on a given date & time*/**

```
DROP TABLE IF EXISTS FACT_PATIENT_IMAGE_DETAIL;
CREATE TABLE   FACT_PATIENT_IMAGE_DETAIL (
  Patient_Id int(10) unsigned NOT NULL,
  Image_Id int(10) unsigned NOT NULL,
  Date_Id int(10) unsigned NOT NULL,
  Time_Id int(10) unsigned NOT NULL,
  Image_Area decimal(10,2) default NULL,
  Mean_Gray_Value decimal(10,2) default NULL,
  Median_Gray_Value decimal(10,2) default NULL,
  Min_Gray_Value decimal(10,2) default NULL,
  Max_Gray_Value decimal(10,2) default NULL,
  Standard_deviation decimal(10,2) default NULL,
  Centroid decimal(10,2) default NULL,
  Shape_Descriptor decimal(10,2) default NULL,
  Perimeter decimal(10,2) default NULL,
  Skewness decimal(10,2) default NULL,
  Kurtosis decimal(10,2) default NULL,
  Diagnosis varchar(200) default NULL,
  Recomendation varchar(4000) default NULL,
  PRIMARY KEY  (Patient_Id,Image_Id,Date_Id,Time_Id),
  KEY FK_fact_patient_image_detail_2 (Image_Id),
```

```
  KEY FK_fact_patient_image_detail_3 (Date_Id),
  KEY FK_fact_patient_image_detail_4 (Time_Id)
) ENGINE=MyISAM DEFAULT CHARSET=latin1;
```

B1 - ETL Mappings for processing data from source files to respective tables in staging schema (Cancer_Staging_Data)

- **Staging_date** (from Date.csv to Date table)

Code:

```
<?xml version="1.0" encoding="UTF-8"?>
<transformation-steps>
 <steps>
  <step>
   <name>Date CSV file</name>
   <type>CsvInput</type>
   <description/>
   <distribute>Y</distribute>
   <copies>1</copies>
       <partitioning>
         <method>none</method>
         <schema_name/>
         </partitioning>
   <filename>E:\My Research Material\Data Files\converted for Kettle\Date.csv</filename>
   <filename_field/>
   <rownum_field/>
   <include_filename>N</include_filename>
   <separator>,</separator>
   <enclosure>&amp;quot;</enclosure>
   <header>Y</header>
   <buffer_size>50000</buffer_size>
   <lazy_conversion>N</lazy_conversion>
   <add_filename_result>N</add_filename_result>
   <parallel>N</parallel>
   <encoding/>
```

```
  <fields>
  </fields>
   <cluster_schema/>
<remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
    <xloc>247</xloc>
    <yloc>213</yloc>
    <draw>Y</draw>
    </GUI>
  </step>
 <step>
  <name>Table output</name>
  <type>TableOutput</type>
  <description/>
  <distribute>Y</distribute>
  <copies>1</copies>
      <partitioning>
        <method>none</method>
        <schema_name/>
        </partitioning>
  <connection>STAGING</connection>
  <schema/>
  <table>Date</table>
  <commit>100</commit>
  <truncate>N</truncate>
  <ignore_errors>N</ignore_errors>
  <use_batch>Y</use_batch>
  <partitioning_enabled>N</partitioning_enabled>
  <partitioning_field/>
  <partitioning_daily>N</partitioning_daily>
  <partitioning_monthly>Y</partitioning_monthly>
  <tablename_in_field>N</tablename_in_field>
  <tablename_field/>
  <tablename_in_table>Y</tablename_in_table>
  <return_keys>N</return_keys>
```

```
   <return_field/>
    <cluster_schema/>
 <remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
     <xloc>563</xloc>
     <yloc>213</yloc>
     <draw>Y</draw>
     </GUI>
   </step>
   </steps>
<order>
 <hop> <from>Date CSV file</from><to>Table output</to><enabled>Y</enabled> </hop>
 </order>
 <notepads>
  </notepads>
</transformation-steps>
```

- **Staging_time** (from Time.csv to Time table)

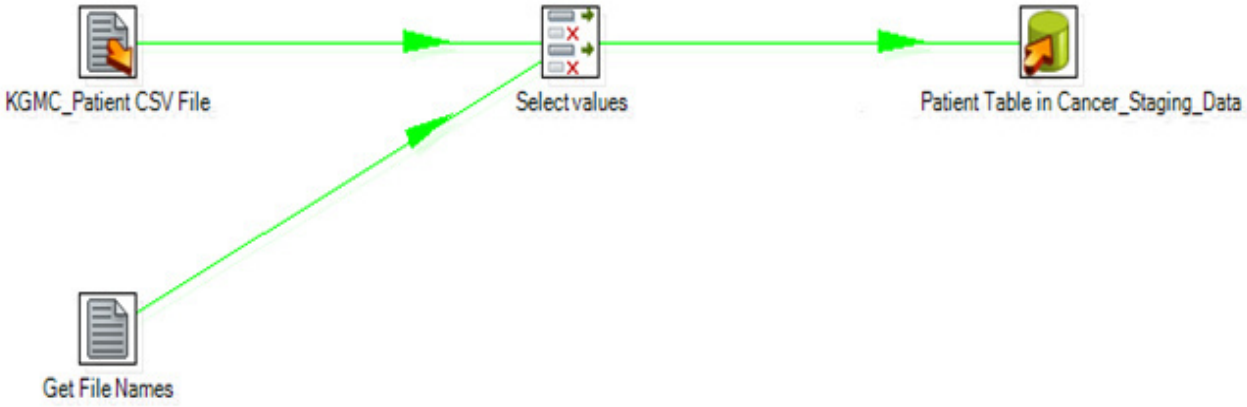
Code:

```
<?xml version="1.0" encoding="UTF-8"?>
<transformation-steps>
 <steps>
  <step>
    <name>Time CSV file</name>
    <type>CsvInput</type>
    <description/>
    <distribute>Y</distribute>
    <copies>1</copies>
        <partitioning>
          <method>none</method>
          <schema_name/>
```

```
        </partitioning>
  <filename>E:\My Research Material\Data Files\converted for Kettle\Time.csv</filename>
  <filename_field/>
  <rownum_field/>
  <include_filename>N</include_filename>
  <separator>,</separator>
  <enclosure>&amp;quot;</enclosure>
  <header>Y</header>
  <buffer_size>50000</buffer_size>
  <lazy_conversion>N</lazy_conversion>
  <add_filename_result>N</add_filename_result>
  <parallel>N</parallel>
  <encoding/>
  <fields>
  </fields>
   <cluster_schema/>
<remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
    <xloc>247</xloc>
    <yloc>213</yloc>
    <draw>Y</draw>
    </GUI>
  </step>
 <step>
  <name>Table output</name>
  <type>TableOutput</type>
  <description/>
  <distribute>Y</distribute>
  <copies>1</copies>
      <partitioning>
        <method>none</method>
        <schema_name/>
        </partitioning>
  <connection>STAGING</connection>
  <schema/>
```

```
<table>Time</table>
<commit>100</commit>
<truncate>N</truncate>
<ignore_errors>N</ignore_errors>
<use_batch>Y</use_batch>
<partitioning_enabled>N</partitioning_enabled>
<partitioning_field/>
<partitioning_daily>N</partitioning_daily>
<partitioning_monthly>Y</partitioning_monthly>
<tablename_in_field>N</tablename_in_field>
<tablename_field/>
<tablename_in_table>Y</tablename_in_table>
<return_keys>N</return_keys>
<return_field/>
<cluster_schema/>
<remotesteps> <input> </input> <output> </output> </remotesteps> <GUI>
<xloc>563</xloc>
<yloc>213</yloc>
<draw>Y</draw>
</GUI>
</step>
</steps>
<order>
<hop> <from>Time CSV file</from><to>Table output</to><enabled>Y</enabled> </hop>
</order>
<notepads>
</notepads>
</transformation-steps>
```

- **Staging_patient** (from patient files obtained from different sources to Patient table)



Code: <?xml version="1.0" encoding="UTF-8"?>

<transformation-steps>

<steps>

<step>

<name>KGMC_Patient CSV file</name>

<type>CsvInput</type>

<description/>

<distribute>Y</distribute>

<copies>1</copies>

<partitioning>

<method>none</method>

<schema_name/>

</partitioning>

<filename>E:\My Research Material\Data Files\converted for Kettle\KGMC_Patient.csv</filename>

<filename_field/>

<rownum_field/>

<include_filename>N</include_filename>

<separator>,</separator>

<enclosure>"</enclosure>

<header>Y</header>

<buffer_size>50000</buffer_size>

<lazy_conversion>N</lazy_conversion>

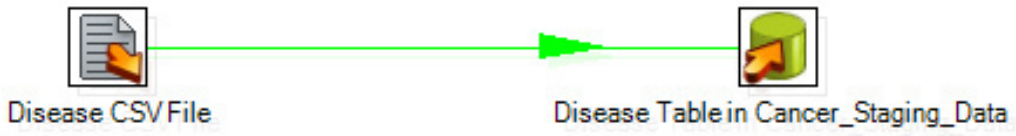
<add_filename_result>N</add_filename_result>

```
<parallel>N</parallel>
<encoding/>
<fields>
</fields>
<cluster_schema/>
<remotesteps> <input> </input> <output> </output> </remotesteps> <GUI>
<xloc>247</xloc>
<yloc>213</yloc>
<draw>Y</draw>
</GUI>
</step>
<step>
<name>Patient</name>
<type>TableOutput</type>
<description/>
<distribute>Y</distribute>
<copies>1</copies>
<partitioning>
<method>none</method>
<schema_name/>
</partitioning>
<connection>STAGING</connection>
<schema/>
<table>Patient</table>
<commit>100</commit>
<truncate>N</truncate>
<ignore_errors>N</ignore_errors>
<use_batch>Y</use_batch>
<partitioning_enabled>N</partitioning_enabled>
<partitioning_field/>
<partitioning_daily>N</partitioning_daily>
<partitioning_monthly>Y</partitioning_monthly>
<tablename_in_field>N</tablename_in_field>
<tablename_field/>
```

```
<tablename_in_table>Y</tablename_in_table>
<return_keys>N</return_keys>
<return_field/>
<cluster_schema/>
<remotesteps> <input> </input> <output> </output> </remotesteps> <GUI>
<xloc>673</xloc>
<yloc>213</yloc>
<draw>Y</draw>
</GUI>
</step>
<step>
<name>Get File Names</name>
<type>GetFileNames</type>
<description/>
<distribute>Y</distribute>
<copies>1</copies>
<partitioning>
<method>none</method>
<schema_name/>
</partitioning>
<filter>
<filterfiletype>all_files</filterfiletype>
</filter>
<rownum>N</rownum>
<isaddresult>Y</isaddresult>
<filefield>N</filefield>
<rownum_field/>
<filename_Field/>
<wildcard_Field/>
<limit>0</limit>
<file>
</file>
<cluster_schema/>
<remotesteps> <input> </input> <output> </output> </remotesteps> <GUI>
```

```
    <xloc>248</xloc>
    <yloc>338</yloc>
    <draw>Y</draw>
    </GUI>
   </step>
  <step>
   <name>Select values</name>
   <type>SelectValues</type>
   <description/>
   <distribute>Y</distribute>
   <copies>1</copies>
       <partitioning>
        <method>none</method>
        <schema_name/>
        </partitioning>
   <fields>       <select_unspecified>N</select_unspecified>
   </fields>    <cluster_schema/>
 <remotesteps>  <input>  </input>  <output>  </output> </remotesteps>   <GUI>
    <xloc>459</xloc>
    <yloc>213</yloc>
    <draw>Y</draw>
    </GUI>
   </step>
   </steps>
<order>
  <hop> <from>KGMC_Patient CSV file</from><to>Select
values</to><enabled>Y</enabled> </hop>
  <hop> <from>Get File Names</from><to>Select values</to><enabled>Y</enabled> </hop>
  <hop> <from>Select values</from><to>Patient</to><enabled>Y</enabled> </hop>
  </order>
  <notepads>
   </notepads>
 </transformation-steps>
```

- **Staging_disease** (from Disease.csv file to Disease table)



Code: <?xml version="1.0" encoding="UTF-8"?>

<transformation-steps>

<steps>

<step>

<name>Disease CSV file</name>

<type>CsvInput</type>

<description/>

<distribute>Y</distribute>

<copies>1</copies>

<partitioning>

<method>none</method>

<schema_name/>

</partitioning>

<filename>E:\My Research Material\Data Files\converted for Kettle\Disease.csv</filename>

<filename_field/>

<rownum_field/>

<include_filename>N</include_filename>

<separator>,</separator>

<enclosure>"</enclosure>

<header>Y</header>

<buffer_size>50000</buffer_size>

<lazy_conversion>N</lazy_conversion>

<add_filename_result>N</add_filename_result>

<parallel>N</parallel>

<encoding/>

<fields>

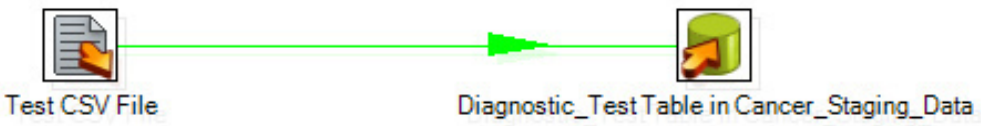
</fields>

<cluster_schema/>

<remotesteps> <input> </input> <output> </output> </remotesteps> <GUI>
<xloc>247</xloc>
<yloc>213</yloc>
<draw>Y</draw>
</GUI>
</step>
<step>
<name>Disease</name>
<type>TableOutput</type>
<description/>
<distribute>Y</distribute>
<copies>1</copies>
<partitioning>
<method>none</method>
<schema_name/>
</partitioning>
<connection>STAGING</connection>
<schema/>
<table>Disease</table>
<commit>100</commit>
<truncate>N</truncate>
<ignore_errors>N</ignore_errors>
<use_batch>Y</use_batch>
<partitioning_enabled>N</partitioning_enabled>
<partitioning_field/>
<partitioning_daily>N</partitioning_daily>
<partitioning_monthly>Y</partitioning_monthly>
<tablename_in_field>N</tablename_in_field>
<tablename_field/>
<tablename_in_table>Y</tablename_in_table>
<return_keys>N</return_keys>
<return_field/>
<cluster_schema/>
<remotesteps> <input> </input> <output> </output> </remotesteps> <GUI>

```
      <xloc>563</xloc>
      <yloc>213</yloc>
      <draw>Y</draw>
      </GUI>
    </step>
    </steps>
  <order>
   <hop> <from>Disease CSV file</from><to>Disease</to><enabled>Y</enabled> </hop>
   </order>
   <notepads>
    </notepads>
  </transformation-steps>
```

- **Staging_test** (from Test.csv file to Diagnostic_Test table)



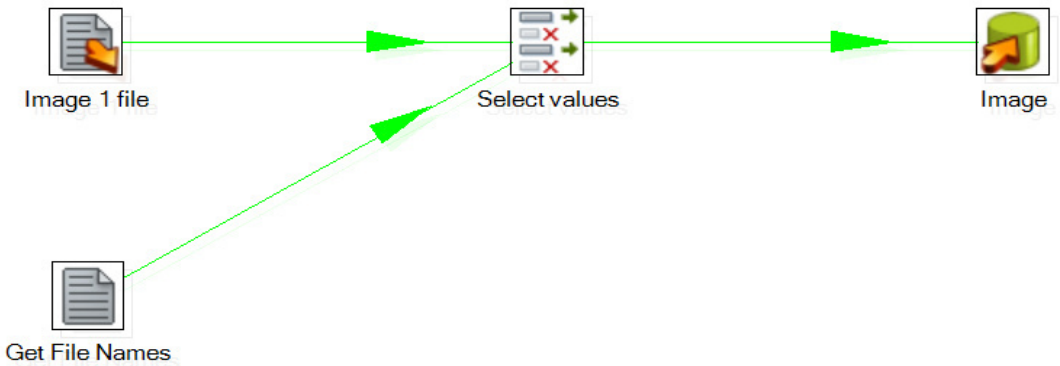
Code:

```
<?xml version="1.0" encoding="UTF-8"?>
<transformation-steps>
 <steps>
  <step>
    <name>Test CSV file</name>
    <type>CsvInput</type>
    <description/>
    <distribute>Y</distribute>
    <copies>1</copies>
        <partitioning>
          <method>none</method>
          <schema_name/>
          </partitioning>
    <filename>E:\My Research Material\Data Files\converted for Kettle\Test.csv</filename>
    <filename_field/>
    <rownum_field/>
```

<include_filename>N</include_filename>
<separator>,</separator>
<enclosure>"</enclosure>
<header>Y</header>
<buffer_size>50000</buffer_size>
<lazy_conversion>N</lazy_conversion>
<add_filename_result>N</add_filename_result>
<parallel>N</parallel>
<encoding/>
<fields>
</fields>
<cluster_schema/>
<remotesteps> <input> </input> <output> </output> </remotesteps> <GUI>
<xloc>247</xloc>
<yloc>213</yloc>
<draw>Y</draw>
</GUI>
</step>
<step>
<name>Diagnostic_test</name>
<type>TableOutput</type>
<description/>
<distribute>Y</distribute>
<copies>1</copies>
<partitioning>
<method>none</method>
<schema_name/>
</partitioning>
<connection>STAGING</connection>
<schema/>
<table>Diagnostic_test</table>
<commit>100</commit>
<truncate>N</truncate>
<ignore_errors>N</ignore_errors>

<use_batch>Y</use_batch>
<partitioning_enabled>N</partitioning_enabled>
<partitioning_field/>
<partitioning_daily>N</partitioning_daily>
<partitioning_monthly>Y</partitioning_monthly>
<tablename_in_field>N</tablename_in_field>
<tablename_field/>
<tablename_in_table>Y</tablename_in_table>
<return_keys>N</return_keys>
<return_field/>
<cluster_schema/>
<remotesteps> <input> </input> <output> </output> </remotesteps> <GUI>
<xloc>563</xloc>
<yloc>213</yloc>
<draw>Y</draw>
</GUI>
</step>
</steps>
<order>
<hop> <from>Test CSV file</from><to>Diagnostic_test</to><enabled>Y</enabled> </hop>
</order>
<notepads>
</notepads>
</transformation-steps>

- **Staging_image** (from compilation of different patient images to Image table)



Code:

```
<?xml version="1.0" encoding="UTF-8"?>
<transformation-steps>
 <steps>
  <step>
   <name>Image 1 file</name>
   <type>CsvInput</type>
   <description/>
   <distribute>Y</distribute>
   <copies>1</copies>
       <partitioning>
        <method>none</method>
        <schema_name/>
        </partitioning>
   <filename>E:\My Research Material\Data Files\converted for Kettle\Image_1</filename>
   <filename_field/>
   <rownum_field/>
   <include_filename>N</include_filename>
   <separator>,</separator>
   <enclosure>&amp;quot;</enclosure>
   <header>Y</header>
   <buffer_size>50000</buffer_size>
   <lazy_conversion>N</lazy_conversion>
   <add_filename_result>N</add_filename_result>
   <parallel>N</parallel>
   <encoding/>
   <fields>
   </fields>
    <cluster_schema/>
 <remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
     <xloc>247</xloc>
     <yloc>213</yloc>
     <draw>Y</draw>
     </GUI>
   </step>
```

```
<step>
  <name>Image</name>
  <type>TableOutput</type>
  <description/>
  <distribute>Y</distribute>
  <copies>1</copies>
      <partitioning>
        <method>none</method>
        <schema_name/>
        </partitioning>
  <connection>STAGING</connection>
  <schema/>
  <table>Image</table>
  <commit>100</commit>
  <truncate>N</truncate>
  <ignore_errors>N</ignore_errors>
  <use_batch>Y</use_batch>
  <partitioning_enabled>N</partitioning_enabled>
  <partitioning_field/>
  <partitioning_daily>N</partitioning_daily>
  <partitioning_monthly>Y</partitioning_monthly>
  <tablename_in_field>N</tablename_in_field>
  <tablename_field/>
  <tablename_in_table>Y</tablename_in_table>
  <return_keys>N</return_keys>
  <return_field/>
   <cluster_schema/>
<remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
    <xloc>673</xloc>
    <yloc>213</yloc>
    <draw>Y</draw>
    </GUI>
  </step>
 <step>
```

```
  <name>Get File Names</name>
  <type>GetFileNames</type>
  <description/>
  <distribute>Y</distribute>
  <copies>1</copies>
      <partitioning>
        <method>none</method>
        <schema_name/>
        </partitioning>
  <filter>
    <filterfiletype>all_files</filterfiletype>
  </filter>
  <rownum>N</rownum>
  <isaddresult>Y</isaddresult>
  <filefield>N</filefield>
  <rownum_field/>
  <filename_Field/>
  <wildcard_Field/>
  <limit>0</limit>
  <file>
  </file>
   <cluster_schema/>
<remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
    <xloc>248</xloc>
    <yloc>338</yloc>
    <draw>Y</draw>
    </GUI>
  </step>
 <step>
  <name>Select values</name>
  <type>SelectValues</type>
  <description/>
  <distribute>Y</distribute>
  <copies>1</copies>
```

```
        <partitioning>
          <method>none</method>
          <schema_name/>
          </partitioning>
    <fields>       <select_unspecified>N</select_unspecified>
    </fields>     <cluster_schema/>
 <remotesteps>   <input>   </input>   <output>   </output> </remotesteps>    <GUI>
      <xloc>459</xloc>
      <yloc>213</yloc>
      <draw>Y</draw>
      </GUI>
    </step>
    </steps>
<order>
  <hop> <from>Image 1 file</from><to>Select values</to><enabled>Y</enabled> </hop>
  <hop> <from>Get File Names</from><to>Select values</to><enabled>Y</enabled> </hop>
  <hop> <from>Select values</from><to>Image</to><enabled>Y</enabled> </hop>
  </order>
  <notepads>
   </notepads>
 </transformation-steps>
```

B2 - ETL Mappings for processing data from tables of staging schema (Cancer_Staging_Data) to tables in functional schema (Clinical_Warehouse).

- **DW_dim_date** (from Date to Dim_Date)

Code: <?xml version="1.0" encoding="UTF-8"?>
<transformation-steps>
<steps>
<step>
<name>Date</name>
<type>TableInput</type>
<description/>
<distribute>Y</distribute>
<copies>1</copies>
<partitioning>
<method>none</method>
<schema_name/>
</partitioning>
<connection>Cancer_Staging_Data</connection>
<sql>SELECT * FROM Date </sql>
<limit>0</limit>
<lookup/>
<execute_each_row>N</execute_each_row>
<variables_active>N</variables_active>
<lazy_conversion_active>N</lazy_conversion_active>
<cluster_schema/>
<remotesteps> <input> </input> <output> </output> </remotesteps> <GUI>
<xloc>173</xloc>
<yloc>213</yloc>
<draw>Y</draw>
</GUI>

```
  </step>
 <step>
  <name>DIM_DATE</name>
  <type>TableOutput</type>
  <description/>
  <distribute>Y</distribute>
  <copies>1</copies>
      <partitioning>
       <method>none</method>
       <schema_name/>
       </partitioning>
  <connection>Functional</connection>
  <schema/>
  <table>dim_date</table>
  <commit>100</commit>
  <truncate>N</truncate>
  <ignore_errors>N</ignore_errors>
  <use_batch>Y</use_batch>
  <partitioning_enabled>N</partitioning_enabled>
  <partitioning_field/>
  <partitioning_daily>N</partitioning_daily>
  <partitioning_monthly>Y</partitioning_monthly>
  <tablename_in_field>N</tablename_in_field>
  <tablename_field/>
  <tablename_in_table>Y</tablename_in_table>
  <return_keys>N</return_keys>
  <return_field/>
   <cluster_schema/>
<remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
    <xloc>686</xloc>
    <yloc>213</yloc>
    <draw>Y</draw>
    </GUI>
  </step>
```

```
<step>
  <name>Add_Date_ID</name>
  <type>Sequence</type>
  <description/>
  <distribute>Y</distribute>
  <copies>1</copies>
     <partitioning>
       <method>none</method>
       <schema_name/>
       </partitioning>
    <valuename>valuename</valuename>
    <use_database>N</use_database>
    <connection/>
    <schema/>
    <seqname>SEQ_</seqname>
    <use_counter>Y</use_counter>
    <counter_name/>
    <start_at>1</start_at>
    <increment_by>1</increment_by>
    <max_value>999999999</max_value>
   <cluster_schema/>
<remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
    <xloc>435</xloc>
    <yloc>213</yloc>
    <draw>Y</draw>
    </GUI>
  </step>
  </steps>
<order>
 <hop> <from>Date</from><to>Add_Date_ID</to><enabled>Y</enabled> </hop>
 <hop> <from>Add_Date_ID</from><to>DIM_DATE</to><enabled>Y</enabled> </hop>
 </order>
 <notepads>
  </notepads>
```

</transformation-steps>

- **DW_dim_time** (from Time to Dim_Time)

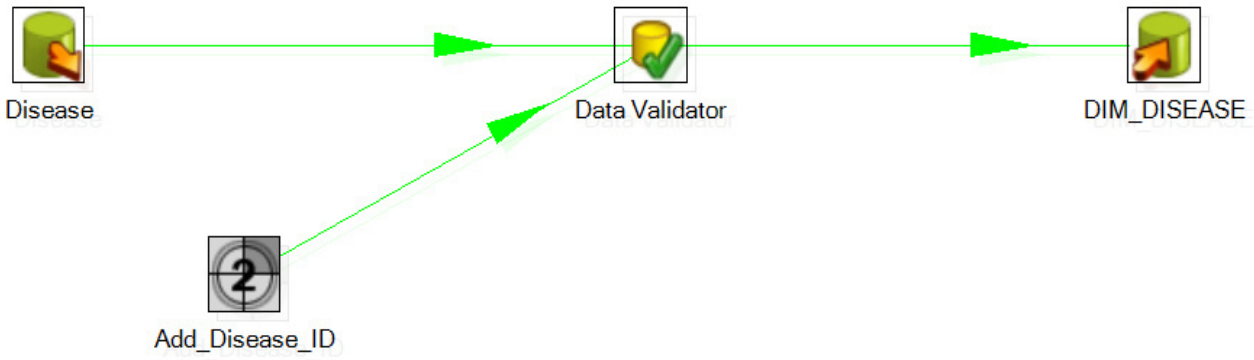
Code: <?xml version="1.0" encoding="UTF-8"?>
<transformation-steps>
<steps>
<step>
<name>Time</name>
<type>TableInput</type>
<description/>
<distribute>Y</distribute>
<copies>1</copies>
<partitioning>
<method>none</method>
<schema_name/>
</partitioning>
<connection>Cancer_Staging_Data</connection>
<sql>SELECT * FROM Time</sql>
<limit>0</limit>
<lookup/>
<execute_each_row>N</execute_each_row>
<variables_active>N</variables_active>
<lazy_conversion_active>N</lazy_conversion_active>
<cluster_schema/>
<remotesteps> <input> </input> <output> </output> </remotesteps> <GUI>
<xloc>173</xloc>
<yloc>213</yloc>
<draw>Y</draw>
</GUI>

```
  </step>
 <step>
  <name>DIM_DATE</name>
  <type>TableOutput</type>
  <description/>
  <distribute>Y</distribute>
  <copies>1</copies>
      <partitioning>
       <method>none</method>
       <schema_name/>
       </partitioning>
  <connection>Functional</connection>
  <schema/>
  <table>dim_time</table>
  <commit>100</commit>
  <truncate>N</truncate>
  <ignore_errors>N</ignore_errors>
  <use_batch>Y</use_batch>
  <partitioning_enabled>N</partitioning_enabled>
  <partitioning_field/>
  <partitioning_daily>N</partitioning_daily>
  <partitioning_monthly>Y</partitioning_monthly>
  <tablename_in_field>N</tablename_in_field>
  <tablename_field/>
  <tablename_in_table>Y</tablename_in_table>
  <return_keys>N</return_keys>
  <return_field/>
   <cluster_schema/>
<remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
    <xloc>686</xloc>
    <yloc>213</yloc>
    <draw>Y</draw>
    </GUI>
  </step>
```

```
<step>
  <name>Add_Date_ID</name>
  <type>Sequence</type>
  <description/>
  <distribute>Y</distribute>
  <copies>1</copies>
     <partitioning>
      <method>none</method>
      <schema_name/>
      </partitioning>
    <valuename>valuename</valuename>
    <use_database>N</use_database>
    <connection/>
    <schema/>
    <seqname>SEQ_</seqname>
    <use_counter>Y</use_counter>
    <counter_name/>
    <start_at>1</start_at>
    <increment_by>1</increment_by>
    <max_value>999999999</max_value>
   <cluster_schema/>
<remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
    <xloc>435</xloc>
    <yloc>213</yloc>
    <draw>Y</draw>
    </GUI>
  </step>
  </steps>
<order>
 <hop> <from>Time</from><to>Add_Date_ID</to><enabled>Y</enabled> </hop>
 <hop> <from>Add_Date_ID</from><to>DIM_DATE</to><enabled>Y</enabled> </hop>
 </order>
 <notepads>
  </notepads>
```

</transformation-steps>

- **DW_dim_disease** (from Disease to Dim_Disease)



Code: <?xml version="1.0" encoding="UTF-8"?>

<transformation-steps>

<steps>

<step>

<name>Disease</name>

<type>TableInput</type>

<description/>

<distribute>Y</distribute>

<copies>1</copies>

<partitioning>

<method>none</method>

<schema_name/>

</partitioning>

<connection>Cancer_Staging_Data</connection>

<sql>SELECT * FROM Disease</sql>

<limit>0</limit>

<lookup/>

<execute_each_row>N</execute_each_row>

<variables_active>N</variables_active>

<lazy_conversion_active>N</lazy_conversion_active>

<cluster_schema/>

<remotesteps> <input> </input> <output> </output> </remotesteps> <GUI>

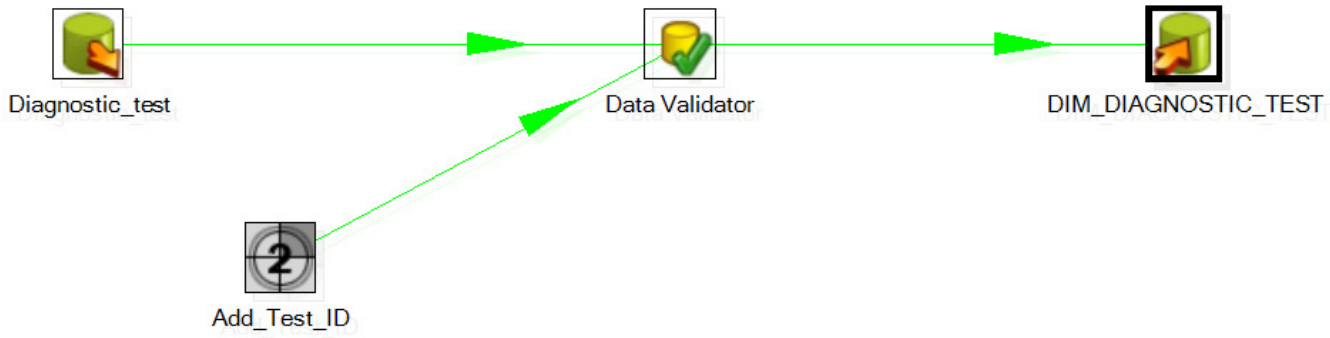
<xloc>173</xloc>

```
    <yloc>213</yloc>
    <draw>Y</draw>
    </GUI>
  </step>
 <step>
  <name>DIM_DISEASE</name>
  <type>TableOutput</type>
  <description/>
  <distribute>Y</distribute>
  <copies>1</copies>
      <partitioning>
        <method>none</method>
        <schema_name/>
        </partitioning>
  <connection>Functional</connection>
  <schema/>
  <table>DIM_DISEASE</table>
  <commit>100</commit>
  <truncate>N</truncate>
  <ignore_errors>N</ignore_errors>
  <use_batch>Y</use_batch>
  <partitioning_enabled>N</partitioning_enabled>
  <partitioning_field/>
  <partitioning_daily>N</partitioning_daily>
  <partitioning_monthly>Y</partitioning_monthly>
  <tablename_in_field>N</tablename_in_field>
  <tablename_field/>
  <tablename_in_table>Y</tablename_in_table>
  <return_keys>N</return_keys>
  <return_field/>
   <cluster_schema/>
<remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
    <xloc>686</xloc>
    <yloc>213</yloc>
```

```
    <draw>Y</draw>
    </GUI>
  </step>
 <step>
  <name>Add_Disease_ID</name>
  <type>Sequence</type>
  <description/>
  <distribute>Y</distribute>
  <copies>1</copies>
      <partitioning>
        <method>none</method>
        <schema_name/>
        </partitioning>
    <valuename>valuename</valuename>
    <use_database>N</use_database>
    <connection/>
    <schema/>
    <seqname>SEQ_</seqname>
    <use_counter>Y</use_counter>
    <counter_name/>
    <start_at>1</start_at>
    <increment_by>1</increment_by>
    <max_value>999999999</max_value>
   <cluster_schema/>
<remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
    <xloc>263</xloc>
    <yloc>313</yloc>
    <draw>Y</draw>
    </GUI>
  </step>
 <step>
  <name>Data Validator</name>
  <type>Validator</type>
  <description/>
```

<distribute>Y</distribute>
<copies>1</copies>
<partitioning>
<method>none</method>
<schema_name/>
</partitioning>
<cluster_schema/>
<remotesteps> <input> </input> <output> </output> </remotesteps> <GUI>
<xloc>450</xloc>
<yloc>213</yloc>
<draw>Y</draw>
</GUI>
</step>
</steps>
<order>
<hop> <from>Disease</from><to>Data Validator</to><enabled>Y</enabled> </hop>
<hop> <from>Add_Disease_ID</from><to>Data Validator</to><enabled>Y</enabled>
</hop>
<hop> <from>Data Validator</from><to>DIM_DISEASE</to><enabled>Y</enabled>
</hop>
</order>
<notepads>
</notepads>
</transformation-steps>

- **DW_dim_test** (from Diagnostic_Test to Dim_Test)



Code:

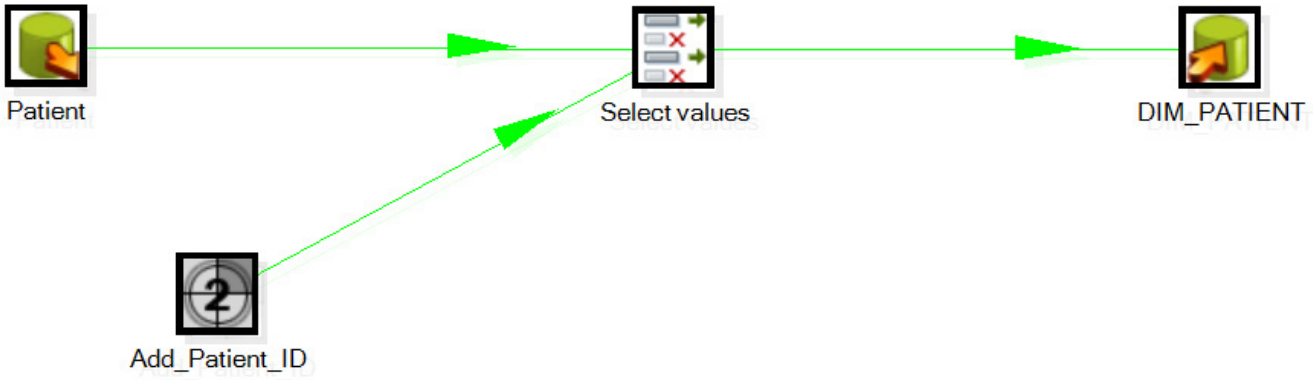
```
<?xml version="1.0" encoding="UTF-8"?>
<transformation-steps>
 <steps>
  <step>
   <name>Diagnostic_test</name>
   <type>TableInput</type>
   <description/>
   <distribute>Y</distribute>
   <copies>1</copies>
       <partitioning>
        <method>none</method>
        <schema_name/>
        </partitioning>
   <connection>Cancer_Staging_Data</connection>
   <sql>SELECT * FROM Diagnostic_test</sql>
   <limit>0</limit>
   <lookup/>
   <execute_each_row>N</execute_each_row>
   <variables_active>N</variables_active>
   <lazy_conversion_active>N</lazy_conversion_active>
    <cluster_schema/>
 <remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
     <xloc>173</xloc>
     <yloc>213</yloc>
     <draw>Y</draw>
     </GUI>
   </step>
  <step>
   <name>DIM_DIAGNOSTIC_TEST</name>
   <type>TableOutput</type>
   <description/>
   <distribute>Y</distribute>
   <copies>1</copies>
       <partitioning>
```

```
        <method>none</method>
        <schema_name/>
        </partitioning>
   <connection>Functional</connection>
   <schema/>
   <table>DIM_DIAGNOSTIC_TEST</table>
   <commit>100</commit>
   <truncate>N</truncate>
   <ignore_errors>N</ignore_errors>
   <use_batch>Y</use_batch>
   <partitioning_enabled>N</partitioning_enabled>
   <partitioning_field/>
   <partitioning_daily>N</partitioning_daily>
   <partitioning_monthly>Y</partitioning_monthly>
   <tablename_in_field>N</tablename_in_field>
   <tablename_field/>
   <tablename_in_table>Y</tablename_in_table>
   <return_keys>N</return_keys>
   <return_field/>
    <cluster_schema/>
<remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
     <xloc>686</xloc>
     <yloc>213</yloc>
     <draw>Y</draw>
     </GUI>
   </step>
 <step>
   <name>Add_Test_ID</name>
   <type>Sequence</type>
   <description/>
   <distribute>Y</distribute>
   <copies>1</copies>
       <partitioning>
        <method>none</method>
```

```
        <schema_name/>
        </partitioning>
    <valuename>valuename</valuename>
    <use_database>N</use_database>
    <connection/>
    <schema/>
    <seqname>SEQ_</seqname>
    <use_counter>Y</use_counter>
    <counter_name/>
    <start_at>1</start_at>
    <increment_by>1</increment_by>
    <max_value>999999999</max_value>
   <cluster_schema/>
<remotesteps> <input> </input> <output> </output> </remotesteps> <GUI>
    <xloc>263</xloc>
    <yloc>313</yloc>
    <draw>Y</draw>
    </GUI>
  </step>
 <step>
  <name>Data Validator</name>
  <type>Validator</type>
  <description/>
  <distribute>Y</distribute>
  <copies>1</copies>
      <partitioning>
        <method>none</method>
        <schema_name/>
        </partitioning>
   <cluster_schema/>
<remotesteps> <input> </input> <output> </output> </remotesteps> <GUI>
    <xloc>450</xloc>
    <yloc>213</yloc>
    <draw>Y</draw>
```

```
    </GUI>
   </step>
   </steps>
<order>
  <hop> <from>Diagnostic_test</from><to>Data Validator</to><enabled>Y</enabled>
</hop>
  <hop> <from>Add_Test_ID</from><to>Data Validator</to><enabled>Y</enabled> </hop>
  <hop> <from>Data
Validator</from><to>DIM_DIAGNOSTIC_TEST</to><enabled>Y</enabled> </hop>
  </order>
  <notepads>
   </notepads>
 </transformation-steps>
```

- **DW_dim_patient** (from Patient to Dim_Patient)



Code:

```
<?xml version="1.0" encoding="UTF-8"?>
<transformation-steps>
 <steps>
  <step>
    <name>Patient</name>
    <type>TableInput</type>
    <description/>
    <distribute>Y</distribute>
    <copies>1</copies>
        <partitioning>
```

```
        <method>none</method>
        <schema_name/>
        </partitioning>
   <connection>Cancer_Staging_Data</connection>
   <sql>SELECT * FROM Patient</sql>
   <limit>0</limit>
   <lookup/>
   <execute_each_row>N</execute_each_row>
   <variables_active>N</variables_active>
   <lazy_conversion_active>N</lazy_conversion_active>
    <cluster_schema/>
<remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
     <xloc>173</xloc>
     <yloc>213</yloc>
     <draw>Y</draw>
     </GUI>
   </step>
  <step>
   <name>DIM_PATIENT</name>
   <type>TableOutput</type>
   <description/>
   <distribute>Y</distribute>
   <copies>1</copies>
       <partitioning>
        <method>none</method>
        <schema_name/>
        </partitioning>
   <connection>Functional</connection>
   <schema/>
   <table>DIM_PATIENT</table>
   <commit>100</commit>
   <truncate>N</truncate>
   <ignore_errors>N</ignore_errors>
   <use_batch>Y</use_batch>
```

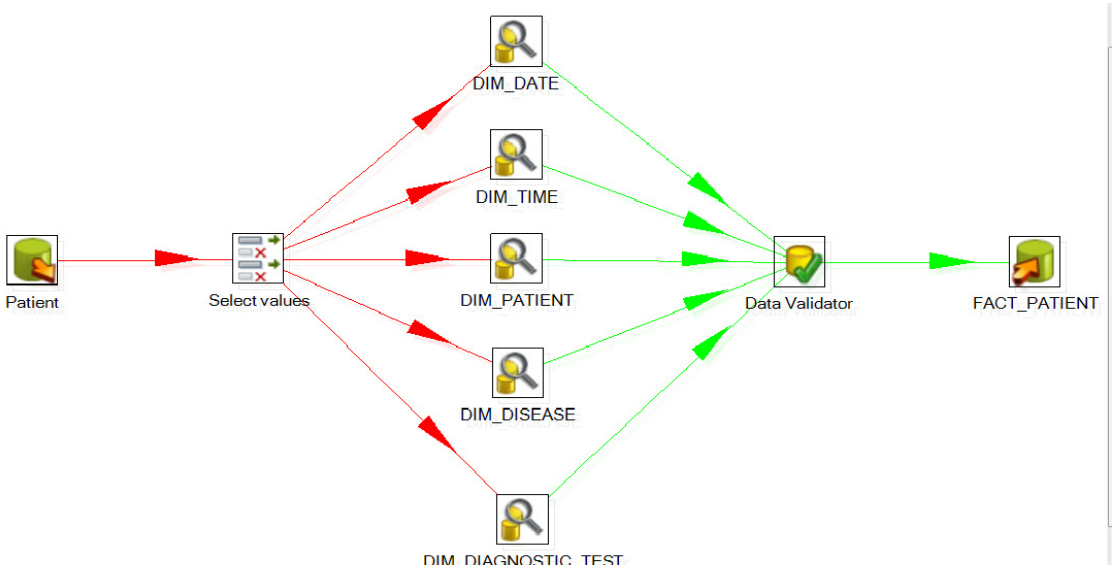
```
<partitioning_enabled>N</partitioning_enabled>
<partitioning_field/>
<partitioning_daily>N</partitioning_daily>
<partitioning_monthly>Y</partitioning_monthly>
<tablename_in_field>N</tablename_in_field>
<tablename_field/>
<tablename_in_table>Y</tablename_in_table>
<return_keys>N</return_keys>
<return_field/>
<cluster_schema/>
<remotesteps> <input> </input> <output> </output> </remotesteps> <GUI>
<xloc>686</xloc>
<yloc>214</yloc>
<draw>Y</draw>
</GUI>
</step>
<step>
<name>Add_Patient_ID</name>
<type>Sequence</type>
<description/>
<distribute>Y</distribute>
<copies>1</copies>
<partitioning>
<method>none</method>
<schema_name/>
</partitioning>
<valuename>valuename</valuename>
<use_database>N</use_database>
<connection/>
<schema/>
<seqname>SEQ_</seqname>
<use_counter>Y</use_counter>
<counter_name/>
<start_at>1</start_at>
```

```
    <increment_by>1</increment_by>
    <max_value>999999999</max_value>
   <cluster_schema/>
 <remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
    <xloc>248</xloc>
    <yloc>321</yloc>
    <draw>Y</draw>
    </GUI>
  </step>
 <step>
  <name>Select values</name>
  <type>SelectValues</type>
  <description/>
  <distribute>Y</distribute>
  <copies>1</copies>
      <partitioning>
        <method>none</method>
        <schema_name/>
        </partitioning>
  <fields>      <select_unspecified>N</select_unspecified>
  </fields>    <cluster_schema/>
 <remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
    <xloc>447</xloc>
    <yloc>214</yloc>
    <draw>Y</draw>
    </GUI>
  </step>
  </steps>
<order>
 <hop> <from>Patient</from><to>Select values</to><enabled>Y</enabled> </hop>
 <hop> <from>Add_Patient_ID</from><to>Select values</to><enabled>Y</enabled> </hop>
 <hop> <from>Select values</from><to>DIM_PATIENT</to><enabled>Y</enabled> </hop>
 </order>
 <notepads>
```

</notepads>

</transformation-steps>

- **DW_fact_patient** (from Patient to Fact_Patient)



Code: <?xml version="1.0" encoding="UTF-8"?>

<transformation-steps>

<steps>

<step>

<name>Patient</name>

<type>TableInput</type>

<description/>

<distribute>N</distribute>

<copies>1</copies>

<partitioning>

<method>none</method>

<schema_name/>

</partitioning>

<connection>Cancer_Staging_Data</connection>

<sql>SELECT * FROM Patient</sql>

<limit>0</limit>

<lookup/>

<execute_each_row>N</execute_each_row>

<variables_active>N</variables_active>

```
<lazy_conversion_active>N</lazy_conversion_active>
<cluster_schema/>
<remotesteps> <input> </input> <output> </output> </remotesteps> <GUI>
<xloc>117</xloc>
<yloc>216</yloc>
<draw>Y</draw>
</GUI>
</step>
<step>
<name>FACT_PATIENT</name>
<type>TableOutput</type>
<description/>
<distribute>Y</distribute>
<copies>1</copies>
<partitioning>
<method>none</method>
<schema_name/>
</partitioning>
<connection>Functional</connection>
<schema/>
<table>FACT_PATIENT</table>
<commit>100</commit>
<truncate>N</truncate>
<ignore_errors>N</ignore_errors>
<use_batch>Y</use_batch>
<partitioning_enabled>N</partitioning_enabled>
<partitioning_field/>
<partitioning_daily>N</partitioning_daily>
<partitioning_monthly>Y</partitioning_monthly>
<tablename_in_field>N</tablename_in_field>
<tablename_field/>
<tablename_in_table>Y</tablename_in_table>
<return_keys>N</return_keys>
<return_field/>
```

```
   <cluster_schema/>
<remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
    <xloc>765</xloc>
    <yloc>217</yloc>
    <draw>Y</draw>
    </GUI>
  </step>
 <step>
  <name>Select values</name>
  <type>SelectValues</type>
  <description/>
  <distribute>N</distribute>
  <copies>1</copies>
      <partitioning>
        <method>none</method>
        <schema_name/>
        </partitioning>
  <fields>      <select_unspecified>N</select_unspecified>
  </fields>    <cluster_schema/>
<remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
    <xloc>263</xloc>
    <yloc>215</yloc>
    <draw>Y</draw>
    </GUI>
  </step>
 <step>
  <name>DIM_DATE</name>
  <type>DBLookup</type>
  <description/>
  <distribute>Y</distribute>
  <copies>1</copies>
      <partitioning>
        <method>none</method>
        <schema_name/>
```

```
        </partitioning>
    <connection>Functional</connection>
    <cache>N</cache>
    <cache_load_all>N</cache_load_all>
    <cache_size>0</cache_size>
    <lookup>
     <schema/>
     <table>DIM_DATE</table>
     <orderby/>
     <fail_on_multiple>N</fail_on_multiple>
     <eat_row_on_failure>N</eat_row_on_failure>
    </lookup>
     <cluster_schema/>
<remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
     <xloc>430</xloc>
     <yloc>71</yloc>
     <draw>Y</draw>
     </GUI>
    </step>
  <step>
    <name>DIM_TIME</name>
    <type>DBLookup</type>
    <description/>
    <distribute>Y</distribute>
    <copies>1</copies>
        <partitioning>
          <method>none</method>
          <schema_name/>
          </partitioning>
    <connection>Functional</connection>
    <cache>N</cache>
    <cache_load_all>N</cache_load_all>
    <cache_size>0</cache_size>
    <lookup>
```

```
    <schema/>
    <table>DIM_TIME</table>
    <orderby/>
    <fail_on_multiple>N</fail_on_multiple>
    <eat_row_on_failure>N</eat_row_on_failure>
   </lookup>
    <cluster_schema/>
<remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
    <xloc>430</xloc>
    <yloc>146</yloc>
    <draw>Y</draw>
    </GUI>
   </step>
 <step>
   <name>DIM_DISEASE</name>
   <type>DBLookup</type>
   <description/>
   <distribute>Y</distribute>
   <copies>1</copies>
       <partitioning>
        <method>none</method>
        <schema_name/>
        </partitioning>
   <connection>Functional</connection>
   <cache>N</cache>
   <cache_load_all>N</cache_load_all>
   <cache_size>0</cache_size>
   <lookup>
    <schema/>
    <table>DIM_DISEASE</table>
    <orderby/>
    <fail_on_multiple>N</fail_on_multiple>
    <eat_row_on_failure>N</eat_row_on_failure>
   </lookup>
```

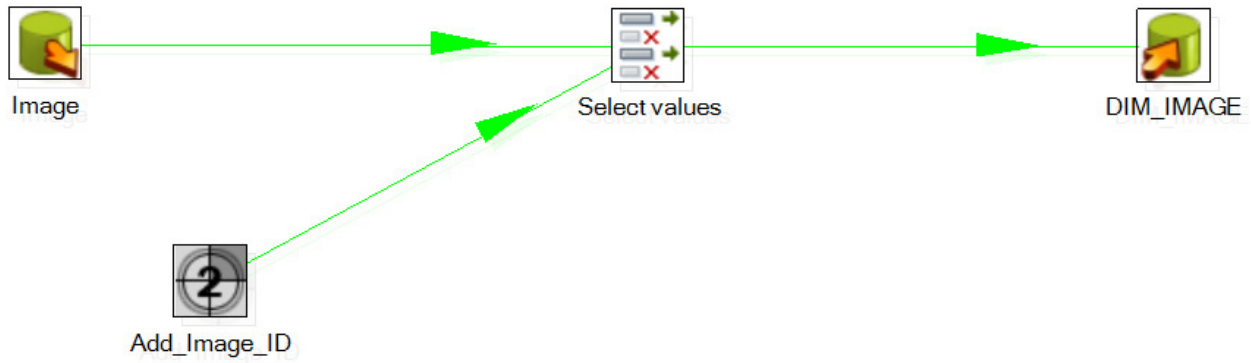
```
    <cluster_schema/>
<remotesteps>   <input>   </input>   <output>   </output> </remotesteps>   <GUI>
    <xloc>431</xloc>
    <yloc>291</yloc>
    <draw>Y</draw>
    </GUI>
  </step>
 <step>
  <name>DIM_DIAGNOSTIC_TEST</name>
  <type>DBLookup</type>
  <description/>
  <distribute>Y</distribute>
  <copies>1</copies>
      <partitioning>
       <method>none</method>
       <schema_name/>
       </partitioning>
  <connection>Functional</connection>
  <cache>N</cache>
  <cache_load_all>N</cache_load_all>
  <cache_size>0</cache_size>
  <lookup>
    <schema/>
    <table>DIM_DIAGNOSTIC_TEST</table>
    <orderby/>
    <fail_on_multiple>N</fail_on_multiple>
    <eat_row_on_failure>N</eat_row_on_failure>
  </lookup>
   <cluster_schema/>
<remotesteps>   <input>   </input>   <output>   </output> </remotesteps>   <GUI>
    <xloc>434</xloc>
    <yloc>388</yloc>
    <draw>Y</draw>
    </GUI>
```

```
  </step>
 <step>
  <name>Data Validator</name>
  <type>Validator</type>
  <description/>
  <distribute>Y</distribute>
  <copies>1</copies>
      <partitioning>
       <method>none</method>
       <schema_name/>
       </partitioning>
   <cluster_schema/>
<remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
    <xloc>613</xloc>
    <yloc>217</yloc>
    <draw>Y</draw>
    </GUI>
  </step>
 <step>
  <name>DIM_PATIENT</name>
  <type>DBLookup</type>
  <description/>
  <distribute>Y</distribute>
  <copies>1</copies>
      <partitioning>
       <method>none</method>
       <schema_name/>
       </partitioning>
  <connection>Functional</connection>
  <cache>N</cache>
  <cache_load_all>N</cache_load_all>
  <cache_size>0</cache_size>
  <lookup>
   <schema/>
```

```
    <table>DIM_PATIENT</table>
    <orderby/>
    <fail_on_multiple>N</fail_on_multiple>
    <eat_row_on_failure>N</eat_row_on_failure>
   </lookup>
    <cluster_schema/>
 <remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
    <xloc>430</xloc>
    <yloc>215</yloc>
    <draw>Y</draw>
    </GUI>
   </step>
   </steps>
<order>
  <hop> <from>Patient</from><to>Select values</to><enabled>Y</enabled> </hop>
  <hop> <from>Select values</from><to>DIM_DATE</to><enabled>Y</enabled> </hop>
  <hop> <from>Select values</from><to>DIM_TIME</to><enabled>Y</enabled> </hop>
  <hop> <from>Select values</from><to>DIM_DISEASE</to><enabled>Y</enabled> </hop>
  <hop> <from>Select
values</from><to>DIM_DIAGNOSTIC_TEST</to><enabled>Y</enabled> </hop>
  <hop> <from>Select values</from><to>DIM_PATIENT</to><enabled>Y</enabled> </hop>
  <hop> <from>DIM_DATE</from><to>Data Validator</to><enabled>Y</enabled> </hop>
  <hop> <from>DIM_TIME</from><to>Data Validator</to><enabled>Y</enabled> </hop>
  <hop> <from>DIM_DISEASE</from><to>Data Validator</to><enabled>Y</enabled>
</hop>
  <hop> <from>DIM_DIAGNOSTIC_TEST</from><to>Data
Validator</to><enabled>Y</enabled> </hop>
  <hop> <from>Data Validator</from><to>FACT_PATIENT</to><enabled>Y</enabled>
</hop>
  <hop> <from>DIM_PATIENT</from><to>Data Validator</to><enabled>Y</enabled>
</hop>
  </order>
  <notepads>
   </notepads>
```

</transformation-steps>

- **DW_dim_image** (from Image to Dim_Image)



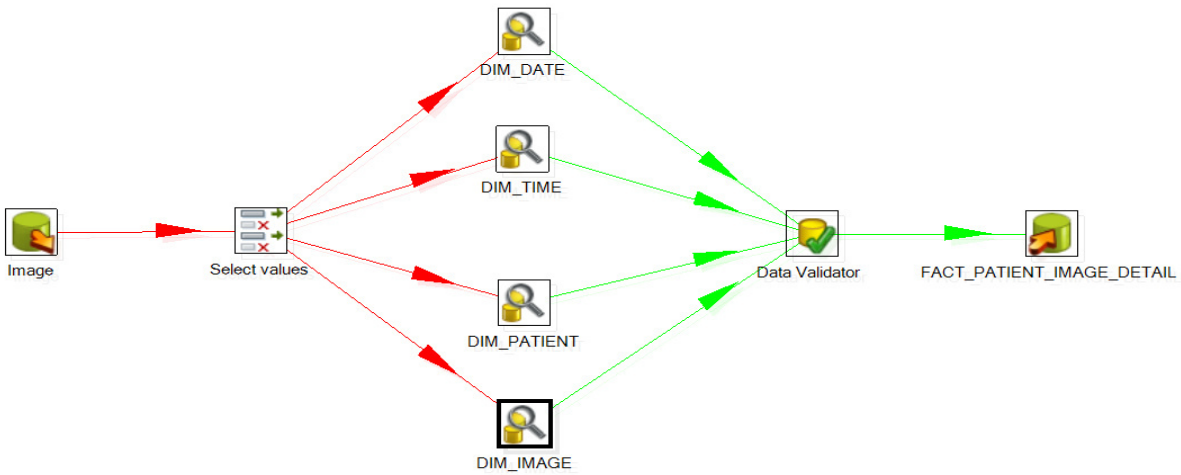
Code: <?xml version="1.0" encoding="UTF-8"?>
<transformation-steps>
<steps>
<step>
<name>Image</name>
<type>TableInput</type>
<description/>
<distribute>Y</distribute>
<copies>1</copies>
<partitioning>
<method>none</method>
<schema_name/>
</partitioning>
<connection>Cancer_Staging_Data</connection>
<sql>SELECT * FROM Image</sql>
<limit>0</limit>
<lookup/>
<execute_each_row>N</execute_each_row>
<variables_active>N</variables_active>
<lazy_conversion_active>N</lazy_conversion_active>
<cluster_schema/>
<remotesteps> <input> </input> <output> </output> </remotesteps> <GUI>
<xloc>173</xloc>

```
    <yloc>213</yloc>
    <draw>Y</draw>
    </GUI>
  </step>
 <step>
  <name>DIM_IMAGE</name>
  <type>TableOutput</type>
  <description/>
  <distribute>Y</distribute>
  <copies>1</copies>
      <partitioning>
        <method>none</method>
        <schema_name/>
        </partitioning>
  <connection>Functional</connection>
  <schema/>
  <table>DIM_IMAGE</table>
  <commit>100</commit>
  <truncate>N</truncate>
  <ignore_errors>N</ignore_errors>
  <use_batch>Y</use_batch>
  <partitioning_enabled>N</partitioning_enabled>
  <partitioning_field/>
  <partitioning_daily>N</partitioning_daily>
  <partitioning_monthly>Y</partitioning_monthly>
  <tablename_in_field>N</tablename_in_field>
  <tablename_field/>
  <tablename_in_table>Y</tablename_in_table>
  <return_keys>N</return_keys>
  <return_field/>
   <cluster_schema/>
<remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
    <xloc>686</xloc>
    <yloc>214</yloc>
```

```
    <draw>Y</draw>
    </GUI>
  </step>
 <step>
  <name>Add_Image_ID</name>
  <type>Sequence</type>
  <description/>
  <distribute>Y</distribute>
  <copies>1</copies>
      <partitioning>
        <method>none</method>
        <schema_name/>
        </partitioning>
    <valuename>valuename</valuename>
    <use_database>N</use_database>
    <connection/>
    <schema/>
    <seqname>SEQ_</seqname>
    <use_counter>Y</use_counter>
    <counter_name/>
    <start_at>1</start_at>
    <increment_by>1</increment_by>
    <max_value>999999999</max_value>
   <cluster_schema/>
<remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
    <xloc>248</xloc>
    <yloc>321</yloc>
    <draw>Y</draw>
    </GUI>
  </step>
 <step>
  <name>Select values</name>
  <type>SelectValues</type>
  <description/>
```

<distribute>Y</distribute>
<copies>1</copies>
<partitioning>
<method>none</method>
<schema_name/>
</partitioning>
<fields> <select_unspecified>N</select_unspecified>
</fields> <cluster_schema/>
<remotesteps> <input> </input> <output> </output> </remotesteps> <GUI>
<xloc>447</xloc>
<yloc>214</yloc>
<draw>Y</draw>
</GUI>
</step>
</steps>
<order>
<hop> <from>Image</from><to>Select values</to><enabled>Y</enabled> </hop>
<hop> <from>Add_Image_ID</from><to>Select values</to><enabled>Y</enabled> </hop>
<hop> <from>Select values</from><to>DIM_IMAGE</to><enabled>Y</enabled> </hop>
</order>
<notepads>
</notepads>
</transformation-steps>

- **DW_fact_image** (from Image to Fact_Patient_Image_Detail)



Code:

```
<?xml version="1.0" encoding="UTF-8"?>
<transformation-steps>
 <steps>
  <step>
   <name>Image</name>
   <type>TableInput</type>
   <description/>
   <distribute>N</distribute>
   <copies>1</copies>
       <partitioning>
        <method>none</method>
        <schema_name/>
        </partitioning>
   <connection>Cancer_Staging_Data</connection>
   <sql>SELECT * FROM Image</sql>
   <limit>0</limit>
   <lookup/>
   <execute_each_row>N</execute_each_row>
   <variables_active>N</variables_active>
   <lazy_conversion_active>N</lazy_conversion_active>
    <cluster_schema/>
 <remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
     <xloc>117</xloc>
     <yloc>216</yloc>
     <draw>Y</draw>
     </GUI>
   </step>
  <step>
   <name>FACT_PATIENT_IMAGE_DETAIL</name>
   <type>TableOutput</type>
   <description/>
   <distribute>Y</distribute>
   <copies>1</copies>
       <partitioning>
```

```
        <method>none</method>
        <schema_name/>
        </partitioning>
   <connection>Functional</connection>
   <schema/>
   <table>FACT_PATIENT_IMAGE_DETAIL</table>
   <commit>100</commit>
   <truncate>N</truncate>
   <ignore_errors>N</ignore_errors>
   <use_batch>Y</use_batch>
   <partitioning_enabled>N</partitioning_enabled>
   <partitioning_field/>
   <partitioning_daily>N</partitioning_daily>
   <partitioning_monthly>Y</partitioning_monthly>
   <tablename_in_field>N</tablename_in_field>
   <tablename_field/>
   <tablename_in_table>Y</tablename_in_table>
   <return_keys>N</return_keys>
   <return_field/>
    <cluster_schema/>
<remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
     <xloc>765</xloc>
     <yloc>217</yloc>
     <draw>Y</draw>
     </GUI>
   </step>
 <step>
   <name>Select values</name>
   <type>SelectValues</type>
   <description/>
   <distribute>N</distribute>
   <copies>1</copies>
       <partitioning>
         <method>none</method>
```

```
        <schema_name/>
        </partitioning>
  <fields>       <select_unspecified>N</select_unspecified>
  </fields>    <cluster_schema/>
<remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
    <xloc>263</xloc>
    <yloc>215</yloc>
    <draw>Y</draw>
    </GUI>
  </step>
 <step>
  <name>DIM_DATE</name>
  <type>DBLookup</type>
  <description/>
  <distribute>Y</distribute>
  <copies>1</copies>
      <partitioning>
        <method>none</method>
        <schema_name/>
        </partitioning>
  <connection>Functional</connection>
  <cache>N</cache>
  <cache_load_all>N</cache_load_all>
  <cache_size>0</cache_size>
  <lookup>
    <schema/>
    <table>DIM_DATE</table>
    <orderby/>
    <fail_on_multiple>N</fail_on_multiple>
    <eat_row_on_failure>N</eat_row_on_failure>
  </lookup>
   <cluster_schema/>
<remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
    <xloc>430</xloc>
```

```
    <yloc>71</yloc>
    <draw>Y</draw>
    </GUI>
  </step>
 <step>
  <name>DIM_TIME</name>
  <type>DBLookup</type>
  <description/>
  <distribute>Y</distribute>
  <copies>1</copies>
      <partitioning>
        <method>none</method>
        <schema_name/>
        </partitioning>
  <connection>Functional</connection>
  <cache>N</cache>
  <cache_load_all>N</cache_load_all>
  <cache_size>0</cache_size>
  <lookup>
    <schema/>
    <table>DIM_TIME</table>
    <orderby/>
    <fail_on_multiple>N</fail_on_multiple>
    <eat_row_on_failure>N</eat_row_on_failure>
  </lookup>
   <cluster_schema/>
<remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
    <xloc>429</xloc>
    <yloc>156</yloc>
    <draw>Y</draw>
    </GUI>
  </step>
 <step>
  <name>DIM_IMAGE</name>
```

```
  <type>DBLookup</type>
  <description/>
  <distribute>Y</distribute>
  <copies>1</copies>
      <partitioning>
        <method>none</method>
        <schema_name/>
        </partitioning>
  <connection>Functional</connection>
  <cache>N</cache>
  <cache_load_all>N</cache_load_all>
  <cache_size>0</cache_size>
  <lookup>
    <schema/>
    <table>DIM_IMAGE</table>
    <orderby/>
    <fail_on_multiple>N</fail_on_multiple>
    <eat_row_on_failure>N</eat_row_on_failure>
  </lookup>
   <cluster_schema/>
<remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
    <xloc>431</xloc>
    <yloc>355</yloc>
    <draw>Y</draw>
    </GUI>
  </step>
 <step>
  <name>Data Validator</name>
  <type>Validator</type>
  <description/>
  <distribute>Y</distribute>
  <copies>1</copies>
      <partitioning>
        <method>none</method>
```

```
        <schema_name/>
        </partitioning>
    <cluster_schema/>
<remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
    <xloc>613</xloc>
    <yloc>217</yloc>
    <draw>Y</draw>
    </GUI>
  </step>
 <step>
  <name>DIM_PATIENT</name>
  <type>DBLookup</type>
  <description/>
  <distribute>Y</distribute>
  <copies>1</copies>
      <partitioning>
        <method>none</method>
        <schema_name/>
        </partitioning>
  <connection>Functional</connection>
  <cache>N</cache>
  <cache_load_all>N</cache_load_all>
  <cache_size>0</cache_size>
  <lookup>
    <schema/>
    <table>DIM_PATIENT</table>
    <orderby/>
    <fail_on_multiple>N</fail_on_multiple>
    <eat_row_on_failure>N</eat_row_on_failure>
  </lookup>
   <cluster_schema/>
<remotesteps>  <input>  </input>  <output>  </output> </remotesteps>  <GUI>
    <xloc>430</xloc>
    <yloc>267</yloc>
```

```
    <draw>Y</draw>
    </GUI>
   </step>
   </steps>
<order>
 <hop> <from>Image</from><to>Select values</to><enabled>Y</enabled> </hop>
 <hop> <from>Select values</from><to>DIM_DATE</to><enabled>Y</enabled> </hop>
 <hop> <from>Select values</from><to>DIM_TIME</to><enabled>Y</enabled> </hop>
 <hop> <from>Select values</from><to>DIM_IMAGE</to><enabled>Y</enabled> </hop>
 <hop> <from>Select values</from><to>DIM_PATIENT</to><enabled>Y</enabled> </hop>
 <hop> <from>DIM_DATE</from><to>Data Validator</to><enabled>Y</enabled> </hop>
 <hop> <from>DIM_TIME</from><to>Data Validator</to><enabled>Y</enabled> </hop>
 <hop> <from>DIM_IMAGE</from><to>Data Validator</to><enabled>Y</enabled> </hop>
 <hop> <from>Data
  Validator</from><to>FACT_PATIENT_IMAGE_DETAIL</to><enabled>Y</enabled> </hop>
 <hop> <from>DIM_PATIENT</from><to>Data Validator</to><enabled>Y</enabled> </hop>
 </order>
 <notepads>
  </notepads>
</transformation-steps>
```

APPENDIX II

Multi Domain Cognitive Screening Test (MDCST)

The procedure for administration of MDCST and the scoring awarded for each parameter is as follows:

- **Orientation:** To assess orientation on the MDCST score, the subject had to answer his/her date of birth, present address, day, last festival, and name of head/director/employer. One point is awarded for each correct answer.
- **Memory registration:** The subject is made to remember 5 logically unrelated words. The points awarded are equal to the number of words the subject remembered after repeating the words twice.
- **Visuospatial executive:** The subject is made to match the words to numerical in the order A to 1, B to 2 and so on. The subject is made to replicate the cube and draw a clock with hour and minute hand showing 0800 h. On correct matching of the alphabets with the numbers, cube drawing and correct positioning of either the hour hand or minute hand, fetched the subject one point each. Drawing a clock with numbers written fetched the subject one point.
- **Object recognition:** The subject is made to identify four unrelated objects. The subject is awarded with 1 point for each correct identification. The subject is also made to recall the name of two school teachers and awarded one point for correct recall.
- **Coordination and learning:** The subject is made to learn a complex task like tying a tie knot in 5 steps. The number of times the subject failed to learn the task is subtracted from 5 to obtain the final score. If the number of attempts are more than 5, the final score for learning the task is 0. If the subject learns the task in the first attempt but takes more than 3 min to complete it in the second and third attempt, is awarded 3 points.
- **Attention:** The subject is made to tap a finger depending on number of times 'A' appeared on different sites of the computer screen or number of times the administrator uttered the alphabet 'A' while reading a jumbled alphabet sequence. The subject is awarded one point on tapping the finger for correct number of times 'A' appeared in the jumbled alphabet sequence. The subject is made to read a sequence of numbers forward and then backward. On correct reading of the sequence in both the directions; the subject gets one point. For assessment of the problem solving ability, the subject is made to subtract 7 sequentially for four times, starting with 100 and one point is awarded for each correct result.
- **Language:** The subjects is made to read 5 sentences in English or in the local dialect as preferred by the subject. Correct reading of each sentence resulted in awarding of one point.

- **Recall:** Delayed recall is assessed by making the subject recall the words that are presented during memory registration. Each correct recall led to award of one point. If the subject fails to recall 2 words, is assessed for cued recall by asking to identify the correct words from a list of ten. On correct identification of all the words from the list, the subject is awarded with one point.
- **Procedural memory:** The subject is given a single attempt to perform the complex task that he is made to learn in five steps for assessment of procedural memory. For each correct step the subject gets one point.

All the individual parameter scores are then summed to determine the total MDCST score. Subjects scoring >34 are considered to be normal, those scoring =<34 are considered to have MCI.

LIST OF PUBLICATIONS AND RESEARCH GRANT

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.

RESEARCH GRANT

1. 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.