Design of Dimensional Model for Clinical Data Storage and Analysis

Dipankar SENGUPTA[*], Priyanka ARORA, Shradha PANT, Pradeep K. NAIK

Dept. of Biotechnology & Bioinformatics, Jaypee University of Information Technology, Waknaghat, Solan, H.P., India. Waknaghat, Solan, Pin – 173234 Himachal Pradesh, India.
E-mails: dipankarsengupta.1982@gmail.com*; priyankaarora1090@gmail.com; shradhapant89@gmail.com; pknaik1973@mail.com
* Author to whom correspondence should be addressed Tel. - +91-1792-239313; Fax. - +91-1792-245362; Mo. - +91-9805986683

Received: 1 March 2013 / Accepted: 6 June 2013 / Published online: 6 June 2013

Abstract
Current research in the field of Life and Medical Sciences is generating chunk of data on daily basis. It has thus become a necessity to find solutions for efficient storage of this data, trying to correlate and extract knowledge from it. Clinical data generated in Hospitals, Clinics & Diagnostics centers is falling under a similar paradigm. Patient's records in various hospitals are increasing at an exponential rate, thus adding to the problem of data management and storage. Major problem being faced corresponding to storage, is the varied dimensionality of the data, ranging from images to numerical form. Therefore there is a need for development of efficient data model which can handle this multi-dimensionality data issue and store the data with historical aspect.

For the stated problem lying in façade of clinical informatics we propose a clinical dimensional model design which can be used for development of a clinical data mart. The model has been designed keeping in consideration temporal storage of patient's data with respect to all possible clinical parameters which can include both textual and image based data. Availability of said data for each patient can be then used for application of data mining techniques for finding the correlation of all the parameters at the level of individual and population.

Keywords: Data Warehouse; Data Mart; Dimensional Model; Slowly Changing Dimension

Introduction

A major problem being faced by most of the organizations & industries around the world is with respect to efficient storage of huge amount of data and its maintenance. Most of the financial services, telecom giants & other service providers hence forth have tried to take help from information technology for getting storage solutions. Besides 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 which can be associated with any organization corresponding to its 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 Data Warehouse (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 Life & Medical Sciences, where huge amount of data is being 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 of 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. However, clinical informatics makes the representation and interpretation of complex medical terms quite simple. 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, and history with the disease, and treatment description etc. The type of data varies from a needle to a sophisticated machine. 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 science, it's a need that basic science and clinical researchers work together on interpretation and application of research data in clinical settings. Data sharing is necessary to improve the quality of healthcare and 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. 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.

Data warehouses are 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 denormalized 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 historical data, a clinical dimensional model is being proposed. This will also 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 1, 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.

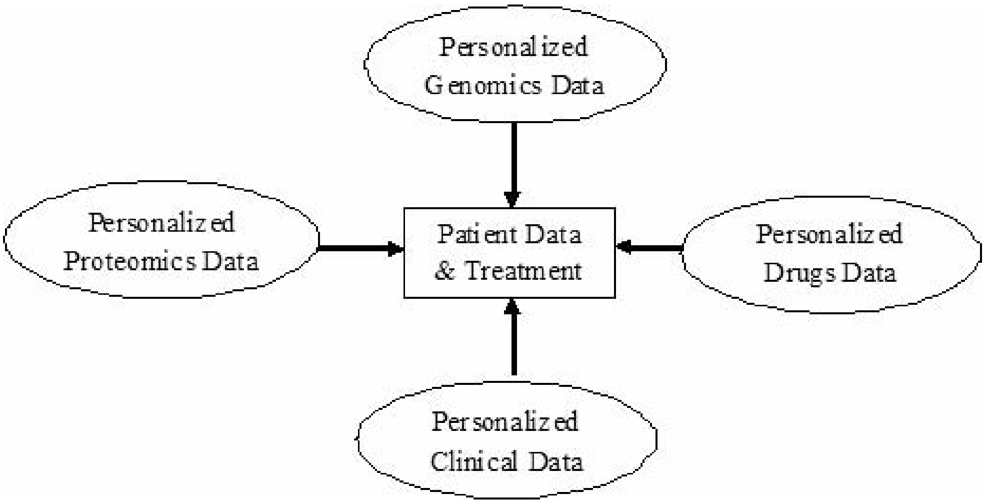
Personalized Genomics Data
Personalized Proteomics Data
Patient Data & Treatment
Personalized Drugs Data
Personalized Clinical Data

Figure 1. Domains associated to health care

(Different domains which can be associated with health care in future)

Material and Method

Proposed Design for the Clinical Dimensional Model

The advanced form of clinical data warehouses as discussed would be complex and time consuming to review a series of patient records. However, they are going to be the efficient data repositories 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 the current advents the clinical domain associated with the health cycle needs major attention. The major problem being faced is of varied dimensionality, ranging from images to numerical form of data which needs to be answered. Based on the same we propose for an appropriate clinical dimensional model (Figure 2 - logical data model representation of clinical dimensional model) for the structure of a clinical data mart which would store the data at granularity of an individual corresponding to time. The given model has been designed using Erwin data modeller (version 8.2) [7] and is represented as star schema representation [8]. While designing the model it has also been taken into consideration that the data in some dimension may change, for which additional attributes have been added, such that it can act as slowly changing dimension (SCD) [9-11]. During its physical implementation for a data mart an SCD-Type II implementation can be made for such dimensions [9]. The data extraction, cleaning & processing process can be carried out using ETL technology and any optimal existing RDBMS package can also be used to physically create the data warehouse.

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 & internet.

Results

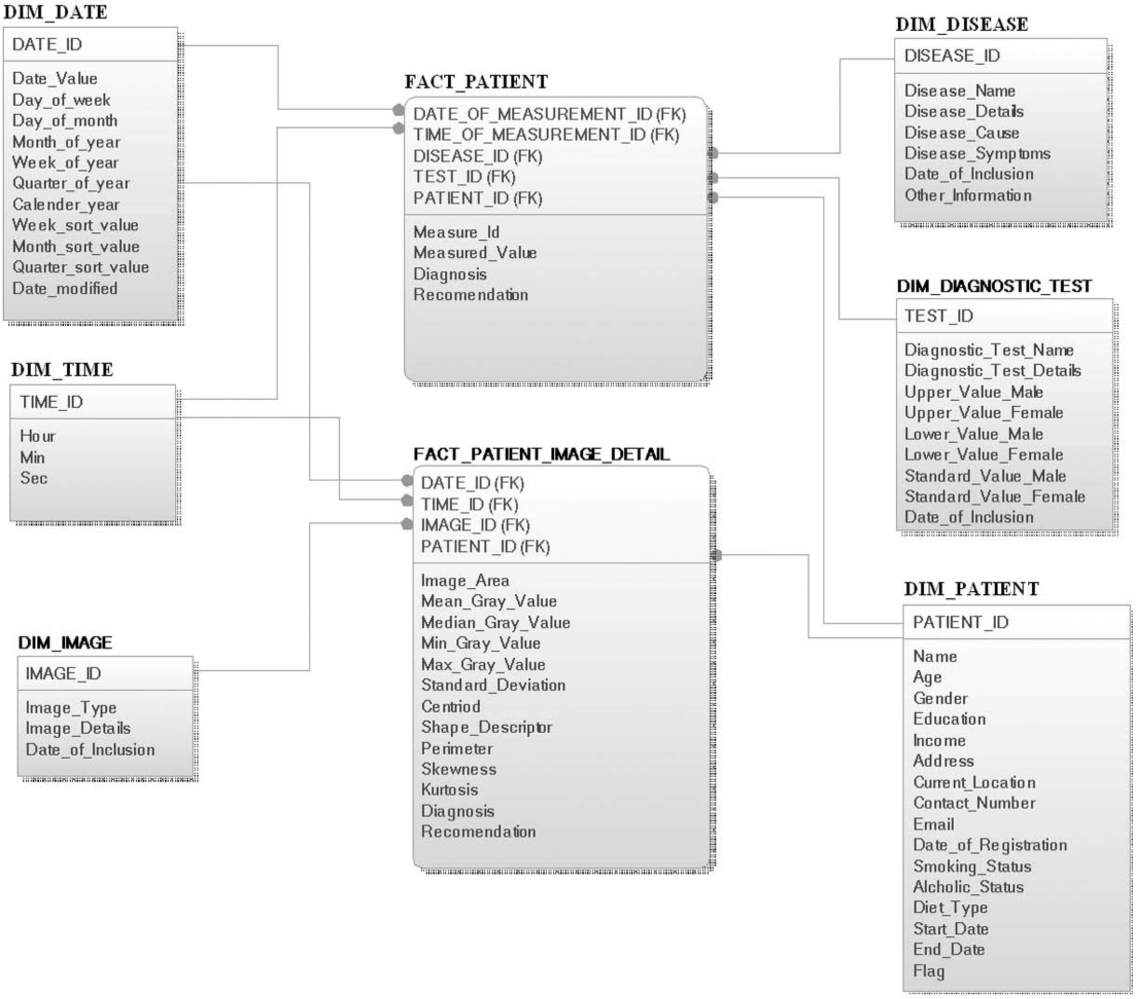

Figure 2. Logical representation of clinical dimensional model (logical data model for the clinical data mart)

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 so that it can act as slowly changing dimension. While designing the clinical dimensional model the historical tracking prospect was taken into consideration, hence forth Date & Time dimension was included. Date_Id is the primary key for Dim_Date dimension, which assigns unique id to each of the date value. The dimension also include various date based attributes like month, week, calendar year, quarter, etc., which can help to make an analysis considering different period. Time_Id is the primary key for Dim_Time which assign unique id corresponding to each second of a minute and hour. Separate inclusion of Time dimension ensures irrespective of number of times a test is conducted for a patient on any given date, each measure would be recorded uniquely in the Fact_Patient table. 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.

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 [12-15]. Nagarajan *et al.* [15] 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 [16]. Some of the researches propose use of knowledge-based terminology for identifying data dimensions in clinical informatics [17] and on the conceptual development of IDs using ontology-based systems for the design and integration of clinical data [18]. 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 [16]. 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 [18]. 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.

Conclusions

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 information of an individual will be correlated with his/her clinical information aspect. As Clinical informatics is the study of information systems (computers and programs) used in the clinical practice of medicine, so our honest attempt might contribute in the following ad-hoc aspects:

- **Data Entry**- The hospitals can keep a complete record of their patients in the form of electronic health record (EHR).
- **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-This is being targeted using the data mining techniques.
- **Decision Support**-Immediate feedback at the time of order entry about any correlation among various parameters of that patient can be shown to reduce both patient morbidity and healthcare costs.

List of abbreviations

DW – Data warehouse
IDs – Integrated Data Systems
SCD – Slowly Changing Dimension
ETL – Extraction, Transformation, Loading
EHR – Electronic Health Record

References

1. Kimball R. The Data Warehouse Lifecycle Toolkit 1998. John Wiley & Sons, Inc., New York.
2. Kimball R, Ross M. The Data Warehouse Toolkit, 2nd edition 2002. John Wiley & Sons, Inc., New York.
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, et al. Slim-Prim: A Biomedical Informatics Database to Promote Translational Research. Pers in Health Inf Manag 2009;6(6).
5. Corey JM, Abbey M, Abramson I, Taub B. Oracle8 Data Warehousing – A Practical Guide to Successful Data Warehouse Analysis, Build, and Roll-Out 1998. Osborne/McGraw-Hill, Berkeley.
6. Ross M. The 10 Essential Rules of Dimensional Modeling - Kimball Group [Internet]. 2009 [updated 2009 Mar 29; 2012 Oct 31]. Available from: http://www.kimballgroup.com/2009/05/29/the-10-essential-rules-of-dimensional-modeling/.

7. CA, Tech. ERWIN DATA MODELER (data modeling software system), version 8.2 2012. http://erwin.com/products/data-modeler. *(software source)*
8. Mozes A. Mining Star Schemas A Telco Churn Case Study [Internet]. 2011 [updated 2011 Feb 8; 2012 Oct 31]. Available from: http://www.oracle.com/technetwork/database/options/advanced-analytics/odm/odmtelcowhitepaper-326595.pdf.
9. Kimball R. Slowly Changing Dimensions, Types 2 and 3. Kimball Group [Internet]. 2009 [updated 2008 Sep 22; 2012 Oct 31]. Available from: http://www.kimballgroup.com/2008/09/22/slowly-changing-dimensions-part-2/
10. Browning D, & Mundy J. Data Warehouse Design Considerations [Internet]. 2001 [updated 2001 Dec; 2012 Oct 31]. Available from: Microsoft. http://msdn.microsoft.com/en-us/library/aa902672%28v=sql.80%29.aspx .
11. Lunexa. White Paper on Slowly Changing Dimensions [Internet]. 2011 [updated 2011; 2012 Oct 31]. http://www.lunexa.com/documents/10923/12405/SlowlyChangingDimensions.pdf.
12. Brazhnik O, Jones J. Anatomy of Data Integration. J of Biomed Inf 2007; 40(3) 252–269.
13. 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. 2007 IEEE: 53-55.
14. Wang K, *et al* . BioMediator Data Integration: Beyond Genomics to Neuroscience Data. AMIA Annual Symposium Proceedings 2005:779-783.
15. 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.
16. Dinu V, Nadkarni P. Guidelines for the Effective Use of Entity–Attribute–Value Modeling for Biomedical Databases. Int J of Med Inf 2007:76(11):769-779.
17. Deshpande AM, Brandt C, Nadkarni PM. Metadata-driven Ad Hoc Query of Patient Data. J of the Ame Med Inf Association 2002;9:369-382.
18. Anhøj J. Generic Design of Web-Based Clinical Databases. J Med Internet Res 2003;5(4):e27.