A Data Warehouse Architecture for Clinical Data Warehousing

Tony R. Sahama and Peter R. Croll

Faculty of Information Technology
Queensland University of Technology
PO Box 2434, Brisbane 4001, Queensland

t.sahama@qut.edu.au

Abstract

Data warehousing methodologies share a common set of tasks, including business requirements analysis, data design, architectural design, implementation and deployment. Clinical data warehouses are complex and time consuming to review a series of patient records however it is one of the 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. The presented data warehouse architectures are practicable solutions to tackle data integration issues and could be adopted by small to large clinical data warehouse applications.

Keywords: Clinical Data Warehouse, Data Integration, Data Warehousing, Data Design, Data Warehouse Architecture.

1 Introduction

Comparative reduction in computing cost together with the explosion and wide spread internet access has led a rapid expansion of Biomedical Knowledge Repository (**BKR**). The vast and complex compendium of molecular biology knowledge is available today in electronic databases, often accessible via the internet [e.g., GenBank, GDB, Swiss-Prot, PDB, OMIM, ENZYME] (Sujansky, 2001; MOLBIO). Also, "the clinical domain is one in which a plethora of data exists in repositories distributed across the globe, crossing institutional, regional and national boundaries. To be able to harness this data and move it across these boundaries has the potential to provide great scientific and medical insight, to the benefit of many protagonists in the field of clinical medicine" (Stell *et al*, 2006). Turning the specific clinical domain information (e.g., **BKR**) to a Clinical Data Warehouse (**CDW**) can facilitate efficient storage, enhances timely analysis and increases the quality of real time decision making processes. Such methodologies share a common set of tasks, including business requirements analysis, data design, architecture design, implementation and deployment (Inmon, 2002) and (Kimball *et al*. 1998).

The **CDW** is a place where healthcare providers can gain access to clinical data gathered in the patient care process. It is also anticipated that such data warehouse may provide information to users in areas ranging from research to management (Sen, 1998). In this connection, establishment of the data design such as data modelling, normalisation and their attributes which facilitate measurements of the effectiveness of treatment, relationships between causality and treatment protocols for systemic diseases and conditions are captured. The realisation of the need to address safety and avoid adverse outcomes in a clinical setting (Wolff & Bourke 2001) has promoted the need of effective **CDW**s (Ledbetter & Morgan 2001) and (Pedersen *et al* 1998). On the other hand, creating breakdowns of cost and charge information or forecasting demand to manage resources from the management perspective are a necessary requirement (Sen 1998).

Currently, a Clinical Data Store (**CDS**) needs to address several issues with Clinical Data Management Systems (**CDMS**). They are namely, data location, technical platforms, and data formats; organisational behaviours on processing the data and culture across the data management population. These factors are vital and unless these barriers are broken, the required levels of quality decision making and analytics can not be achieved when designing practical data warehouse architecture. Furthermore, it is a practicable strategy considering the time factor for those issues when integrating different data locations. For example, the fate of a patient's record from admission and throughout their lifetime and even beyond will need careful consideration. Hence, some of this information must be captured into the **CDW** over the long term. Storage of such sequences of information will raise another series of queries as to how long such information is required to be stored in the **CDW**. Furthermore we should establish whether this information is time dependent (which means, is it non-volatile data?)

The **CDS**s are containing "islands'' of information across various departments, laboratories and related administrative processes, which are time consuming and laborious tasks to separately access and integrate reliably. Clinical practices and their routines in different institutions, e.g. public verses private hospital, differ significantly and could benefit greatly from the integration of these information islands however the existence of heterogeneity of the data sources often delays such effort. Integration of those kinds of data stores are challenging tasks and an important problem to tackle and resolve in the **CDW** arena. This effort would be a timely solution for present-day health care requirements.

Data acquisition and information dissemination in a knowledge-intensive and time-critical environment presents a challenge to clinicians, medical professionals, statisticians and researchers. As computer technology becomes more powerful, it becomes possible to collect data in volume, and to a level of detail that could not even be imagined just a few years ago. At the same time, it offers a growing possibility of discovering intelligence from data through database marketing, information retrieval and statistical techniques such as Exploratory Data Retrieval, Data Analysis and Data Mining. A recent development in information technologies (Figure 1) in particular Database Management Systems (DBMS) has been extensively used for "Decision Support". Such Decision Support Systems (DSS) allow analytical queries, statistical queries and real time reporting from data collected for many applications especially in Online Transaction Processing Systems (Friedman, 1997).

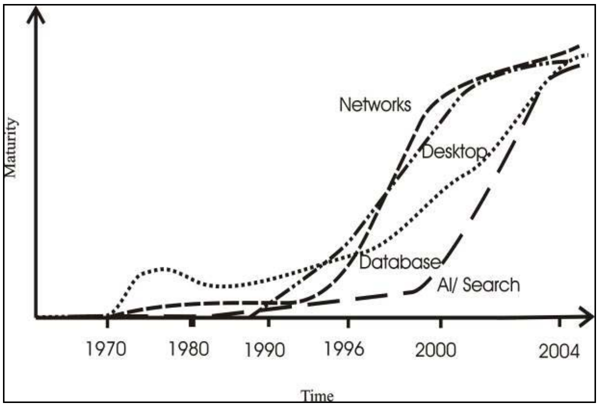

Figure 1: Technological Maturity [primary source: Dhar and Stein (1997)]

A DSS requires the construction of a "Data Warehouse" in order to complete its life cycle. A DW unifies the data scattered throughout an organization into a single centralized data structure with a common format. A fundamental concept of a DW is the distinction between data and information. Data is composed of observable and recordable facts that are often found in operational or transactional systems. A DW is a repository of integrated information, available for querying and analysis (Inmon, 1992). Thus, data warehousing may be considered a "proactive" approach to information integration, as compared to the more traditional "passive" approaches where processing and integration starts when a query arrives.

For instance, healthcare organisations practicing evidence-based medicine strive to unite their data assets in order to achieve a wider knowledge base for more sophisticated research as well as to provide a matured decision support service for the care givers. The central point of such an integrated system is a data warehouse, to which all participants have access (Stolba et al. 2006).

Of another situation similar to healthcare organisation, where building medical data warehouses for research purposes are worth exploring. Szirbik *et al.* (2006) used rational unified process (RUP) framework when designing a medical data warehouse for elderly patient care systems. Such methodology emphasized current trends, as early identification of critical requirements, data modelling, close and timely interaction with users and stakeholders, ontology building, quality management, and exception handling. This medical data warehouse delivered stakeholders to perform better collaborative negotiations that brought better solutions for the overall systems investigated. As a result, better decision making processes were established that led to a social impact and enhanced global outcomes.

1.1 Building a Data Warehouse (DW)

The **DW** is a data structure that is optimized for distribution, mass storage and complex query processing (Figure 2). It collects and stores integrated sets of historical data from multiple operational systems and feeds them to one or more Data Marts, which are data structures that are optimized for faster access. It may also provide end-user access to support enterprise views of data. A **DW** can potentially provide numerous benefits to an organization with quality improvement, and decision support by enabling quick and efficient access to information from legacy systems and linkage to multiple operational data sources. Recent research shows that the key factors for successful **DW** implementation are organisational in nature; management support and adequate resources are most important because these address political resistance. The **DW** is that portion of an overall Architected Data Environment that serves as the single integrated source of data for processing.

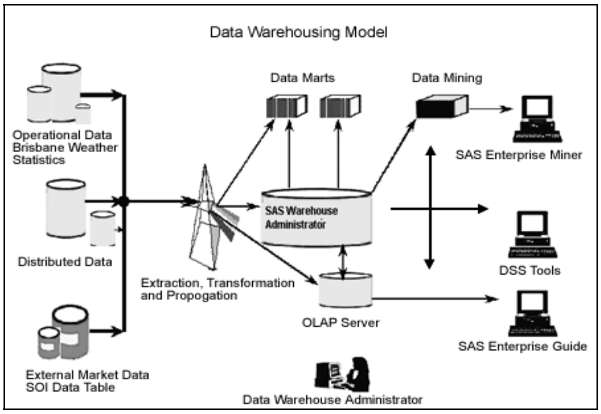

Figure 2: Example Data Warehousing Model [primary source: Marempudi, (2001)]

The **DW** concept has been around since 1989. Its definition largely depends on the background and views of the definer. According to Kimble and Inmon (1996), the **DW** is **Subject-Oriented**, **Integrated**, **Time-Variant**, **Non-Volatile** data in support of management decisions.

- **Subject-oriented** means that all relevant data about a subject is gathered and stored as a single set in a useful format. Information is presented according to specific subjects or areas of interest;
- **Integrated** refers to data being stored in a globally acceptable fashion with consistent naming conventions, measurements, encoding structures, and physical attributes, even when the underlying operational systems store the data differently;
- **Non-volatile** means stable information that does not change each time an operational process is executed. Information is consistent regardless of when the warehouse is accessed;

- **Time-variant** means that the data warehouse contains a history of the subject, as well as current information. Data warehouse data represents long-term data from five to ten years in contrast to the 30 to 60 day time period of operational data;

Data warehousing is a process requiring a set of hardware and software components that can be used to better analyse the massive amounts of data that organisations, companies and research disciplines are accumulating to make better operational and/or strategic decisions. The data warehousing process does not consist of just adding data to the DW, but also requires the architecture and tools to collect, query, analyse and present information. "Data warehousing is a process, not a product, for assembling and managing data from various sources for the purpose of gaining a single, detailed view of part or all of a business"(Stephen 1998).

2 The approach

Although there are several technical issues as indicated above that challenge building a data warehouse solution and designing data warehouse architecture. Our approach was to experiment with the known and available **BKR** (e.g., Oncology and Mental Care). This approach had been taken by the mutual understanding of a Queensland base industry partner who provide Information Technology solutions to health care providers. Due to the confidentiality of healthcare data, and the privacy policy of the participating health care organisation, the proposed experimental data and information is not augmented physically. The data structure and alias names is used instead. Most of the data design and attributes in this experiment is an abstract only. We maintain such status of the data in order to preserve the privacy and protect intellectual properties as agreed with the collaborating industry partner.

We explored and experimented with the a few data warehousing methodologies (Figure 3) proposed by Sen and Sinha (2005). We have taken brutal steps not to follow the conventional, relational database paradigm such as normalisation (utilising structures that break available information into pieces) and minimise data duplications. There are no longer issues and disadvantages with duplicating the data as storage is effectively free or very low cost. The duplicated data must be consistent throughout the process when ever necessary to maintain the data integrity (Kroenke, 2005).

During the design and planning stage of the application phase, we used a business analytics approach where a small team comprised of the data warehouse architect, business analyst and expected users of the **CDW** to understand the key processes of the business. In this connection, it is understood that the architect typically works with a business analyst, business leaders and expected users of the **CDW** to understand the key processes of the business and the questions business leaders and other users of the warehouse would ask of those processes (Gray, 2004). Patient management scenarios in the Oncology is somewhat general to the patient care process however, one application area might be unit census, where analysis is conducted on admissions, discharges and transfers by patient demographic, diagnosis, severity of illness, and length of stay. Another application area might be the care planning process, where problems, planned interventions, and expected outcomes are compared against standard care plans and expected results. The data (in fact the information) for those areas are complex and there are hundreds of duplicated data attributes.

In contrast, patient management scenarios in the Mental Health discipline are different. In this context, it is an essential element to integrate strategic use of information to plan service delivery for a non-integrated environment. This environment includes paucity of useful information to monitor health service activities and investigate patient outcomes. The middle and/or senior management could not effectively monitor levels of team activities, or determine which factors were predictive of the clinical outcomes of mental health patients. Providing such reports or reporting capabilities are essential for planning and to improve future service deliveries. Enabling data integration solutions must provide capabilities of producing summary reports by identifying the clinical activity of mental health teams within a given period, predictive measure of quality (good or poor) clinical outcomes of mental health patients and schedule for routine monitoring of the clinical outcomes of mental patients by senior management.

The ability to integrate all of this data for purposes of analysis and actionable knowledge defines the emerging technical arena of clinical intelligence. Leveraging years of experience in the broader business community with extensive data warehousing and business intelligence initiatives, the healthcare industry now stands on the brink of an exciting new era in which lower costs and higher quality of care can exist side by side. No longer is it necessary to manually select data from the different (and often proprietary) silos in order to create the documentation that the business requires. In business analysis, the healthcare decision maker may wish to manipulate parameters and rerun the data, or generate a report that cross-references the cost of delivering a particular service in a particular demographic to a particular patient population. Whatever the business question, it is essential to realise that today's healthcare organizations are being evaluated not only on the quality and effectiveness of their treatment, but also on waste and unnecessary cost. By effectively leveraging enterprise-wide data on labour expenditures, supply utilisation, procedures, medications prescribed, and other costs associated with patient care, healthcare professionals can identify and correct wasteful practices and unnecessary expenditures. These changes benefit the bottom line and can also be used to differentiate the healthcare organisation from its competition.

3 Creating a Data Model

Having considered, the nature of both **BKR**s (e.g., Oncology and Mental Care) developing reasonably good **CDW** is challenging. The oceans of electronic data from both **BKR**s are largely decentralised by their processes and somewhat difficult to coordinate practically. Furthermore, it was a difficult task to screen potential and realised values and mix and match available software tools around to proceed with data integration. With our past experiences and availability of a promising technical development module on **CDW** in particular medical software engineering (**MSE**) capabilities, we use the SAS Data Warehouse Administrator (SAS©, 2002). The validities of this warehouse module are their flexibilities to integrate external data repositories, to facilitate the hassle free **ETL** requirements, its ability to accommodate analytics using Enterprise Miner (**EM**), and its usage to explore and report a majority of the clients' requirements via the SAS© Enterprise Guide (**EG**).

At this stage, the data integration steps were followed using the concept of "*Integration for application portability*" discussed by Sujansky (2001). This has been revealed as the standardisation of access to semantically similar information at disparate sources. For example, a universal database interface for decision-support applications that allows them to be shared across institutions with no modifications to their implementations (Sujansky *et al.* 1994).

Our experiment was to design an appropriate **CDW** by implementing a few of the data warehousing methodologies (Figure 3) discussed by Sen and Sinha (2005) by keeping the data attributes for application portability and sharing across institutions. During the design phase we encountered issues as such some of the data warehousing methodologies does not qualify for the proposed **CDW**. We experimented with all possible combinations and finally decided to implement Enterprise Warehouse with Operational Data Store Architecture (Figure 4) and Distributed Data Warehouse Architecture (Figure 5) using the SAS© Data warehousing administrator software module (SAS© 2002). We chose this avenue with an extension of including several data marts (Figure 6) for different administration and management operations (e.g., summary reports capable of being executed by team leaders, identifying the clinical activity within a given period, factors predicting the quality of clinical outcomes and routine monitoring of the clinical outcomes by senior management etc.). Furthermore, OnLine Analytical Processing (**OLAP**) tables were created to accommodate team analysis. Figures 7 and 8 depict the DW architecture implemented for Oncology patient management and Mental Health patient management CDW respectively.

Once our experiment concluded selecting a appropriate data warehouse design, a mechanism to move data from their source systems to the **CDW** was established. This step is typically referred to as the Extraction-Transformation-Load (**ETL**) which is generally known as data transformation in the DW application development. To be able to fulfil these requirements, there are several third party tools around. A summary of 15 different data warehousing methodologies classified by their core-technology, infrastructure and information modelling were presented by Sen and Sinha (2005) should provide further information. We used Microsoft Excel, SAS External File Interface (**EFI**) and SAS© Enterprise Guide (**EG**) to clean and cleanse related data. The **EG** and **EM** of the same software module had been used for reporting and further data analysis. This snap-on approach was practically achievable using SAS modules.

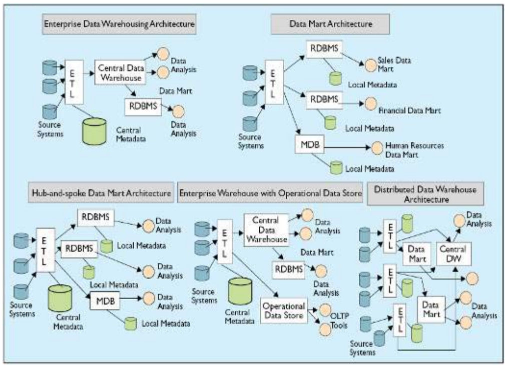

Figure 3: Different types of DW Architectures (source: Sen and Sinha, 2005)

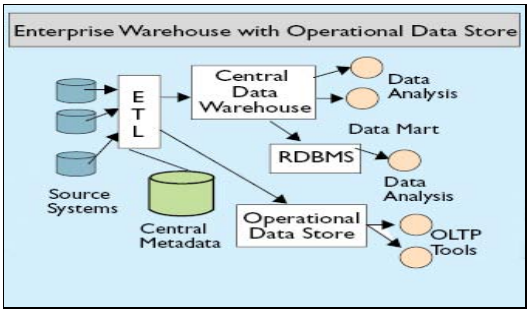

Figure 4: Enterprise DW Architecture

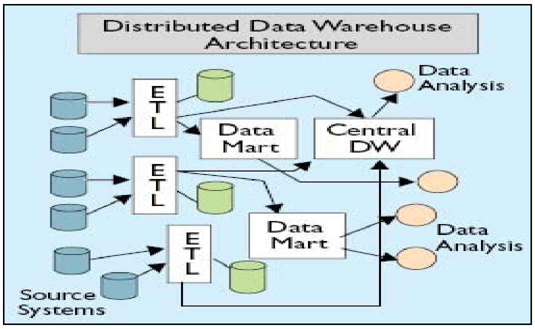

Figure 5: Distributed DW Architecture

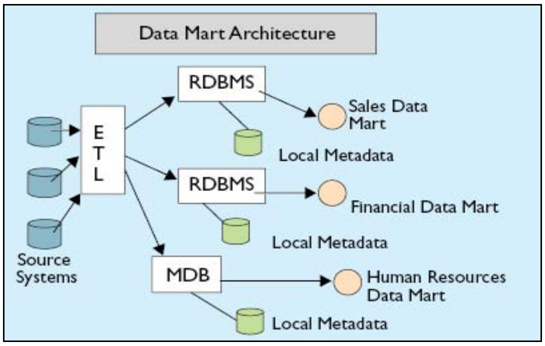

Figure 6: Data Mart Architecture

3.1 Step-by-step processes for building a DW using SAS© Warehouse Administrator:

By viewing the DW Model and its mechanics (Figures 4, 5 and 6) as related to SAS, it is possible to understand how data is transported through the system. We modified the design depicted in Figure 4 to accommodate SAS© Data Warehouse architecture as illustrated in Figures 7 and 8. The data is imported into the SAS Warehouse via the SAS Warehouse Administrator from all relevant sources using Microsoft Excel; SAS **EFI** tools and SAS Component Language (**SCL**). The SAS Warehouse Administrator is then used to create the required Data Cubes and tables needed for analysis. In the SAS environment this has being identified as Subjects Group where representing higher order and their child nodes as a Operational Data Definition (**ODD**). This can be utilised for either report analysis where graphs or tables are created or for pictorial representation using the SAS Enterprise Guide (Figure 8, e.g., POS OLAP Table). The SAS Enterprise Miner is used for predictive analysis or database analysis of data as structured by the SAS Warehouse Administrator. The following steps are an expanded view of a completed DW created in the SAS DW module using relevant clinical data sets from both Oncology and Mental Health. Completed **ODD**s are presented in Figures 7 and 8 respectively.

Step—1:

The data is first imported into SAS to allow the format to be standardized into SAS table format. The opportunity for data manipulation is available at this stage to standardize formats, create or delete relevant column data, etc. SAS can integrate with most table formats available in the commercial environment.

Step—2:

Once the data is imported into SAS, metadata is created using the **ODD**s. This allows the metadata formatting to be set (*The Warehouse Engine*).

Step—3:

The Data Tables are now created and loaded. These tables can be a mixture of any or all of the relevant data. Specific tables can be created using the SAS process editor, containing data targeted for predictive analysis or database analysis. Data marts can be created that are targeted to a specific audience, and multi-dimensional cubes and/or relational tables can be created for report generation.

By using the analysis tools within SAS, such as **EM** and **EG**, patterns which offer insights into relationships between the data that may never have surfaced without firstly warehousing the data can be visualized. These patterns may otherwise be hidden by the overpowering size of the data sets. In addition, the data warehouse data patterns can be graphed for report production and equipped for further Data Mining applications.

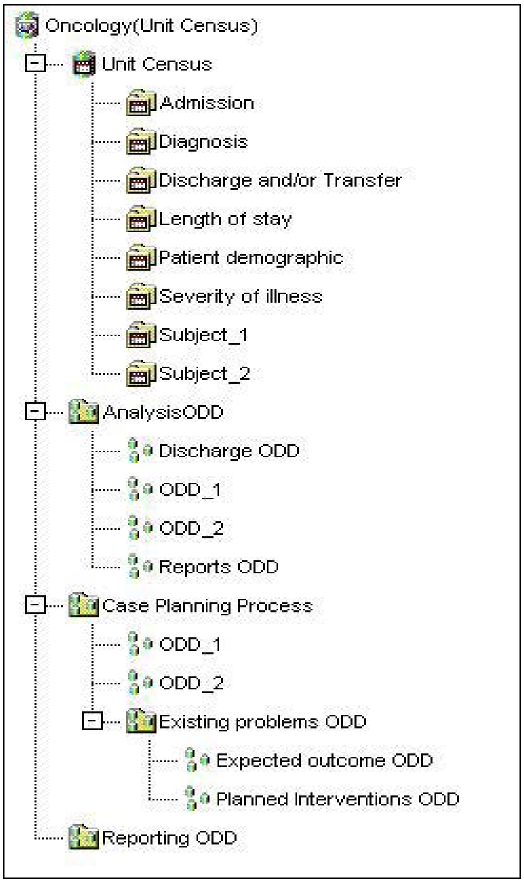

Figure 7: Oncology Patient Management CDW

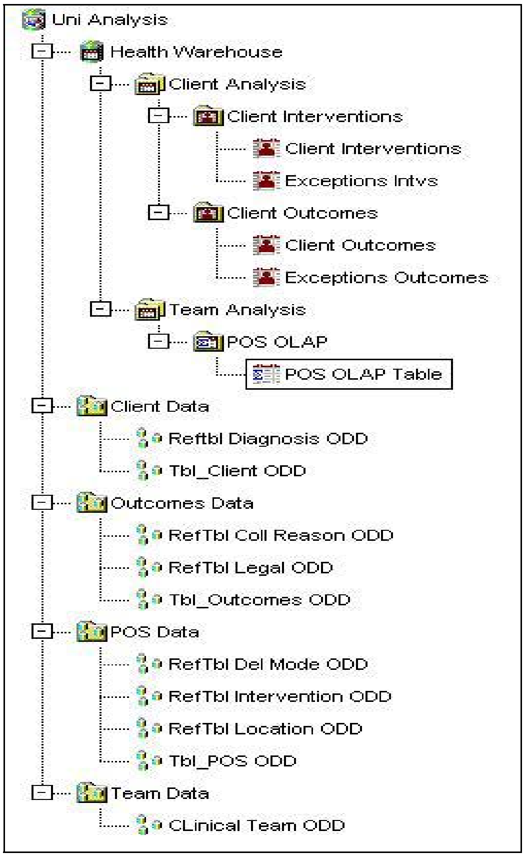

Figure 8: Mental Health CDW

4 Discussions

Data warehousing is becoming an established discipline and a valuable alternative to traditional "passive" approaches for integrating and accessing data from autonomous, heterogeneous information sources that are widely spread with mass volume which could be a bridge to a single application. The warehousing approach is particularly useful when high query performance is desired, or when information sources are expensive or transitory. A **DW** is a driving force for consolidation and integration of data structure designed for swift information access in almost all disciplines. A **DW** speeds up the data acquisition step. Although, most heterogeneous data repositories in **BKR**s are in the research stages, the Enterprise **DW** model with scattered Data Marts used to design and develop Oncology and Mental Health **CDW** were proof of such a concept bringing the decision-support with a business analytic approach. This could be adopted to small to large scale **CDW** applications.

The first step in data preparation is data acquisition, where the relevant data is identified, accessed, and retrieved from various sources, converted, and then consolidated. In many cases, the data acquisition step takes so long that there is little time left for other preparation tasks such as cleaning and transformation etc. The use of data "***integration for application portability***" approach successfully tackles these issues using SAS **EFI**, **SCL** and of course Microsoft Excel application programs. It is often remarked that data preparation and integration takes 90% of the effort for a given **CDW** project. The truth is that the modeling process could benefit from more effort than is usually given to it, but after a gruelling data preparation phase using SAS **DW** administrator, the developed **CDW** overcame most of those constraints. However, there is often not enough time left to spend on refining the shared environment that facilitates across institutions (e.g., with no modifications to their interface implementation and/or prediction models). This is another challenging task which would be opening another research area with security issues of federated data warehouses.

In the case of database analysis, SAS **EM** can be used to interpret statistical results and relationships between data not necessarily linked by anything more tenuous than date. The proposed **DW** model handles this analysis well.

The ideas presented in this paper are based on the assumption that it is undesirable to start a data integration task for a given **CDW** without prior investigation and clear understanding of the **BKR**s or data stores in bio-medicine in general. This could be resolved not just by domain specific knowledge but also through politically expedient solutions.

It would be a significant move to investigate secure access **DW** models suitable for healthcare decision support systems driven by business analytics that would accommodate case-based, evidence-based and role-based data structure. This system should be able to benefit a majority of health care sectors.

5 References

Dhar, V. and Stein, R (1997). "*Seven Methods for Transforming Corporate Data into Business Intelligence*": Prentice Hall.

Friedman, J. M. (1997). "Data Mining and Statistics: What's The Connections?". *29th Symposium on the Interface in Data Mining and the analysis of large data sets*, Houston, TX.

Gray, G.W. (2004): Challenges of Building Clinical Data Analysis Solutions. Journal of Critical Care **19**(4):264-270.

Inmon, W. (2002): Building the Data Warehouse, 3rd edition, Wiley-New York.

Inmon, W.H. (1992). EIS and the data warehouse: a simple approach to building an effective foundation for EIS. *Database Programming and Design*, **5**(11): 70—73.

Kimbal, R., Reeves, L., Ross, M. AND Thronthwaite, W. (1998): The Data Warehouse Lifecycle Toolkit, Wiley- NY.

Kimball. R. and Inmon, W.H. (1996). *The Data Warehouse Toolkit.* John Wiley: New York

Kroenke, D.M.(2005): Beond the relational database model. *Computer,(IEEE Computer Society)*, **38**(5):89-90.

Ledbetter C. S. & Morgan M. W. (2001). "Toward best practice: leveraging the electronic patient record as a clinical data warehouse." Journal of Healthcare Information Management 15(2): 119-31

Marempudi, A. (2001): Data Warehousing for better decisions. Intellibusiness, www.datawarehousing.com

Torben Bach Pedersen and Christian S. Jensen (1998). Clinical Data Warehousing - A Survey. VIII Mediterranean Conference on Medical and Biological Engineering and Computing (Medicon98). Limassol, Cyprus.

SAS© (2002): Building a Data Warehouse Using SAS/Warehouse Administrator®, Software Course Notes (Book code58787). SAS Institute Inc., Cary, NC 27513, USA.

Sen, A. and Sinha, A. P. (2005): A Comparison of Datawarehousing Methodologies, Communication *of the ACM*, **48**(3), 79-84.

Sen, A. and Jacob, V. S. (1998): Industrial Strength Data Warehousing, Communication of the ACM, **41**(9), 28-31.

Stephen R. (1998) .Building the Data Warehouse., Communications of the ACM, 41(9), 52-60 (September 1998).

Stell, A., Sinnott, R. and Ajayi, O. (2006): Secure Federated Data Retrieval in Clinical Trials. Proc. IASTED International Conference on Telehealth, Banff, Alberta, Canada, July 3-5, 2006.

Stolba, N., Banek, M. and Tjoa, A.M. (2006): The Security Issue of Federated Data Warehouses in the Area of Evidence-Based Medicine. Proc. of the First International Conference on Availability, Reliability and Security (ARES'06, IEEE), 20-22 April, 2006.

Sujansky, W. (2001): Methodological Review-"Heterogeneous Database Integration in Biomedicine". Journal of Biomedical Informatics. **34**:285-298.

[MOLBIO] http://molbio.info.nih.gov/molbio/db.html

Szirbik, N.B., Pelletier, C. and Chaussalet, T. (2006): Six methodological steps to build medical data warehouses for research. International Journal of Medical Informatics **75**(9):683-691.

Wolff A.M, Bourke J, Campbell, I & Leembruggen DW (2001). "Detecting and reducing hospital adverse events: outcomes of the Wimmera clinical risk management program." Medical Journal of Australia 174: 621-625.