INTEGRATING CLINICAL DATA FROM MULTIPLE SOURCES, USING A DATA MART

Case Study : MENTORS Project , Infectious Disease Institute Limited

KUTEESA DAVID MANSEN K

RegNo: 2012-M132-20008

Uganda Martyrs University

August 2016

INTEGRATING CLINICAL DATA FROM MULTIPLE SOURCES, USING A DATA MART

Case Study : MENTORS Project of Infectious Disease Institute Limited

A postgraduate dissertation
presented to the Faculty of Science in partial
Fulfillment of the requirements for the award of
the degree of Master of Science in Information Systems

Uganda Martyrs University

KUTEESA DAVID MANSEN K

RegNo: 2012-M132-20008

August 2016

DEDICATION

This dissertation is lovingly dedicated to my mom, Christine Musisi Kanwa. For her endless support, encouragement, and constant love all through my experiences.

ACKNOWLEDGEMENT

First of all, I would like to thank God, the Almighty, for having made everything possible by giving me strength and courage to do this work.

My deepest gratitude to Dr Ssembatya Richard, my supervisor, for his unselfishness, encouragement and guidance accorded to me during my study.

Sincere thanks to my family and friends who all gave me courage and support.

I am also deeply indebted to all those who did the proofreading. They too have also contributed towards my success.

DECLARATION

I Kuteesa David Mansen declare that this work has been produced by me. I hereby state that this work is my own. It has not been submitted to any other institution for another degree or qualification, either in full or in part. Throughout the work I have acknowledged all sources used in its compilation.

Name of Researcher

…………………………………………………………………………………………………

Researcher's Signature

…………………………………………………………………………………………………

This work has been produced under my supervision

Name and signature of Supervisor

…………………………………………………………………………………………………

Date of Submission

………………………………………………………………..

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS AND ABBREVIATIONS

CDMS	- Clinical Database Management System
CDR	- Clinical Data Repository
EAI	- Enterprise application integration
EDW	- Enterprise Data Warehouse
EII	- Enterprise information integration
ETL	- Extraction, Transformation, Load
IDI	- Infectious Disease Institute
IHE	- Integrated Health Enterprise.
HL7	- Health Level Seven
OLTP	- Online Transaction Processing System
RAD	- Rapid Application Development
RIM	- Reference Information Model
SSAS	- SQL Server Analysis Services
SSRS	- SQL Server Reporting Services
UML	- Universal Modeling Language

ABSTRACT

Many healthcare organisations have seen tremendous increase in the volume and complexity of healthcare data. Oftentimes healthcare data is stored in many dispersed databases, duplicated in many cases and stored in a variety of formats. The process of producing information from these multiple dispersed systems is quite difficult and time consuming. Consequently essential information required to make medical decision and improve healthcare service delivery is not readily available for medical practitioners and healthcare providers.

Delivering quality healthcare requires the integration of healthcare information from many different sources, and healthcare providers must be able to readily access and use the right information at the right time in order to improve the quality of health service delivery. A variety of approaches are available to integrate healthcare information stored in many different sources. However they are limited in their application.

This research uses a data mart model to integrate clinical data from the multiple clinical data sources gathered by MENTORS project at the Infectious Disease Institute Ltd (IDI) to ease analysis of clinical data and improve availability of information used for decision making and clinical research.

CHAPTER ONE

INTRODUCTION

1.0 INTRODUCTION

Integration of data sources refers to the process of creating a common schema as well as data transformation solution for a number of data sources with related content (Koeller, 2006). In the clinical domain, data integration involves the capture, cleansing and storage of data from clinical data sources (HIMSS, 2013). Clinical data sources manage data used for clinical studies or trials and the data stored in the repositories is used for supporting patient care and retrospective clinical research (Ogbuji, 2009). To support clinical functions, clinical data must be gathered from varying data sources (Ogbuji, 2009). However, oftentimes data is stored in many dispersed databases, duplicated in many cases and stored in a variety of formats (Boterenbrood, et al., 2014). Data required to make medical decisions is not properly integrated or fully utilized. Consequently provision of healthcare is hindered, due to unavailability of clinical data to be utilized by care providers to perform activities such as diagnostics, prognostics and optimization to improve patient care.

Provision of effective healthcare requires integrated, high quality data (Oracle, 2011). However integrating data stored in disparate data sources into a central repository to create one single view for all users is a major information technology challenge (Chenhui, et al., 2008). The larger number and size of modern data sources makes manual approaches to data integration increasingly impractical (Koeller, 2006). In addition, information system processing issues arise, when massive sets of data are requested from their dispersed source systems for decision making purposes (Boterenbrood, et al., 2014).

Bridging the gaps resulting from deployment of disparate systems, are data warehouses which provide a powerful solution for data integration and information access problems (Sheta and Eldeen, 2013). Data warehousing can help to partially or fully automate the data integration process. (Koeller, 2006). Inmon (1993), defines a data warehousing as subject oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process.

Due to the constraint of time, associated with implementing a data warehouse, this research emphasizes the use of data marts, as an alternative solution for integrating disparate data sources. Building a data warehouse can take longer period of time hence most organizations opt for data marts. Data marts support only the requirements of a particular department or functional area and can therefore be built more rapidly (Connoly and Begg, 2005).

1.1 Background to the Study

In recent years, there has been a tremendous increase in the volume and complexity of data available to the health research community (Branson, et al., 2008). Branson, et al. (2008) further argues that to enable the use of this information in clinical studies, users generally require an integrated view of health data across a number of data sources. However, due to the constraints of time and resources, most organizations adopt a "one-thing-at-a-time" approach to developing islands of information systems, which results into a number of uncoordinated and often inconsistent databases (Hoffer, et al., 2011). Besides, Riazati

(2012), states that dispersed information is difficult to use effectively and therefore is not of high knowledge value. In the clinical domain, the end users of medical data analysis systems do not understand storage structures and access mechanisms for the dispersed data sources. Thus they need simplified mechanisms for integrating various disparate data sources in order to have a holistic view of patient information and thereby deliver personalized healthcare (Branson, et al., 2008).

Obtaining a single view of the customer still remains an elusive goal for many organization (Ballard, et al., 2003) and hence many fields of research have shown a great need to integrate data from different sources (Koeller, 2006). The clinical domain (laboratory test, medication, so on) being a data rich environment, it was thus chosen for this research. This domain is composed of a multitude of different data entities and in addition data are generated from various information system, often from different vendors (Patel and Weng, 2009). Integrating such a wide variety of data streams into a common information model is a challenging task. On the other hand, developing a consistent view of clinical data collected from various clinical areas would facilitate efficient storage, enhance timely analysis and increase the quality of real time decision making in the clinical domain (Sahama and Croll, 2007).

The infectious Disease institute (IDI), is a non-governmental organization dedicated to the treatment of patients and education of physicians from all over Africa. Over the years the organization has expanded its operations, with several projects and studies having been implemented, or are ongoing. Most of projects are focused on clinical research and they come with different data management demands. Clinical data is a key item of the data management process in the organization, thus projects collect massive volumes of clinical data from various project sites to support their reporting function and also back research efforts. But due to limited time and resources, many projects develop fragmented data stores which are usually in various formats such as flat files, relational databases and xml. In particular the MENTORS project which is the main focus of this research study, is run under the training department of IDI. The project operates in 10 health facilities across the country with the main objective of improving care and treatment of TB and HIV/AIDS patient in low resource setting. In order to carry out project planning, decision making, data analysis and dissemination of research finding, the MENTORS project staff mainly rely on clinical data gathered from the different implementation sites. The clinical data is distributed in numerous operational clinical data system and stored in different database formats such as Microsoft Access and MySQL server, which makes the process of building a comprehensive view of clinical information tailored for decision making and research purposes difficult.

The MENTORS project utilizes a reporting application developed using Ms Access, but this application accesses the different clinical databases separately and has limited analytical capabilities. Furthermore, when decision makers at various levels need information to get an insight into multi-site performance, this information is not readily available, and if successfully generated, it does not guarantee high quality data due to human errors introduced in the system during data entry, and the duplication of patient records stored in the different data systems. Much of the difficulty with existing information management process is due to the unavailability of tools which can easily and accurately avail information required for the decision making and research purposes.

Sahama and Croll (2007) argues that clinical data stores containing islands of information across various operating requirements are time consuming and laborious tasks to separately access and integrate reliably. However, Sahama and Croll (2007) propose the clinical data warehousing solution as one that can facilitate efficient storage, enhances timely analysis and increases the quality of real time decision making processes. IL-Yeol (2009) defines a data warehousing system (data mart or data warehouse) as an environment that integrates diverse technologies into its infrastructure. It was therefore against the aforementioned background that the researcher suggested the adaptation of the data mart design approach to develop a solution capable of consolidating the multiple clinical data sources gathered by the MENTORS Project of IDI into single data repository, to ease data analysis process and improve availability of information for decision making and research purposes.

1.2 Problem Statement

The MENTORS project of IDI requires clinical data for planning, decision making and dissemination of research findings which contribute to the country's healthcare policy. However there is limitation in the accessibility of clinical data gathered by the project from the different implementation sites due to the existing setup, whereby clinical data is distributed in multiple operational clinical data sources and in different formats. Consolidating the multiple clinical data sources for reporting is almost difficult to achieve as it involves extensive manual processes and is time consuming. In addition validation and cleaning of data errors resulting from data entry is also a difficult and lengthy process.

Kimball and Ross (2013) offer suggestion to healthcare organizations struggling with the many disparate systems, to integrate the information in the various systems in order to deliver more effective patient care. Sahama and Croll (2007), consider data warehousing as the practical approach to information integration, as compared to the more traditional static approaches where processing and integration starts when a query arrives.

1.3 Objectives

1.3.1 Main Objective

The main objective of this project is to build a consolidated view of clinical data from multiple data sources, using a data mart design approach, to ease analysis of clinical data, improve the availability and quality of information required for decision making and clinical research in the MENTORS project at the Infectious Disease Institute Ltd (IDI).

1.3.2 Specific Objectives

1. To study and analyze the current system, review existing literature on clinical domain data integration.
2. To design a Data Mart that would provide an integrated view of multiple clinical data sources.
3. To implement a Data Mart to enable users to flexibly access the integrated view of clinical data from multiple data sources.
4. To test and validate the Data Mart using a case study.

1.4 Scope of Study

This research was carried out in the MENTORS project, running under the training department of the Infectious Disease Institute. The research project focused majorly on the design, development and implementation of a data mart system to provide a single physical repository of the multiple clinical data sources collected by the MENTORS project from its 10 study sites across the country. The data mart can be used to support in-depth data analysis, efficient reporting and querying of information that can shared to intended users for decision making support. The intended users of the developed data mart system will the project manager, project section heads, project clinical staff and project data staff.

1.5 Justification and Research Contribution

Data access and management have become some of the prime areas of importance in healthcare management (Evans, 2013). Data about the effectiveness of treatments, the accuracy of diagnoses, and the practices of health care providers is crucial to organizations that strive to maintain and improve health care delivery (Leitheiser, 2001). However many organizations have their data in more than one repository and to be able to gain meaningful insight, all relevant data has to be available (Soderlund, 2011).

The Infectious Disease Institute runs several clinical studies which collect tremendous volumes of clinical data from different study sites, for purposes of supporting research outcomes and providing project funders with quality information on the study activities and achievements. The various projects store disease specific data in fragmented information systems which are used to generate analytical reports to support the research outcomes and decision making. In particular the MENTORS project which is the focus of this research, runs different databases for the specific diseases under study. To enable the use of knowledge in clinical studies, users generally require an integrated view of medical data across a number data sources (Branson, et al., 2008). However the data integration within the project is largely manual and a time consuming process with significant input from domain experts. Henceforth end users of research data cannot quickly analyze existing data in time to aid in improved clinical care and for supporting research outcomes.

Data quality is also important for generation of reliable statistical reports, but the hospitals which provide the primary data captured into electronic databases are a chaotic environment with multiple providers taking care of a single patient record. This results into generation of inconsistent information which is often difficult to reconcile. Completeness of patient records is critical for clinicians to choose the most appropriate treatment plan for the patient. Hence validating electronic data should be an ongoing process and requires a substantial investment in time.

Given the unique data issues and analysis problems evident with clinical databases, developing a solution to integrate disparate clinical data sources, would enable end users of clinical domain data gain the following benefits:-

I. Significantly reduce the complexity, risks, time and resources for data integration and management of clinical data sources.

II. Provision of better performance of clinical transactional processing databases by moving reporting to the data mart.

III. Systematic processing of errors in clinical data source systems would be more readily identified and further corrected, reducing efforts necessary to clean clinical data.

IV. Provide end users of clinical data with the ability to formulate queries and undertake analysis on integrated clinical data sources in order to contribute to improved clinical care and outcomes.

V. With quality data, clinicians, researchers, healthcare managers, can make better decisions for influencing healthcare policy.

Finally lessons learned as a result of undertaking this research will be applicable to broader clinical data source integration.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter offers a critical review of prior studies relevant to clinical data integrating and data warehousing. The first section gives an overview of information management in health care (2.2). This covers data and information management issues related to healthcare and health information systems. The second section (2.3), discusses the diffusion innovation theory which is used to evaluate the success and significances of this project. The third section (2.4) discusses background theory of data warehousing and data mart in general and gives detailed information how data warehouses differ from operational systems. The fourth section (2.5) discusses data integration and then gives detailed information about the different data integration approaches. The next section (2.6) gives background information on data warehousing architectures and further discusses in the detail data mart designs in the subsections. Sixth section (2.7) reviews the literature on data marts and health, then covers the relevancy of data warehousing and data marts in health care. Section (2.8) discusses related work on data marts in healthcare and gives some real examples of data mart implementation in healthcare including benefits and limitations. Section (2.9) discusses related work on existing approaches to integration of clinical data in dispersed systems with real examples of clinical data integration projects and also mentioning their benefits and limitation. Finally based on the results of the literature, section (2.10) concludes this chapter by justifying the integration approaches and data warehousing architecture (data mart) selected in this research to integrate data from multiple data sources.

2.2 Managing Information in Healthcare

Information management is defined in different ways by different authors. Synott and Gruber state, the information management function provides control and management over information resources. Also, Scheyman (2002) states information management “refers to information characteristics such as information ownership, content, quality and appropriateness.

Today many healthcare organizations appear to be trying increasingly to re-organize their processes and improve the effectiveness and efficiency of their services, in order to be more competitive and reduce their costs, while also ensuring the provision of better and more personalized patient care. Information is considered to be a valuable resource and a vital element in this drive for efficiency and effectiveness (Colesca and Dobrica, 2009). Within the health industry large volumes of data are collected, containing valuable information about patients, procedures, treatments and etc (Sheta and Eldeen, 2013). However given that, decades of successful application of information technology has occurred in other information intensive industries, data in the health industry continues to be processed manually (Colesca and Dobrica, 2009). Raghupathi and Raghupathi (2014) argues that while most data is stored in hard copy form, the current trend of data storage is towards rapid digitization of these large amounts of data. Wager, et al, (2005) state that the last decade has witnessed the model for maintaining healthcare information shift from the current, primarily paper-based medical record system, in which information is often incomplete, illegible, or unavailable where and when it is needed, to a system in which the patient’s clinical and administrative information is integrated, complete, stored electronically, and available to the patient and authorized persons anywhere, anytime.

Organizations have always looked to their own data for help in making important business decisions. Revest, et al. (2005) argue that ability to efficiently store and analyze business data is a necessity for every large organization. In healthcare, the management of clinical domain data for supporting patient care and for supporting retrospective clinical research requires a repository with functions necessary for managing them over their lifetime (Ogbuji, 2009). Today, virtually every healthcare organization uses computer technology to collect, store, or retrieve all or portions of a patient's healthcare data (Nanette, 2013). The most common choice is to store data in databases, but other solutions, such as storing it in plain files, also exists (Soderlund, 2011). Besides there are numerous electronic systems from which health care providers may choose. For example, a provider can choose one electronic system for lab results and another to track medications. However the management and operation of these systems for any single health care provider is very complex and requires a team of individuals to operate and maintain, including getting the different systems "to communicate" with one another (HSCC Clinical Data Warehouse, 2013). For instance, healthcare organizations 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 (Sahama and Croll, 2007).

The current situation on data management in healthcare can have negative consequences on business efficiency and patient safety. Loper, et al. 2013 argues that transferring information from paper to a digital format always implies additional time and effort. Furthermore, this manual work is error-prone and hence might be harmful for the patient's healthcare. The huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods (Koh and Tan , 2005). Oftentimes, using clinical database management Systems (CDMS) as a pseudo data warehouse, augmented by file-based approaches, inherently fragments the data. (Palmer, 2013). Consequently, healthcare workers who wish to analyze large amounts of patient data are faced with technical challenges of integrating scattered, heterogeneous data. In addition, data exploration across related files becomes a complex undertaking as data are stored in separate files, requiring specialist programming skills to access and analyze.

Conversely, integrating data from diverse data source systems for purposes of meeting the goals of healthcare organizations poses special problems for data quality (Leitheiser, 2001). Much as electronic medical record systems have provided a platform for consistent data capture, but the reality is data capture is anything but consistent (LeSueur, 2014). Healthcare data comes from many sources and is delivered in many forms including spreadsheets and several data formats. In addition the enormous variety of data, structured, unstructured and semi structured data make healthcare data challenging. The quality of data especially unstructured data is highly variable and all too often incorrect (Raghupathi and Raghupathi, 2014). Verhulst (2006) states that dirty data in healthcare can consequently lead to medical errors, which can kill or cause long-term damage to the health of patients. Verhulst (2006) further argues, despite the severity of the problem, the risks posed by dirty data often go unrecognized and in many ways the problem of inaccurate data comes as a low priority for organizations. Henceforth it is critical to develop strategies for managing data inaccuracies and the potential harm they cause.

Because healthcare data is so uniquely complex, it’s clear that traditional approaches to managing data will not work in healthcare. A different approach is needed that can handle the multiple sources, the structured and unstructured data, the inconsistency, the variability, and the complexity within an ever-changing environment (LeSueur, 2014). Thus, data warehousing may be considered a “proactive” approach to information integration, as compared to the more traditional “passive” approaches (Sahama and Croll, 2007). The concept of data warehousing provides a powerful solution for data integration and information access problems (Sheta and Eldeen, 2013). In addition data warehousing can facilitate efficient storage, enhances timely analysis and increases the quality of real time decision making processes (Sahama and Croll, 2007).

2.3 Diffusion of Innovation Theory

The Diffusion of Innovation Theory (DOI) is one of the most popular theories for studying adoption of information technologies (IT) and understanding how IT innovations spread within and between communities (Rogers, 1995). According to the theory, an innovation is any idea, process or object perceived as new by the intended user (Rogers, 1995). Diffusion is the process by which an innovation is communicated through certain channels over time among the members of a social system (Rogers, 1995). Similarly, Daft (1978) defines an organizational innovation as “the adoption of an idea or behavior that is new to the organization adopting it”. Therefore, an innovation need not necessarily refer to a technology. It may refer to a renewal in terms of thought and action as well (Thong, 1999).

Roger (1995) described steps that an individual go through from his first knowledge of such innovation, his attitude to acceptance or rejection, to the decision to implement the new idea. The user’s final decision to accept and implement an innovation is influenced by different factors. Rogers (1995) identified the five user-perceived attributes that consistently proved to be determinants of success of an IT innovation: relative advantage, compatibility, complexity, trialability and observability. Relative advantage is the degree to which the user perceives benefits or improvements on the existing technology by adopting an innovation. Compatibility deals with how the intended users perceives the innovation will fit into their set of values, needs and experiences. Complexity is the ease or the way the innovation might by be learned by the intended user. Trialability is the ability of an innovation to be put on trial without total commitment and with minimal investment. Finally, observability refers to the positive outcomes intended users can see from the implementation of such innovation.

An important part of the analysis and design of any information system is justifying and demonstrating the effectiveness the information system (Gonzales and Bagchi, 2011). This research builds on the DOI theory to evaluate the success and significances of integrating clinical data in multiple data sources using a data mart in a healthcare setting. The research further uses the DOI framework to define the kind of information system developed as a result this of research undertaking, and this is an information system that is in the early stages of diffusion, thus allowing timely and practical feedback to be given for further implementation of the information system.

2.4 Data Warehousing and Data Marts

The concept of "data warehousing" arose in mid 1980s with the intention to support huge information analysis and management reporting (Teh Ying Wah and Ong Suan Simn 2009). Hoffer, et al, 2011 argues that the development of data warehousing was a result of the recognition of the fundamental differences between operational systems and informational systems. Operational systems support the day-to-day conduct of the business, and are optimized for fast response time of predefined transactions, with a focus on update transactions. In contrast, informational systems are used to manage and control the business.

Inmon (2005) defines a data warehouse as a subject-oriented, integrated, nonvolatile and time-variant collection of data in support of management's decision. Inmon (2005) explains each of the parts of this definition:

I. Subject-oriented: Data is organized around major subject areas of the company. Each type of company has its own unique set of subjects.

II. Integrated: Data is fed from multiple, disparate sources into the data warehouse. As the data is fed, it is converted, reformatted, re-sequenced and summarized. The result is that data – once it resides in the data warehouse – has a single physical corporate image.

III. Nonvolatile: Data warehouse data is loaded and accessed, but it is not updated. Instead, when data in the data warehouse is loaded, it is loaded in a snapshot, static format. When changes occur, a new snapshot record is written. In doing so, a historical record of data is kept in the data warehouse.

IV. Time-variant: Every unit of data in the data warehouse is accurate as of some moment in time. In some cases, a record is time stamped. In other cases, a record has a date of transaction. But in every case, there is some form of time marking to show the moment in time during which the record is accurate.'

According to Kimball and Ross (2002), a data warehouse is a system that extracts, cleans, conforms, and delivers source data into a dimensional data store and then supports and implements querying and analysis for the purpose of decision making".

The challenge for an organization is to turn its archives of data into a source of knowledge, so that a single integrated/consolidated view of the organization's data is presented to the user. The concept of a data warehouse was deemed the solution to meet the requirements of a system capable of supporting decision-making and receiving data from multiple operational data sources (Connolly and Begg 2005).

Today, data warehouses are not only deployed extensively in banking and finance, consumer goods and retail distribution and demand-based manufacturing, it has also became a hot topic in noncommercial sector, mainly in medical fields, government, military services, education and research community etc.

The data warehouse is significantly different from a conventional operational or transactional database in several aspects. A data warehouse is typically a read-only dedicated database system created by integrating data from multiple databases and other information sources. A data warehouse is separate from the organization's transactional databases (i.e., OLTP databases). It differs from transaction systems in that (Gray and Watson 1998):

I. It covers a much longer time horizon (several years to decades) than do transaction systems.

II. It includes multiple databases that have been processed so that the warehouse's data are subject oriented and defined uniformly (i.e., ‘‘clean prearranged data’’).

III. It contains non-volatile data (i.e., read-only data) which are updated in planned periodic cycles, not frequently.

IV. It is optimized for answering complex queries from direct users (decision makers) and applications.

Organizations often build enterprise-wide data warehouses, where a central data warehouse serves the entire organization, or they create smaller, decentralized warehouses called data marts (Laudon and Laudon, 2012). Bonafati et al (2001) argues that in order to standardize data analysis and enable simplified usage patterns, data warehouses are normally organized as problem-driven, small units, called “data marts”, where each data mart is dedicated to the study of a specific problem. Inmon (2005) defined a data mart as a data structure that is dedicated to serving the analytical needs of one group of people and states that the data mart structure will be fed from the granular data found in the data warehouse. A data mart shares the characteristics of a data warehouse, such as being subject-oriented, integrated, non-volatile, and a time-variant collection of data (Inmon, 2005).

The data mart can be standalone or linked centrally to the corporate data warehouse. As a data warehouse grows larger, the ability to serve the various needs of the organization may be compromised. The popularity of data marts stems from the fact that corporate-wide data warehouses are proving difficult to build and use (Connolly and Begg 2005). Laudon and Laudon, 2012) states that because a data mart mainly focuses on a single subject area or line of business, so it usually can be constructed more rapidly and at lower cost than an enterprise-wide data warehouse.

The pro and cons for using a data warehouse or data mart to integrate data from multiple sources depends upon the business requirements and project scope. Hence this research opted to use the data mart strategy given the scope of project and critical time deadlines to achieve short-term goals.

2.5 Data Integration

Most organizations have different databases for different purposes some for transaction processing in different parts of the enterprise, some for local, tactical, or strategic decision making and some for enterprise-wide coordination and decision making (Hoffer, et al., 2011). To break down the silos of data, organizations need data integration solutions which can integrate data wherever it resides in order to fully support their business requirements. Abello, et al. (2002) defines data integration as the process of combining data residing at different sources and providing the user with a unified view of this data. Cali, et al, 2005 argues that task of a data integration system is to combine the data residing at different sources, and providing the user with a unified view of these data, called global schema. The global schema is therefore the interface by which users issue their queries to the system. The system answers the queries by accessing the appropriate sources, thus freeing the user from the knowledge on where data are, and how data are structured at the sources.

Rashmi, et al. (2014) noted that data integration appears with increasing frequency as the volume and the need to share existing data increases. Building a data integration system provides a uniform query interface to a multitude of data sources, thereby freeing the user from the tedious task of interacting and combining data from individual data sources (McCann, et al., 2003). The process of data source integration has two major components. These include; schema matching which refers to the task of identifying related fields across two or more databases (Rahm and Bernstein, 2001) and data transformation in which data in matching fields must be translated into a common format (Koeller, 2006).

Core to any method of data integration are technologies to capture changed data, so only data that have changed need to be refreshed by the integration methods (Hoffer, et al., 2011). White (2005) mentioned that data integration involves a framework of applications, techniques, technologies and products for providing a unified and consistent view of enterprise business data. He further states that this view can be created via three different techniques (data consolidation, data federation and data propagation). Hoffer, et al, (2011) argues that these three techniques form the building blocks for any data integration approach.

2.5.1 Data Consolidation

This data integration approach captures data from multiple sources and integrates it into a single persistent data store (White, 2006). The data store could be for example, data warehouse that is used for reporting and analysis or content repository containing unstructured information such as documents, images, and web pages. Ballard, et al, (2003) argues that data sources brought together into one place in advance, leads to user queries not being distributed. White, (2006) states that business applications that process the consolidated data store can query, report on, and analyze the data in the store.

Data consolidation is the traditional approach to integrating information (Ballard, et al., 2003). The advantage of data consolidation is that it allows large volumes of data to be transformed (restructured, reconciled, cleansed, and/or aggregated) as it flows from source systems to the target data store White, (2006).

In addition data consolidation creates a second, local copy of the data, pre-processed as required, thus reducing the need for extensive data manipulation and remote access within the user query (Ballard, et al., 2003). The disadvantages of this data consolidation are the computing resources required to support the data consolidation process and the amount of disk space required to support the target data store (White, 2006). Data consolidation is the main approach used by data warehousing applications to build and maintain an operational data store and an enterprise data warehouse (White, 2006). The data consolidation typically uses either extract, transformation or load (ETL) or replication functionality (Ballard, et al., 2003). ETL is the technology used in a data warehousing environment to support data consolidation (White, 2006). The ETL process provides a single, authoritative source for data that support decision making (Hoffer, et al, 2011). The ETL process consists of extraction that is reading data from one or more databases, transformation that is converting the extracted data from its previous form into the form in which it needs to be so that it can be placed into a data warehouse or simply another database, and the load process puts the data into the data warehouse. Hoffer, et al, (2011) states that advantages of the ETL process include the isolation of users from conflicting workloads on source systems, especially updates; it is possible to retain history not just current values; a data designed for specific requirements can be accessed quickly; it works well when the scope of the data needs are anticipated in advance. However the limitation of ETL are that; network, storage, and data maintenance costs can be very high and in addition, performance can degrade when the data warehouse becomes quite large.

2.5.2 Data Federation

This integration approach provides a virtual view of integrated data without actually bringing the data all into one physical, centralized database (Hoffer, et al, 2011). Ballard, et al, (2003) states that, it is a logical integration that typically takes place in real time. White (2006) mentions that when a business application issues a query against this virtual view, a data federation engine retrieves data from the appropriate source data stores, integrates it to match the virtual view and query definition, and sends the results to the requesting business application. Hence, by definition, data federation always pulls data from source systems on an on-demand basis. White (2006) further notes that Enterprise information integration (EII) is an example of a technology that supports a federated approach to data integration. Hoffer, et al, 2011 states that the federation approach has an advantage of providing access to current data. In addition the approach hides the intricacies of other applications and the way data are stored in them from a given query or application. However, the workload can be quite burdensome for large amounts of data or for applications that need frequent data integration activities and write access to data sources may not be possible. Furthermore, White (2006) notes that Data federation, is not well suited for applications where there are significant data quality problems in the source data. Although this integration approach can still be used when the cost of data consolidation outweighs the business benefits it provides (White, 2006).

2.5.3 Data Propagation

This approach duplicates data across databases, usually with near-real-time delay (Hoffer, et al, 2011). White (2006) explains that updates to a source system may be propagated asynchronously or synchronously to the target system. But regardless of the type of synchronization used, propagation guarantees the delivery of the data to the target, which is a key distinguishing feature of data propagation. The major advantage of the data propagation approach to data integration is the near-real-time cascading of data changes throughout the organization (Hoffer, et al, 2011). White (2006) states that Enterprise application integration (EAI) and Enterprise data replication (EDR) are examples of technologies that support data propagation.

In clinical domain, a patient is subjected to repeated scans, blood tests or other medical examinations. Misinterpreted or erroneous data may lead to erroneous decision making, putting the patients' health on risk (Boterenbrood, Krediet and Goossen, 2014). Henceforth, Boterenbrood, Krediet and Goossen, (2014) further suggest that achieving reliable views on data is paramount.

To create multiple views on standardized clinical data, a data integration strategy is required (Boterenbrood, Krediet and Goossen, 2014). Doan, Halevy and Ives (2012), noted that given the variety of possible architectures for data integration, most systems fall somewhere on the spectrum between warehousing and virtual integration. Therefore the choice of selecting a data integration architecture is dependent on whether data will be loaded into a centralized warehouse (or operational data store) (Bobak, 2012) or will it be made available through real-time integration services, directly accessing the data in the source databases (Rotem-Gal-Oz, 2012).

Given that clinical data collected by MENTORS project at the Infectious Disease Institute is used for healthcare research purposes, Bellika, et al. 2007 argue that, the use of data for research implies the use of massive queries, which may result in performance conflicts in the source systems if those queries are executed in the source systems directly. Therefore, for this research, the use of the data consolidation integration approach is preferred given that it is the main approach used by data warehousing application to build and maintain operational data store, data mart or an enterprise data warehouse (White, 2006), which is also the preferred the architecture for providing the central point for accessing integrated information .

2.6 Data Warehousing Architectures

IL-Yeol (2009) defines a data warehousing architecture as an infrastructure by which components of a data warehousing environments are organized. The two primary paradigms for data warehousing architectures are enterprise data warehouse design in a top-down manner and the data mart design in the bottom-up manner.

The top down implementation requires more planning and design work to be completed at the beginning of the project (Ballard, et al., 1998). In addition the top-down approach emphasizes more coordination and an enterprise-wide perspective (Hoffer, Ramesh, and Topi, 2011). But, this approach has many problems such as high costs, difficulty of the analyzing and collecting of all sources, difficulty of collecting all specific needs of all the organizational departments and more development time. In the bottom up approach implementation involves the planning and designing of data marts without waiting for a more global infrastructure to be put in place (Ballard, et al., 1998). Furthermore, Ballard, et al. (1998) mention that by adopting the bottom-up approach, it does not mean that a more global infrastructure will not be developed; it will be built incrementally as initial data mart implementations expand. The bottom-up approach is the more widely accepted for most users, because immediate results from the data marts can be realized and used as justification for expanding to a more global implementation.

According to Ariyachandra and Watson, (2005), the different data warehousing architectures stress the need to start small and deliver short term "wins" but have a long term plan. The data warehouse architecture design approaches can be broadly classified into the enterprise wide data warehouse design and data mart design. The reviewed literature in this section shows that the bottom-up approach is the most used due to its simplicity, in addition to the immediate results which the approach can yield.

2.6.1 Design of Data Marts

Rashmi and Pahwa (2014) mentioned two main approaches for designing data marts. These are the dependent data marts and independent data marts. Whereas IL-Yeol (2009), gives the data mart bus design with conformed dimensions as the other data mart design approach. These data mart design approaches are explored in the following sections.

2.6.1.1 Dependent Data Mart

In this design, a single enterprise data warehouse is created with a set of dimensional data marts that are dependent on the enterprise data warehouse (IL-Yoel, 2009). According to this approach the data marts are treated as the subsets of a data warehouse (Inmon, 2005). These dependent data marts as represented in figure 2.1, extract the necessary data from the enterprise data warehouse. The data warehouse provides a single version of truth for the enterprise, and each data mart addresses the analytic needs of a business unit (IL-Yoel, 2009). The dependent data marts design approach advocates for enterprise data coordination and integration (Kimball and Ross, 2013). The primary weakness of this architecture is that it requires significant up-front costs and time for developing the data warehouse due to its scope and scale.

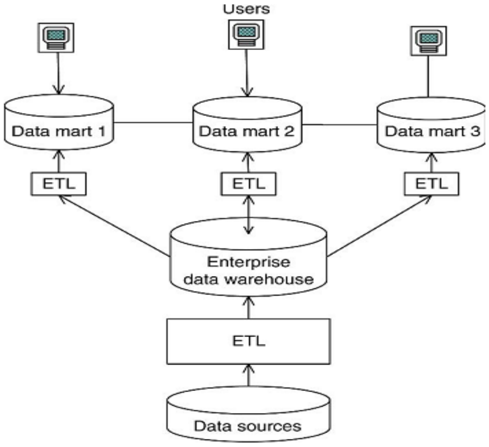


Figure 2.1: Dependent data mart derived from (IL-Yoel, 2009)

2.6.1.2 Independent Data Mart

In this design represented in figure 2.2, multiple data marts are created independently of each other. The independent data marts are developed to meet the needs of the individual organizational units (Ariyachandra and Watson, 2005). Kimball and Ross (2013) states that with the independent data marts approach, analytic data is deployed on a departmental basis without concern to sharing and integrating information across the enterprise. Thus, there is no unified view of enterprise data in this architecture. As the number of data marts grows, maintenance of consistency among data marts is difficult. In the long run, this architecture is likely to produce silos of data marts (IL-Yoel, 2009).

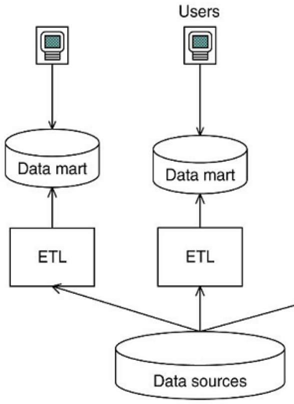


Figure 2.2: Independent data mart (IL-Yoel, 2009).

2.6.1.3 Data Mart Bus with Conformed Dimensions

In this design represented in figure 2.3, instead of creating a single enterprise level data warehouse, multiple dimensional data marts are created that are linked with conformed dimensions and measures to maintain consistency among the data marts (Kimball and Ross, 2013). The data mart bus architecture is designed according to the business requirements of the organization (Ponniah, 2010). At the beginning, data mart architecture is designed with dimensions and measurements and later on, measurement data marts are added to it. The data marts consist of atomic and summarized data and are organized in star schemas

(Ponniah, 2010).

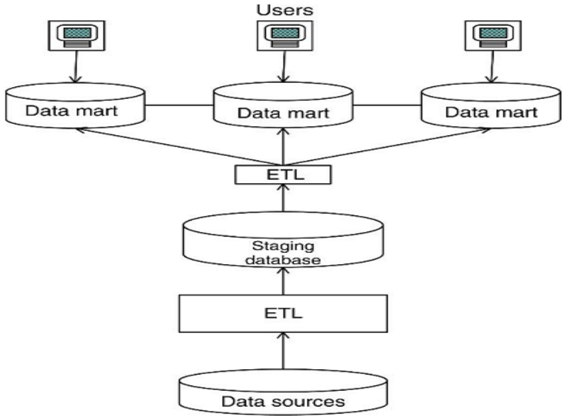


Figure 2.3: Data mart bus with conformed dimensions

2.7 Health Care and Data Marts

The literature covered in this report revealed that despite collecting large volumes of data containing valuable information about patients, procedures, treatment etc, data in health organizations are still stored in operational databases that are not useful for decision makers or executives. In addition many healthcare organizations still have standalone systems that do not communicate with each other (Ado, et al., 2014). Consequently, healthcare providers, face the challenge of managing numerous standalone information systems, including getting the different systems to communicate with one another. HSCC Clinical Data Warehouse (2013) argue that the problem of data integration is of varying significance in every healthcare organization.

Given the data management challenges highlighted in the literature review, health care organizations require data warehousing solutions in order to integrate the valuable patient data fragmented across multiple information systems within the organization. Palmer (2013) states that clinical data warehousing is becoming ever more important, acting as a central hub for information storage, correlation and archiving. Today data warehousing is conventional wisdom in information processing (Inmon, 2007). A clinical data warehouse is a repository where healthcare providers can gain access to health care data gathered in the patient care process (Ado et al, 2014). In addition, Ado et al, 2014 further anticipates that a clinical data warehouse may also provide information to users in areas ranging from research to management. Extracting medical domain information to a data warehouse can facilitate efficient storage, enhances timely analysis and increases the quality of real time decision making processes.

Clinical domain has certain unique data requirements such as high volumes of unstructured data (e.g. digital image files, voice clips, radiology information, etc.) and data confidentiality. Data warehousing models should accommodate these unique needs. According to Pedersen and Jensen (1998) the task of integrating data from several Electronic Health Record (HER) systems is a hard one. This creates the need for a common standard for EHR data.

According to Kerkri et al (2001), the advantages and disadvantages of data warehousing are given below.

Advantages:

1. Ability to allow existing legacy systems to continue in operation without any modification
2. Consolidating inconsistent data from various legacy systems into one coherent set
3. Improving quality of data
4. Allowing users to retrieve necessary data by themselves

Disadvantages:

1. Development cost and time constraints

2.8 Related Work for Data Marts in Health Care

This section reviews related work on data mart development projects drawn from the healthcare and clinical data warehousing literature and in addition highlights the strength and weaknesses of the data mart design approaches used in the selected projects to provide a central repository for clinical data integrated from multiple data sources.

2.8.1 Data Mart based Research in Heart Surgery: Challenges and Benefit

In this case, Arnrich et al, (2004), discusses the typical challenges for the integration of real-time and legacy data stored in multiple unconnected hospital information systems (HIS). The case presents the Heart Institute Lahr a highly specialized hospital, which performs about 2000 open heart operations per year. The hospital for historical reasons choose to operate independent clinical information systems since its inception.

The case highlights the problems and challenges for building a comprehensive research oriented medical database which include; isolated data sources mainly in disconnected HIS, data with partial redundancy and partial consistency, departments prefer to retain autonomy, minimize work flow risk and protect previous investment, obeying privacy protection regulations, legacy data being very valuable. However the case study presents a solution which integrates and consolidates all research relevant data in a data mart without imposing any considerable operational or maintenance liability risk for the existing HIS. The presented data mart architecture proved useful and effective.

The case study revealed that in the past, possibilities to perform retrospective comprehensive studies in the heart center was extremely time consuming and thus limited. Queries to extract and connect data were often rebuilt and modified. Consequently the semantics and definitions of the jointed data changed from one study to the other. With the implementation of the presented data mart, the time and effort consuming process was replaced and thus this lead to reproducible results. Registered user can access the data mart system via a web based information portal in the intranet.

The described data mart solution in the case study demonstrated the benefit that can achieved when clinical information in unconnected hospital information system is consolidated. The case study mentions one of the benefits of the solution as being able to carry out surgical quality assessment through inter-hospital and inter-surgeons comparison of mortality rates after cardiac surgery.

In summary this solution increases performance and accessibility to information when analyzing distributed medical data in a specialized healthcare environment like cardiac surgery whereas its limitation may lie in its suitability to be extended to others disease.

2.8.2 A Health Care Claims Data Mart: Construction and Exploitation

In this case study Scerbo (2009) demonstrated how the CHPDM, University of Maryland overcame the challenges of handling health claims data, through the development of a Data Mart. The case study highlighted some key issues; like any health organization, CHPDM has a number of people looking for different information. They have to sort through large volumes of data in order answer queries from insurers, providers, analysts and other people within the organization. The usual method of getting information was for a user to submit a request to a programmer to extract the data and format it into a report.

The case study revealed that against this background CHPDM decided to simplify and automate this process by developing a data mart to contain this data and give the users access to the data. This would allow analyst to find their answers without having to wait on the backlog of programming requests. In addition the developed data mart enabled users to access data that they could never see before and in the functionality that allowed the users to see the data in ways they had never thought they could see.

In summary the data mart solution enabled user to discover new knowledge regarding healthcare cost and determining profitability of products. Also users were able to access data in a way that satisfied their business needs quickly and efficiently.

2.8.3 Building a Diabetes Data Warehouse to Support Decision Making in Healthcare Industry

In this case study Ado et al (2014), proposed an architecture for healthcare data house for diabetes disease which could be used to monitor diabetes disease, measure cost of infections and to detect prescription errors. In addition the data warehouse would be used by healthcare executive managers, doctors, physicians and other healthcare professionals to support the capture of healthcare processes and analysis of data and offer the potential of altering the practical and delivery of healthcare and medical research.

The case study presents a number of important key issues, one of these is that one of these is that executive managers or physicians should base their decision on information during decision making if an organization is to succeed. In addition given the large volumes of valuable information that healthcare organizations gather, this information is still stored in standalone systems that do not communicate with one another. The case suggests that today, data warehousing is conventional wisdom in data processing and that clinical data warehouse is a place where healthcare providers can gain access to clinical data gathered in the patient care process. Clinical data warehouse can facilitate efficient storage, enhance timely analysis and increase the quality of real time decision making.

The case study used the technology ETL and OLAP in designing the diabetes data warehouse, with the objective of facilitating real time analysis. Finally the case concluded that, building the diabetes data warehouse allowed the health care providers to make informed healthcare decision regarding the treatment of diabetes. This helped the healthcare providers to improve the care they provide to patients.

In summary the main benefit of clinical data warehousing is timely data analysis and increase in the quality of fact based decision making whereas the drawback is the complexity and time required to develop the system.

2.9 Related Work on Existing Approaches to Integration of Clinical Data in Dispersed Systems

This sub-chapter reviews related work on integrating clinical data from multiple systems in order to achieve a consolidated view of clinical data stored in disparate data source for purposes of improved quality and timely analytical reports for use in clinical research and decision making.

2.9.1 Integrating of Biomedical Data Using Federated Databases

The debug IT project (Teodoro, et al., 2009), is a pilot system developed to integrate biomedical data from several healthcare centers across Europe. The system aimed at solving complex problems arising, firstly from the technical and semantic heterogeneity common to biomedical data sources and secondly from the unreliability of the distributed systems. The project used clinical and operational information from existing clinical information systems located in several hospitals across Europe with a view of advancing the healthcare battle through the use of information technology.

Access to these heterogeneous data is gained through a virtualized, fully integrated clinical data repository (CDR). The architecture of the CDR constitute a simple data integration model via the MySQL federated engine to integrate distributed data. The federated data integration approach as stated by White (2005), has a main advantage of providing access to current data and removes the need to consolidated sources into another data store. However the data federation approach is not suited for retrieving and reconciling large amount of data or applications where there are significant data quality problems in the sources data. The federation approach also lacks the possibility to store historical patient data and facilitate a huge knowledge base which could be queried in any desired way by the patient care providers, business and administrative staff. In addition there are performance issues in case the need arises to access multiple data source at runtime.

2.9.2 eHealth Integrator – Clinical Data Integration in Lower Austria

Stolba and Schanner 2007, presented a case study about the "NOMED WAN PatientenIndex" project, in which the Integrated Health Enterprise (IHE) based healthcare network for exchange of patient's documents was developed. IHE is an initiative designed to stimulate the integration of information systems that support modern healthcare institutions. The IHE is organized by clinical and operational domains. For each domain the integration and information sharing preferences are defined. The aim of each IHE domain is to promote the implementation of standard based interoperability solutions in its specific area, to improve information sharing workflow and patient care.

The NOMED WAN PatientenIndex was implemented in lower Austria and it involved the stepwise creation of an electronic health network. The aim of the project was the integration of 27 hospitals, and building of a share directory data about patients' treatments, medical summaries, hospital stays and diagnosis. The consolidation of mostly heterogeneous hospital information enabled physicians to view all the existing examination findings and reconstruct the medical history of their patients.

The IHE based approach enables seamless data exchange beyond clinical and local healthcare boundaries. In addition the approach is capable of improving the quality of care through efficient access to patient's medical history. However the limitation of the IHE approach include; the data stored in the document repositories are not suitable for querying, and no statistical analysis can be run on this data. Medical records, lab tests etc. are still stored as static documents (.pdf files) and can still be helpful to the medical caregivers to reference patient medical history, but still such information is not appropriate for the support of clinical decision making (Stolba, , 2007).

2.9.3 Clinical Data Integration of Distributed Data Sources Using Health Level Seven (HL7) v3-RIM mapping.

Viangteeravat et al, 2011 presented the design and prototype implementation of the H999-L7 v3-RIM mapping for information integration of distributed clinical data sources. The HL7 is a standard development organization which created standards like the Reference Information Model (RIM) which is a standardized abstract representation of HL7 data across all the domain of healthcare.

The created prototype implementation of HL7 v3 RIM mapping of information integration between distributed clinical data sources promotes collaborative healthcare and translational research. The implementation enables the user to retrieve and search information that has been integrated using HL7 v3 RIM technology from disparate healthcare systems. In addition the prototype effectively and efficiently ensures the accuracy of the information and knowledge extractions for systems that have been integrated. A limitation in building the HL7 system is the laborious manual construction of mappings between the HL7 RIM and the local clinical databases (Umer, et al, 2010). Manually creating mappings is extremely tedious and error prone.

2.10 Summary of Literature Review

This chapter discussed the management of information in healthcare, highlighting the challenges of managing healthcare data due to its uniqueness and complexity. The diffusion of innovation theory is discussed and also chosen as the framework to evaluate the success and significances of integrating clinical data in multiple data sources using a data mart. Data warehousing was considered as more practical approach to managing healthcare data compared to the more traditional static approaches. The concept of data warehousing and data marts was also explained, in addition to how it differs from the traditional online transaction processing systems. A theoretical discussion of the data integration and data warehousing was included. For data integration the discussion was centred on the different data integration approaches and an argument was presented why the data consolidation approach to integrating disparate clinical data sources was preferred for this research. The data warehousing discussion focused on the different architectural designs and in particular data mart design approach was the selected approach for this research.

Furthermore an overview of related work on data mart implementation in healthcare brought out some similarities to this research. It can be seen that consolidating clinical data from multiple data sources into a single data repository improves availability of information required to deliver quality healthcare. However there are noticeable differences between this research and the related work on healthcare data mart implementation discussed in the literature, these include the difference in the specific clinical data being integrated and the environment in which the data mart is implemented. This research proposed a data mart system, integrating clinical data made up of Outpatients medical encounters, TB treatment, HIV treatment and Antenatal. In addition the proposed data mart system can used to readily avail clinical information that can used to improve health care service delivery in a low resource settings. The literature also discussed existing approaches to clinical data integration. The limitations of the various data integration approaches used in the studies mentioned in the literature were highlighted. The key limitations included; the in ability to store historical data or trend data, data access performance and availability. This contributed to the reason for selecting the data mart design approach as the most suitable solution to integrate clinical data stored in multiple data sources for this research.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

This chapter describes the steps and procedures followed in order to achieve the aims and objectives of this project. Section 3.2 discusses the research methodology and the sub sections discuss the research design used in the study, sampling, the selected methods used to collect the data and data analysis, section 3.3 discusses the development methodology and subsections discusses in details the different development phases, system analysis, system design, system implementation and finally system testing and validation phases of the research project and chapter conclusion.

3.2 Research Methodology

In general research methodologies can be classified as quantitative and qualitative. Quantitative approaches are concerned with 'what' is contained in the research, while qualitative approaches are focused on an insight into the deeper question of 'why' (Williams and Gunter, 2005). Instead of getting the answer for 'what', this study probes for the answer of 'why' and 'how'. Silverman (1999) argues that there is a common belief that qualitative research can provide a 'deeper' understanding of social and environmental phenomena, rather than quantitative data alone.

Data integration being the focus of this research goes beyond figures. It emphasises the experiences of human beings. Data warehousing technology is a complex data flow system built upon customised needs. The needs of data warehousing cannot be illustrated numerically.

Based on the nature of this research, the qualitative method was chosen as the most appropriate method to use. Qualitative research methodology analyses data which is not expressed numerically only. Instead, it emphasises the use of human experiences. Further the qualitative method can describe a situation or problem easily and using this method the information gathering becomes an easy task as well.

There are a variety of different techniques that can be involved when doing qualitative research, such as action research, ethnography and case study.

3.2.1 Research Design

A case study was the research design approach for this particular project. Case studies are an obvious choice when question of type 'how' or 'why' are raised. According to Sekaran and Bourgie (2010), case studies involve in-depth contextual analyses of similar situations in other organizations, where the nature and definition of the problem happens to be the same as experienced in the current situation. Henceforth the study design used in this project will provide more evidence as what factors might be operating in the current situation and how the problem of managing multiple clinical data sources can be solved within other healthcare organizations facing the same challenges.

3.2.2 Sampling

Sampling is a practice of selecting and inquiring from a fraction of the total population for purposes of making the conclusions about the population as a whole, Oxford (2011).

The reasons why the researcher decided to carry out sampling are:

i. It was time saving because case study was conducted in an identified institution – in this case the MENTORS project in Infectious Disease Institute.

ii. It is a proven practical and realistic means of population representation that provided appropriate response rates instead of attempting to cover a whole set of all projects and the different departments in the Infectious Disease Institute.

iii. It eliminated the need and extra costs that would be incurred over using a large number of interviewers and research assistants who are expensive and difficult to control.

3.2.3 Sampling Technique

Purposive sampling was the chosen sampling technique. It is a non-probabilistic technique of sampling that has researchers selecting a sample based on their own judgment towards a specific purpose. The MENTORS project was selected to represent the Infectious Disease Institute as a whole.

The main justification behind the selection of the above mentioned technique was to ensure the successful development of a data mart system that would successfully suit the requirements of one institution considering that Infectious Disease Institute runs many different clinical projects, which function with similar standard operating procedures. As such developing a data mart system to integrate multiple clinical data sources for the MENTORS project would fits well in any clinical based project within the Infectious Disease Institute.

3.2.4 Sample Area

The study was conducted focusing on key project staff of the MENTORS project of IDI, particularly those staff involved in the decision making or in supporting the decision making and research processes in the healthcare practice of the MENTORS project. The justification for selecting this area included;

i. The MENTORS project where the researcher is an employee was experiencing the challenge of integrating clinical data gathered from its multiple sites for use in decision making and research processes. This provided the researcher the opportunity to study the challenges and drawbacks of working with multiple clinical data sources in a healthcare setting, including the merits that can be gained by creating a single repository for the multiple clinical data sources.

ii. The project also allowed the researcher to conduct interviews with the various project staff and as a result, personnel reliability gained the researcher a chance to explore more in-depth information concerning the working of the entire institution, its processes and procedures, in addition to end user requirements and testing.

The MENTORS project staff who were interviewed included; unit managers, clinicians, laboratory technologists and data staff. The sample of the project stakeholders selected for interviews was based on a cross-section of roles the stakeholders had with the various clinical data sources which were to be integrated as a result of this project. In addition the MENTORS project operating from 10 health centres across the country provided the different clinical data sources to work with and these were stored in database management platforms from different vendors (Ms Access and MySQL server).

3.2.5 Sample Size

The study was conducted with sample size of 15 participants. The small sample size of the research was as a result of the availability of the participants and the time limitation of the project. As there are only a few staff members involved in the various processes of implementing the project activities at the MENTORS supported sites, the participants selected represented a good cross section of relevant staff related to the MENTORS project decision making processes and research undertakings. These people provided accurate, credible and consistent information about the various processes involved in the implementation of the MENTORS project activities in the different health facilities.

3.2.6 Data Collection Techniques and Tools

This section describes those techniques and tools that were used in the data collection phase of this project. It involved gathering information from various informants (respondents) as already defined in the sample size. The techniques that were employed to generate qualitative data in this study included; interviews, document review and observation. These are further discussed below:

Interviews

Interviews were carried out with the key personnel of the MENTORS project under IDI, directly involved in decision-making or in supporting the decision-making processes to gain both high-level and low-level perspectives on issues regarding information management within the MENTORS project.

The objective of the interview was to obtain information from organization users, on the status of the clinical data systems, current data management processes, on the issues and challenges in the current data management process and to identify the data mart prototype development requirements. Interview process helped to identify the issues in the utilization of clinical data for decision making and research, including the identification of the main technical requirements for the development of data mart prototype.

The interview technique was chosen as opposed to administering questionnaires, because the targeted respondents had tight schedule related to implementing field based project activities.

The interview questions were composed of both closed and open ended questions. Pre-designed interview guide (Appendix B) was used to gather facts, opinions and speculations. The responses were noted down on paper, analysed, processed and used during the design and implementation of the project. Part of the data collected during interviews includes:

i. Which kind of data sources do you use to assist decision-making and research?

ii. How do you collect or access data from the mentioned data sources to support decision making and research?

iii. The kind of reports used for decision making and research, how these reports are obtained?

iv. What are the main information related problems you have identified in the Decision-making process supporting clinical service management in your area?

v. Do these various data sources store sufficient data fields for your decision-making processes?

vi. What difficulties are experienced when generating reports?

Document Review

Existing data collection instruments and documents were reviewed and these included; paper reports, policy manuals, current system documentation, user training reports, organization charts, case report forms and procedure manuals. The document reviewed provided a broad coverage and helped collect all the necessary information and variables required for the new system. The reviewed documents also provided an insight into the problems within the existing system and the direction of the organization. Additionally, review of existing literature relevant to this study was also done. Areas of concentration included; health information management, clinical data integration, data warehousing and data marts, data warehousing development techniques, design methodologies and application of data warehousing in healthcare. This enabled the researcher to ensure that the most recent and relevant information was used in the preparation of this study.

Observation

Further information related to the current system was gathered through direct observation of the different data processing activities. The process of how users generated information from the various system to make reports for both internal and external consumption was observed. This provided an insight into the problems faced by users in their data processing activities and enabled the researcher to suggest solution to these problems. Observation also provided a good way to check the validity of the information gathered from other sources such as interviews.

3.2.7 Data Analysis

The key data collected from the interview process which was mainly qualitative, was analyzed according to four sections. Firstly current data sources and decision making process, and secondly the issues related with current decision making process in the different units of the MENTORS project of IDI, thirdly analysis was done to gather information on the data storage and analysis process and lastly results were also analyzed to gather information for the technical details of the data mart prototype development. The results from the analysis process gave a better understanding of the current situation, that is, how the current system worked to visualize known problems. Business events, rules and processes were investigated as an input to the specification of the new proposed system. The problems associated with the current environment in terms of data integration, availability of information and reports required for the support of decision making and research process were obtained. Additionally existing process/information flow chart were established. Logical system specifications and requirement specifications were also identified and derived.

3.3 Development Methodology

Due to the fact that the project is based on the business requirements, the development methodology was based on three major phases which are analysis, design and implementation.

3.3.1 System Analysis

This phase was used to analyse the requirements collected and use the analysis to build models that were input for the design phase and these included, Use case diagrams, data flow diagrams, functional and non-functional requirements.

3.3.2 System Design

The main focus of this phase was to translate the systems requirements into a set of specifications through deriving logical and physical data models for the data mart. The specifications were then used to generate other component such as data mart extractors and transformation, data integration tools and so on.

Facts and dimensions tables for data mart were designed using dimensional modelling techniques. A conceptual star schema design for the data mart was developed on the idea of multidimensional model (constellation) in which data marts are composed of several facts and dimensions (Mussa, et al, 2014). Each dimension is shared between facts and it can be associated with one or more hierarchies thus facilitating comparison between several measures/facts (Teste, 2010). The bottom-up approach was adopted to model the resulting tables into a star schema for the data mart implementation.

The design of the data mart design was done by first specifying the measure. The measure are the foundation and feedback information that the decision makers require. The requirements were reconciled with what is available in the source system (OLTP). For the purpose of this project, the star schema was used for the data mart design. The star schema is a relational database schema used to hold measures and dimensions in a data mart. The measures are stored in a fact table and the dimensions are stored in dimension tables. For each data mart, there is only one measure surrounded by the dimension tables, hence the name star schema.

The centre of the star is formed by the fact table. The fact table has a column or the measure and the column for each dimension containing the foreign key for a member of that dimensions. The key for this table is formed by concatenate all of the foreign key fields. The primary key for the fact table is usually referred to as composite key. It contain the measures, hence the name “Fact”.

The dimensions are stored in dimension tables. The dimension table has a column for the unique identifier of a member of the dimension, usually an integer of a short character value. It has another column for a description. In this project the naming convention which was followed to name the dimension tables was based on the information they contained and prefixed with “Dim”.

3.3.3 System Implementation

In this phase of the project, the actual implementation of the analysis and design was carried out. This phase involved the design of the data mart (facts and dimension tables), the ETL (Extract, Transform and Load) and the front end application for this project. The project schedule that was followed during system implementation is outlined in table 3.1 (appendix A).

A collection of software were used to build the data mart system and these included the following:

I. Microsoft SQL Server 2012
II. Microsoft SQL Server Data Tools
III. Microsoft Visual Studio 2010

The various steps taken in the development of the data mart system included;

i) Created physical database and setup dimensional and fact tables to support the Data Mart.

ii) Extracted, Transformed and loaded clinical data from the various clinical data sources into the data mart tables.

iii) Designed queries for most of the reports as obtained from user expectations and made them available.

iv) Developed front end application to data mart system.

Particularly, MS SQL server 2012 was used at the back-end to develop the data mart objects like tables, stored procedures and above all, Microsoft SQL server 2012 has a comprehensive data warehouse platform with inbuilt Extraction, Transformation & Load (ETL), and can provide business query services.

3.3.4 System Testing and Validation

The system developed was tested to ensure that the system functional requirements were correctly understood as specified by the users and that the system met the required objectives set by the researcher. The selected users for the testing the system were those involved in interviews process regarding the existing system/processes with the researcher. All these users received basic training on the developed system to enable them contribute effectively to the testing and validation process of the new system. Given the nature of sensitivity and confidentiality of clinical data, de-identified data was used to test the system.

Validation of the system was done to ensure that the designed system delivered the results as expected by the users in the MENTORS project of IDI. The users involved in the validation step were asked to give a report on problems and omissions in the designed system. The validation criteria involved testing the system to confirm its ability to extract, transform and load clinical data from multiple clinical data sources into the designed system, and the accuracy of reports generated from the system. Revisions to the system were done basing on the feedback from the users. This process was iterative until a working system that could provide ease of access to consolidated clinical data in MENTORS project at IDI was obtained. A User Acceptance Test document and a Likert scale attitude statement were used in the process of testing and evaluation.

3.4 Conclusion

This chapter described the research and the various methodologies used to gather comprehensive system requirement specifications. It also gave an overview of the next phases, the deliverables and a detailed schedule for system development. The next chapter is the analysis phase where the requirements were determined and analyzed to determine the appropriate business processes for the data mart system.

CHAPTER FOUR

SYSTEM ANALYSIS

4.1 Introduction

This chapter describes the requirements specification for the data mart system. Section 4.2 gives the description of the problem, section 4.3 describes the requirement gathering process, section 4.4 discusses the results from the data analysis, section 4.5 discusses the current organisation structure including illustrations, section 4.6, 4.7, 4.8 and 4.9 describe the current business processes and challenges, the new process and benefits of the new process, section 4.10 details the proposed system and finally section 4.11, 4.12 and 4.13 outlines the proposed system requirement both functional and non-functional, the user requirement are also outlined and system requirements.

4.2 Problem Description

Information stored in different sources hinders the availability of information required to support the decision making process aimed at improving healthcare service delivery. The Infectious Disease Institute (IDI) undertakes several clinical studies or projects with the aim of strengthening health systems in Africa. The various studies collect massive volumes of clinical data from various study sites to support the reporting functionality of the studies and also back research undertakings. In particular the MENTORS project which is the main focus of this research runs under the training department of IDI. The project operates in 10 health centre IVs across the country with the main objective of improving care and treatment of TB and HIV/AIDS patients in low resource settings.

Typically the MENTORS project of IDI requires clinical data for planning, decision making and dissemination of research findings contributing to the country’s healthcare policy. However there is limitation in the accessibility of clinical data gathered by the project from the different implementation sites due to the existing setup, whereby clinical data is distributed in multiple operational clinical data sources and in different formats. Consolidating the multiple clinical data sources for reporting is almost difficult to achieve as it involves extensive manual processes and is time consuming. In addition validation and cleaning of data errors resulting from data entry is also a difficult and lengthy process.

Given the prevailing situation, there is a need to build a data integration solution that would consolidate clinical data from disparate sources to ease the data analysis process and reduce the complexity of delivering information for research purposes and decision making.

4.3 Requirement Gathering

End-User requirements gathering was carried out using interviews as the main tool. Interviews were conducted with key MENTORS project personnel and were geared towards understanding different stakeholders’ opinions about the current data management system and what they would like to see and have as improvements. An interview guide (Appendix B) was administered to all the potential end users of the data mart System. In addition, document review was carried out and observation was used as a mechanism to establish the process flow and shadowing users, to see what systems they use and how they are used within the overall data management process in the MENTORS project.

4.4 Data Analysis Results

The data collected from the interview process was categorized basing on the responses obtained from the end users interviewed and the results of the analysis are explained according to areas; Current data source and decision-making process, current decision making issues in the different units of the MENTORS project of IDI and the data storage and analysis processes. The analysis is detailed as below:

Data Sources and Decision-Making Process

Data collected on questions 3-5 of the interview guide was analysed to provide information on the various data sources and decision making process in the MENTORS project.

According to the responses from the end users of the multiple clinical databases of the MENTORS project, these databases are gathered from the different supported sites of the project on a monthly basis. End users use data from outside the clinical data repositories for the decision making process. Clinical end users, mentioned that they would like to have a holistic view of patient treatment information. However combining patient data stored in multiple databases is largely a manual process and this creates inefficiencies in the clinical decision making process. According to the results of the interview process, some aspects of the current decision-making processes involving multiple sources of data are as follows,

i. If clinicians a requires TB/HIV co-infected patient information, this entails gathering information from different databases by contacting the data staff to extract and provide the data.

ii. If the Monitoring and Evaluation manager needs summary information on key performance indicators, in order to prepare monthly and quarterly report dissemination of the MENTORS project performance across the supported sites, they have to contact the data manager to generate the summaries for the specific indicators.

iii. All unit managers need to contact the data staff in order to collect specific information for the decision making process.

When considering the end user decision making process, it can be seen that currently there is a high degree of repetitive manual processes related to data access and acquisition. The clinician or section managers collect the data separately by contacting data staff and individually integrate and assemble the data for analysis through laborious and time consuming linking processes.

Current Decision Making Issues

Data collected on questions 6-8 of the interview guide was analysed to provide information on the current decision making issues in the MENTORS project.

Participants in the interview process mentioned many issues related to their current decision-making processes. The participants were not satisfied with the support provided by the current information systems for decision-making.

According to further findings from the interviews with the end users from the different sections, systems need to be integrated and should provide easier access to data. Also, clinician and laboratory technologists from the clinical unit of project, pointed out the difficulty in accessing data held in the multiple clinical data sources for use in clinical care quality improvement sessions conducted at the various MENTORS project supported sites. Another respondent from the clinical section mentioned that there is a need of comprehensive data availability at all stages of the decision making and the current systems did not support this. Furthermore, detailed response from the clinical unit manager stated that “there is lack of support available for the current decision-making process from the current information systems and there was need for a centralized data management (process) to improve decision-making and also provide timely data for clinical research undertakings”.

More end users interviewed on the questions related to the current decision-making issues, mentioned integration of data from multiple clinical data sources as the main problem for current decision-making. The end users such as clinicians and unit managers frequently know which information they require, and from where the information is available, but they do not have effective methods to integrate the data. As mentioned before the clinicians or unit managers contact the data staff to collect required data separately. This is a time consuming and often complex process for both parties. For instance, from the clinicians point of view, they have to analyze the data collected from separate clinical intervention area, from the data managers point of view, it takes some time to obtain and integrate information for complex ad hoc queries.

Limited accessibility to data and lack of data availability was another problem pointed out by the end users. As mentioned before, there is limited access to databases or some end users may have difficulty obtaining authority to access data repositories. According to the interviews held with end users, the main reasons identified are, security and confidentiality issues of the MENTORS project information related policies. End users also pointed out a lack of efficient reporting tools and lack of time and resources to undertake analysis as two other problems. According to the data collected from the interview process, analysis tools employed by the different sections are SPSS, STATA and Microsoft Excel. However, in further interviews, end users indicated that there is a need to implement better reporting and data analysis tools. Respondents in the interview process also mentioned some data quality issues identified in the collected data as; data completeness, data accuracy and consistency in the gathered clinical data. Data staff mentioned that cleaning the multiple clinical databases from the various project sites was a challenge and hence it was very difficult to completely eliminate the data quality issues, given the manual approach which was employed to clean the databases. This in a way affected the quality of decision making and research findings disseminated to stakeholders.

4.5 Current Organisation Structure

The MENTORS project under the training department of the Infectious Disease Institute Ltd is headed by the Project Manager who reports to the training department head who is also the project principle investigator. The project is functionally structured into sections headed by section managers and supervisors respectively. The project is fully depended on the multiple clinical OLTP systems gathered from its 10 implementation sites, for all the information required for decision making. The project had about 25 users, and the different clinical OLTP system store over 1,000,000 rows of data related to HIV/AIDS, PMTCT, Tuberculosis, Laboratory diagnosis and Outpatients.

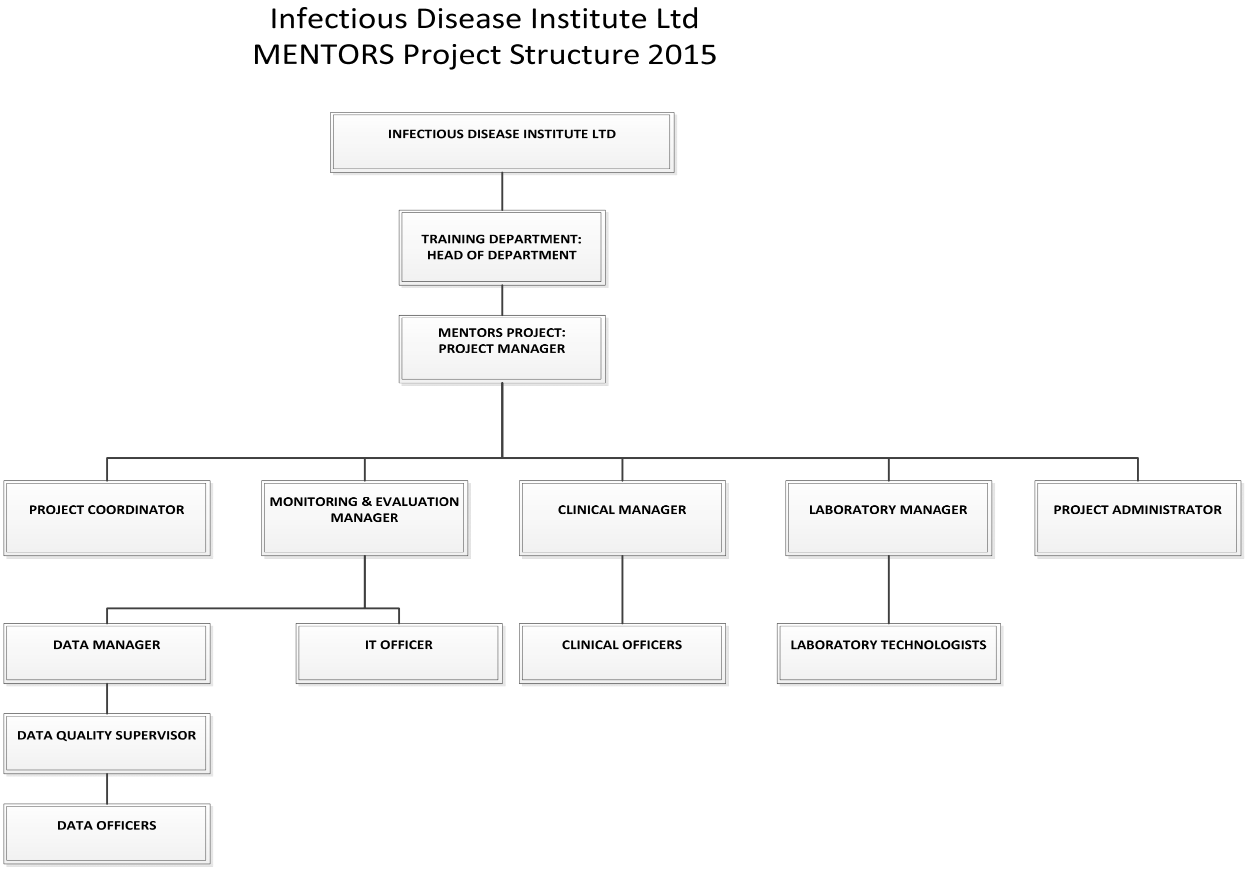

Infectious Disease Institute Ltd
MENTORS Project Structure 2015
INFECTIOUS DISEASE INSTITUTE LTD
TRAINING DEPARTMENT: HEAD OF DEPARTMENT
MENTORS PROJECT: PROJECT MANAGER
PROJECT COORDINATOR
MONITORING & EVALUATION MANAGER
CLINICAL MANAGER
LABORATORY MANAGER
PROJECT ADMINISTRATOR
DATA MANAGER
IT OFFICER
CLINICAL OFFICERS
LABORATORY TECHNOLOGISTS
DATA QUALITY SUPERVISOR
DATA OFFICERS

Figure 4.4: MENTORS Project Structure

As clearly seen from figure 4.4 above, the MENTORS project is directly under the supervision of the training department head of the Infectious Disease Institute Ltd. The project structure is hierarchical and flexible whereby reporting can be done at any level. For example, there existed many scenarios where the Project manager or any other senior person in the structure directly asks for information from any other data personnel or manager without following the chain of command.

Current Roles in the IDI – MENTORS Project

Table 4.2 Current roles in the IDI – MENTORS Project

Title	Roles
MENTOR Principle Investigator (PI)	• Overall scientific and managerial oversight of the project • Interpretation, reporting and publication of the research data • Liaison with stakeholders including the Ministry of Health
Project Manager	• Technical lead and works in collaboration with the PI • Provide day to day support and oversight to ensure quality implementation of the project. • Manage project staff and ensure timely delivery of the intervention • Prepare quarterly and annual reports.
Project Coordinator	• Oversee all project related activities and ensure they are implemented uniformly across all the sites. • Ensure high quality implementation of project activities • Provides technical oversight and support to all the team members of implementing sites. • Prepare project activity reports
Biostatistician	• Directly handles data analysis and also offers technical support to help determine sample size estimations for power and precision of the study
Clinical Manager	• Oversees the all project clinical activities • Ensures high quality implementation of all project clinical activities across all the sites. • Directly supervises all project clinical officers • Prepares clinical site performance reports
Laboratory Manager	• Ensure implementation of standard, practices and procedures across all the sites. • Interpret test results and authorize written reports • Develop and implement quality control measures in the labs across the project sites. • Develop and coordinate laboratory training programs • Direct supervision of laboratory technologists
Monitoring and Evaluation Manager	• Responsible for project progress and ensures that effective M&E systems are established • Monitors project implementation and ensures that timely decisions on corrective actions are made.

	• Identifies Key performance questions and parameters for monitoring performance and comparing targets. • Prepares project monthly, quarterly and annual reports.
Clinical officers	• Deliver on-site clinical support and training to staff at the various project sites.
Lab Technologists	• Deliver on-site laboratory support and training to laboratory staff at the various sites.
Data Manager	• Design data collection tools • Initiate the quality improvements • Set-up programs for proper storage of project data • Track data quality across all sites. • Conduct routine data analysis • Prepare monthly, quarterly and annual reports
IT Officer	• Ensure that all IT equipment and infrastructure are maintained in sound working order.
Data Quality Supervisor	• Ensure data quality across all the sites • Symmetrically track data quality issues across all the sites • Conduct data validation and cleaning at the various sites • Participate in routine data analysis and reporting of project data
Data Officers	• Ensure all data gathered at sites is entered into project databases • Ensure quality of collected data at the sites • Perform data analysis and reporting of project data at respective sites • Ensure proper storage of all soft and hard copies of gathered project data.

4.6 Current Business Processes

4.6.1 On-site support and training

Currently clinical and laboratory technologists are involved in implementing both clinical and laboratory on-site support and training across the 10 sites of the MENTORS project of IDI. The clinical officers deliver clinical training sessions in diagnosis, treatment, patient care and management, while the laboratory technologists equip and improve the capacity of site laboratory staff to conduct various diagnostic tests related to malaria, TB and HIV. In addition the on-site support includes a session on data driven continuous quality improvement which is provided for all healthcare centre staff. Data from the TB clinical and laboratory registers, HIV, maternity and outpatients medical encounter forms is used to identify bottle necks to patient care and laboratory management, in order to develop quality improvement work plan solutions. However this data is stored in different OLTP systems which include MS access and MySQL databases, hence building a compressive view of patient information is difficult. For example building a holistic view of a patient who is on both HIV and TB treatment is done manually since TB and HIV data is stored in different information's systems.

4.6.2 Data Management

The transactions in this process include; data collection, data entry and processing, data quality, data storage and archiving, reporting and analysis.

Data Collection

Collection of data at the various sites of the MENTORS project is done by healthy facility staff who complete records for every patient who visits the relevant clinics of the health facility. The medical forms used for data collection include; the outpatients medical encounter forms, HIV care/ART cards, TB treatment and laboratory registers and Antenatal registers.

Data Entry and Processing

A project data officer is stationed at each project site to support with electronic data entry on a daily basis at the facility. The collected secondary data which mainly includes; outpatients medical encounters, HIV care, TB treatment and laboratory , and Antenatal is entered into electronic data entry systems built in Ms Access and MySQL server relational databases.

Data Quality

The various secondary data sources which include registers and medical record forms are checked for accuracy and completeness. Data discrepancies are identified manually at any point during data processing or checked by the electronic data systems automatically at entry or after entry. The project data officer frequently carryout manual review of data and medical forms. The data quality supervisor conduct monthly data quality assurance exercises of electronically captured data usually in the form of data audits against the registers and medical record forms. The Data Manager checks for data quality in the merged dataset using systems external to the electronic data management system. All identified data quality issues and resolutions are reviewed under lead of the data manager and the information is fed back to the source.

Data Storage and Archiving

Presently softcopies of clinical data gathered by the MENTORS project are maintained in relational databases. Outpatients medical encounters, TB and Antenatal are stored in an Ms Access database and HIV care data is stored in Open Medical Record system built on MySQL server database platform. Hard copy data is maintained at the project sites by the project data officers where all forms are filed and shelved on a daily basis. To maintain confidentiality and protection of patient data, data access is regulated to only site and project data staff and clinical personnel. All soft copies of data are backed up on a daily basis on optical drives and kept under key and lock mode. A soft copy of gathered data is transmitted to the project data manager for integration and analysis on monthly basis. Password protected soft copies of gathered data are transmitted monthly to the project data manager and data quality supervisor through black berry 3G data network for purposes of data integration and analysis across the sites.

Data Analysis and Reporting

Analysis of collected clinical data is done to monitor and evaluate the performance of various health facilities. A mixture of both manual and electronic methods such as excel, STATA, SQL, and Open Medical Record reports are used for analysis. Analyzed data provides information on key performance indicators which measure quality of care related to HIV/AIDS, Malaria, TB and Antenatal, as well as site's performance in supporting functions like laboratory services. Reports prepared are used to influence policy and improve the care and treatment of patients within the project supported health facilities and country at large. The stakeholders who need to make use of the information from resulting analysis process include all MENTORS project staff, IDI corporate monitoring and evaluation team, external stakeholders include Centre for Disease Control (CDC), Ministry of Health, District Health Officials and health facility in-charges.

4.7 Challenges within the Current Process

The clinical data sources used in the current process are not integrated into one data repository for fast access to information which the MENTORS project rely on to make decisions. The traditional approach being used, of building together reports that pull in data from many various source systems is laborious and time consuming. Analysing clinical information from a holistic view point is difficult since information is distributed in many information systems. The decision making process is delayed because users are unable to carryout data analysis without relying on the domain experts to construct queries. Data in the various source data sources is captured in many different ways which makes it impossible for users to easily and accurately use the data without a significant cleaning effort.

4.8 The New Process

The new process aims at improving the current process. The introduction of the proposed system will become a model that will be utilized to build a fully operational data warehouse for the various clinical based projects within the infectious Disease Institute in a sequential phases approach. The developed system will be capable of consolidating the multiple clinical data sources gathered by the MENTORS project towards easing the data analysis process, improving data quality and reducing the complexity of delivering information for research and decision making within the MENTORS project. The key processes in the newly developed system include;

i) Data Acquisition: This process involves integrating clinical data from the different MENTORS projects sites, stored in multiple clinical sources into the data mart. The architectural components which makeup this process include; source data and data staging. Source data is made up of Outpatients medical encounters, TB and Antenatal data stored in an Ms Access database and HIV care data stored in Open Medical Record system built on MySQL server database platform. In this process data comes from two different source data categories in different formats, so its need to convert those data into data mart suitable formats. This is done by the Data Staging component. The function and services for the data acquisition include; Data Extraction, Data Transformation and Data Loading, which involve extracting key information from the different clinical data source, transforming data into clean, consistent, and usable data, and loading it into the data mart.

ii) Data Storage: In this process, clinical data from the various data sources is loaded into the data mart. The data mart contains the data structure in highly normalize form for fast and efficient processing. The data storage in data mart is kept separate for quick retrieval of individual pieces of information. The Data mart is a read-only data repository.

iii) Information Delivery: In this process the user collect information from data mart. To collect the information from data mart, information delivery components is used to make it easy to access. Different levels of users are able to collect information from the data mart. There are different information delivery methods for different user. Ad hoc reports are predefined reports primarily meant for non-technical users. Provision for complex queries, multidimensional analysis and statistical analysis cater to the needs of the business analysts and power users.

4.9 Benefits of the Proposed Process

The proposed system will provide the MENTORS project stakeholders with a single data repository for fast, easy access and analysis of clinical information used in the decision making process, and research aimed at improving the care and treatment of patients within the various health facilities of the project and the country at large. Given automation of the integration process, of the clinical data sources, data staff will spend less time hunting and gathering data, but instead spend more time understanding and analysing data in order to fulfil the data needs of the decision makers in a timely manner. Clinical staff will be able to have a comprehensive view of patient information. Non-technical users will be able to carry out analysis and build reports easily without relying on the technical users to construct queries. A user will be able to create detailed queries very easily, because all the clinical data elements are linked. The system will enable upper and middle managers who are non-technical in data warehousing, to view reports related to key performance indicators in a user friendly manner. The data staff will spend less time on cleaning data due to the systematic processing of errors within the new system. The data staff can use error identified during data integration to improve the quality of data in the source systems. The improved quality of clinical data will enable clinicians, researchers, healthcare managers make better quality and data driven decisions.

4.10 Proposed System

The new process will be supported by the developed Data Mart system. The system will be used by all stakeholders involved in the decision making process of the MENTORS project at the various level of management within the project. These include; Top level managers, Middle level managers and operation staff. The diagram figure 4.5 below highlights the system functionalities and associated stake holders.

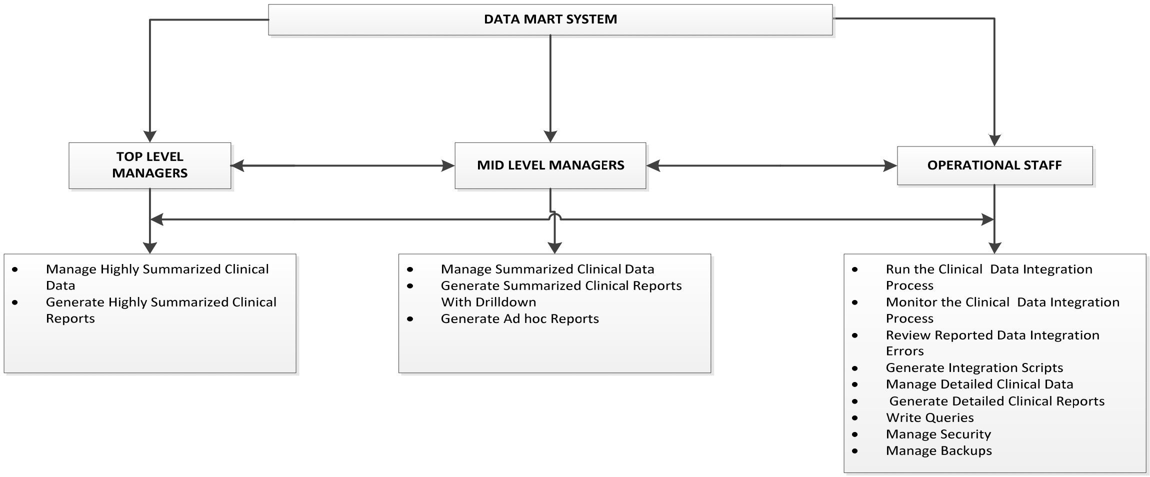

DATA MART SYSTEM
TOP LEVEL MANAGERS
MID LEVEL MANAGERS
OPERATIONAL STAFF
Manage Highly Summarized Clinical Data
Generate Highly Summarized Clinical Reports
Manage Summarized Clinical Data
Generate Summarized Clinical Reports With Drilldown
Generate Ad hoc Reports
Run the Clinical Data Integration Process
Monitor the Clinical Data Integration Process
Review Reported Data Integration Errors
Generate Integration Scripts
Manage Detailed Clinical Data
Generate Detailed Clinical Reports
Write Queries
Manage Security
Manage Backups

Figure 4.5: System Functionalities and Stakeholders

Use Case diagrams

The Use-Case diagrams figure 4.6, illustrates system user / actors and respective actions on the Data Mart System.

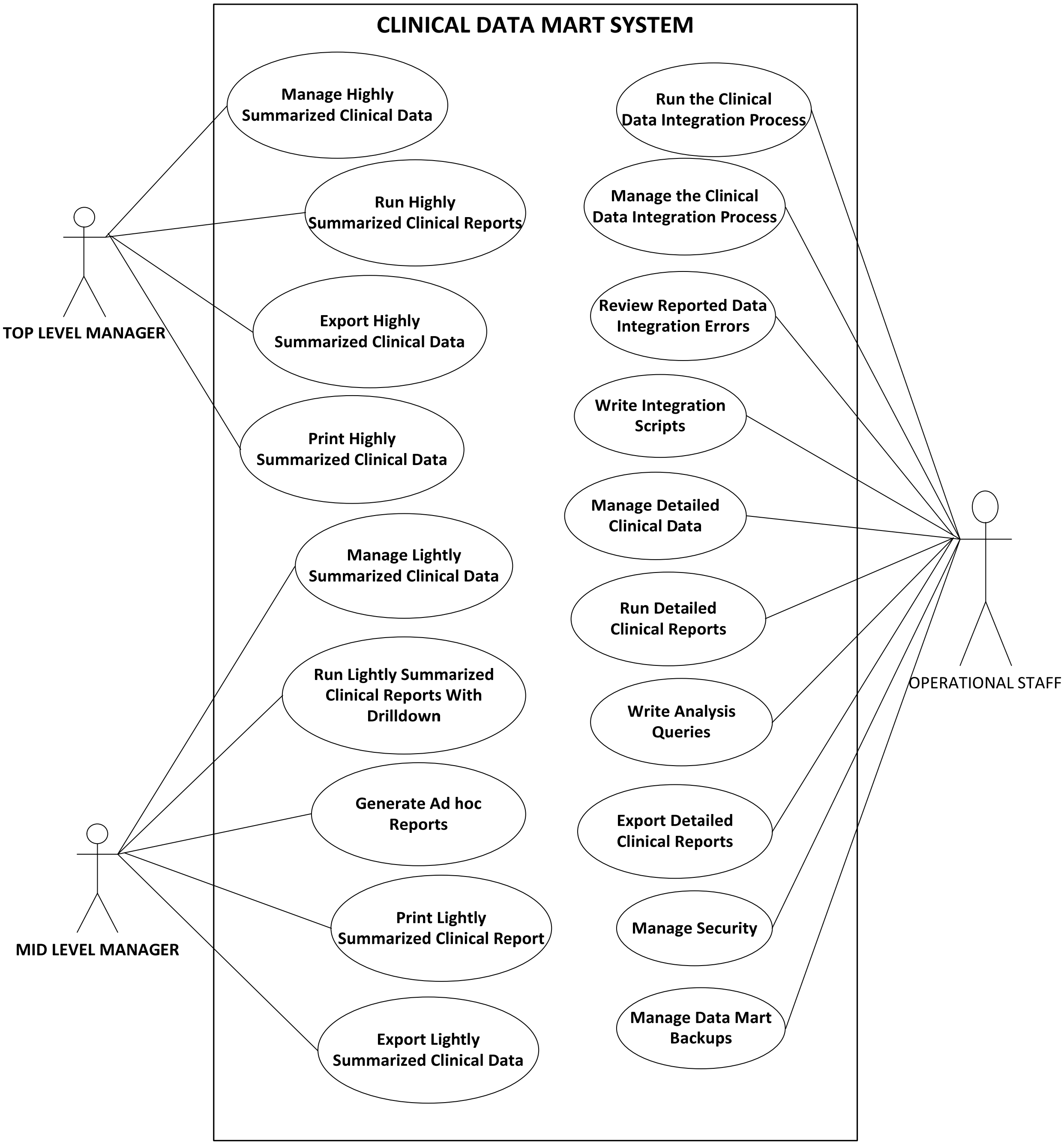

CLINICAL DATA MART SYSTEM
Manage Highly Summarized Clinical Data
Run Highly Summarized Clinical Reports
Export Highly Summarized Clinical Data
Print Highly Summarized Clinical Data
Manage Lightly Summarized Clinical Data
Run Lightly Summarized Clinical Reports With Drilldown
Generate Ad hoc Reports
Print Lightly Summarized Clinical Report
Export Lightly Summarized Clinical Data
Run the Clinical Data Integration Process
Manage the Clinical Data Integration Process
Review Reported Data Integration Errors
Write Integration Scripts
Manage Detailed Clinical Data
Run Detailed Clinical Reports
Write Analysis Queries
Export Detailed Clinical Reports
Manage Security
Manage Data Mart Backups
TOP LEVEL MANAGER
MID LEVEL MANAGER
OPERATIONAL STAFF

Figure 4.6: use case diagram for the Clinical Data Mart System

Process Flow of Clinical Data Integration

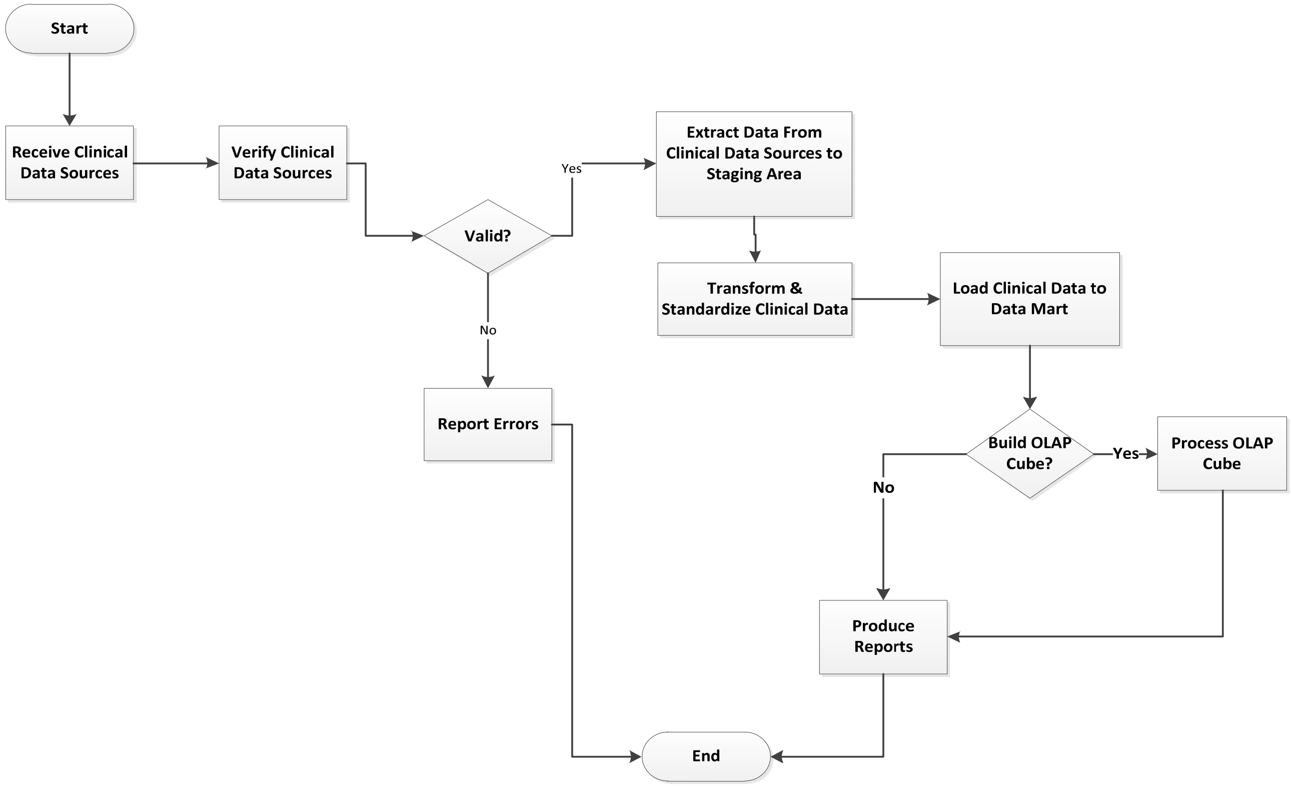

Start
Receive Clinical Data Sources
Verify Clinical Data Sources
Valid?
Yes
No
Report Errors
Extract Data From Clinical Data Sources to Staging Area
Transform & Standardize Clinical Data
Load Clinical Data to Data Mart
Build OLAP Cube?
Yes
Process OLAP Cube
No
Produce Reports
End

Figure 4.7: Process flow diagram for Clinical Data Integration

4.11 Proposed System Requirements

The system requirements were gathered as already mentioned in preceding sections. These requirements were further categorized as functional and non-functional requirements.

4.11.1 Functional Requirements

These requirements defines functions of the data mart system and its component that will work together to accomplish system objectives. A tabular representation of the functional requirements resulting from the interview process is as below:

Functionality	Description	Required by
Easy Generation of clinical reports by the users in the MENTORS project themselves and not relying on technical users	The reports included; Outpatient Encounter Summaries, Disease Incidences, Admission rates, HIV antenatal attendance summaries, TB treatment, HIV treatment encounter summaries, Laboratory test Summaries	Project Manager, Section Leaders, clinical staff and data team.
Data consolidation	Integrating the clinical data from the gathered multiple clinical data sources into a central repository with ease	Data/IT team.
Data Security	Not every staff of the MENTRORS project should have access to clinical data. System should only allow authorised users to access the data in the system	Project Manager, Section Leaders, clinical staff and data team.
Fast Query processing	Fast retrieval of data for analysis using other analytical software's	Section Leaders, clinical staff, data team.
User Friendly system	System should be easy to use	Project Manager, Section Leaders, clinical staff and data team.

Table 4.3: Functional Requirements.

4.11.2 Non-Functional Requirements

The major non-functional requirements include:-

i) System performance: the system should be able to handle at least 20-40 concurrent end-users.

ii) System accessibility: The system users should be able have easy access to the data mart in a more user friendly manner without having to be dependent technical users.

iii) Information security: Only users with the required privileges should be able to access the information in the designed data mart.

iv) Software operability: The initial system should be able to make use of the software environment within MENTORS project of IDI, and therefore should be able to run on the windows operating system.

4.12 User Requirements

Below is a summary of the basic requirements of the system as described by the users:-

i) Use the Data Mart to generate clinical reports.

ii) The system needs to enable centralization of data and information retrieval from various clinical data sources

iii) The system needs to have a friendly user interface.

iv) The System should not allow non-authorised personnel to access the information.

v) The system needs to have provision for aggregations and generating of summaries

vi) Ability to carry out trend analysis

vii) Provision of reliability of at least 98 percent uptime.

4.13 System Requirements

Since the data mart database software should run on the windows operating system platform, Microsoft SQL Server 2012 is recommended. The Microsoft SQL Server 2012 has the SQL Server Management Studio and SQL Business Intelligence Development Studio that contains the Server 2012 Integration Services (SSIS), SQL Server 2012 Analysis Services (SSAS) and SQL Server 2012 Reporting Services (SSRS) that are ideal for the data mart development. To run SQL Server 2012, the following hardware and software are required.

i) VGA or higher resolution;

ii) A Microsoft mouse or compatible pointing device;

iii) Microsoft Internet Explorer 11.0 SP1 or later;

iv) Internet Information Services (IIS) 6.0 or later;

v) ASP.NET 3.5;

vi) Windows Installer 3.5or later;

vii) Microsoft Data Access Components (MDAC) 2.8 SP1 or later;

viii) Itanium processor or higher;

ix) Duo Core processor 3 GHz or much better;

x) Memory (RAM) of at least 4 GB;

xi) Windows server 2008, or higher Operating system.

4.14 Conclusion

In this chapter findings from the data collected were analysed and presented. Analysis results become a basis for identifying user requirements. It’s these requirements that will guide system development. Also actors to use the system were identified as well as system functionalities and business rules. In the following chapter, we shall discuss the proposed system design.

CHAPTER FIVE

SYSTEM DESIGN

5.1 Introduction

Considering the data and information management challenges evident from the analysis of the current state of information management in the MENTORS project of the infectious disease institute, the need to create a solution that redresses the shortcomings of information management in the project is paramount. This chapter details the steps that were used to develop the data mart system that will be used to provide the solutions to some of the problems identified during the previous phase and the detailed specifications of the system elements.

5.2 Data Mart Architecture Design

The purpose of the data mart was to build a system capable of creating a consolidated data repository of the multiple clinical data source gathered by the MENTORS project, to make information readily available for decision making and research purposes, in addition to addressing the analytical needs of the different sets of users within the MENTORS project. The architectural design of the new system, that shows how data from the different sources flows throughout the system, is presented in figure 5.8. The major two processes of a data mart are data load and access. The loading of the data mart was done through the use of ETL process, while data was accessed using OLAP tools.

In data mart architectural design, data is extracted from clinical data sources gathered from the various project sites across the country. These include; Ms Access databases containing outpatients medical encounters, HIV care, TB treatment and laboratory, and Antenatal data and MySQL server databases containing HIV Care data. The extracted data is loaded into the data staging area for further manipulations which include numerous data transformation such as cleansing the data (correcting misspellings, dealing with missing elements, or parsing into standard formats), combining data from multiple sources, and de-duplicating data. Finally data is loaded into the data mart storage were real integrated clinical data can be obtained. The data mart together with the loaded data serve as the back end database for users' access to integrated clinical data through various tools and interfaces provided by the information delivery environment. End-users at various levels within the MENTORS project are able to interact with the data mart from the information delivery area, where they are able to analyse the data and come up with their reports using different tools.

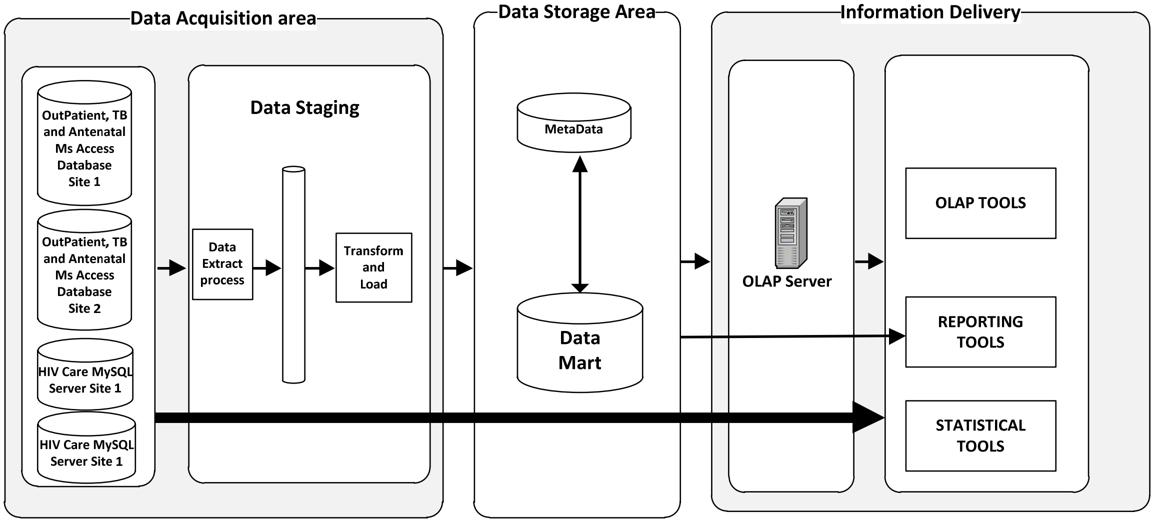


Figure 5.8: Data Mart System Architecture derived from Bhattacharyya, et al (2013)

5.3 Logical Data Mart Design

The logical model is a representation of the data in a way that it can be presented to the business as well as serve as a road map for the physical implementation. The main elements of a logical model are entities, attributes, and relationships. The design of the data mart was accomplished through the fact and dimension tables.

5.3.1 Facts and Dimensions Tables

The Fact table contains the performance measurements associated with a specific business process. A record in a fact table corresponds to a measurement event. These events usually have numeric measurements that quantify the magnitude of the events. These numbers are called facts which represent business measures. Dimensions are the foundation of the dimensional model, describing the objects of the business such as patient, health facility, medical provider, disease and other dimension table to be used in this design of the data mart. Dimension tables also represent the different ways that data can be organised such as health facility, date and patient numbers.

For the MENTORS project data mart, the star schema approach for designing dimensional models was used to come up with the logical dimensional data model for the data mart shown in figure 5.9. The model is composed of five fact tables and eight key dimensions, where each dimension is shared between facts and it can be associated with one or more hierarchies, thus facilitating comparisons between several measures or facts.

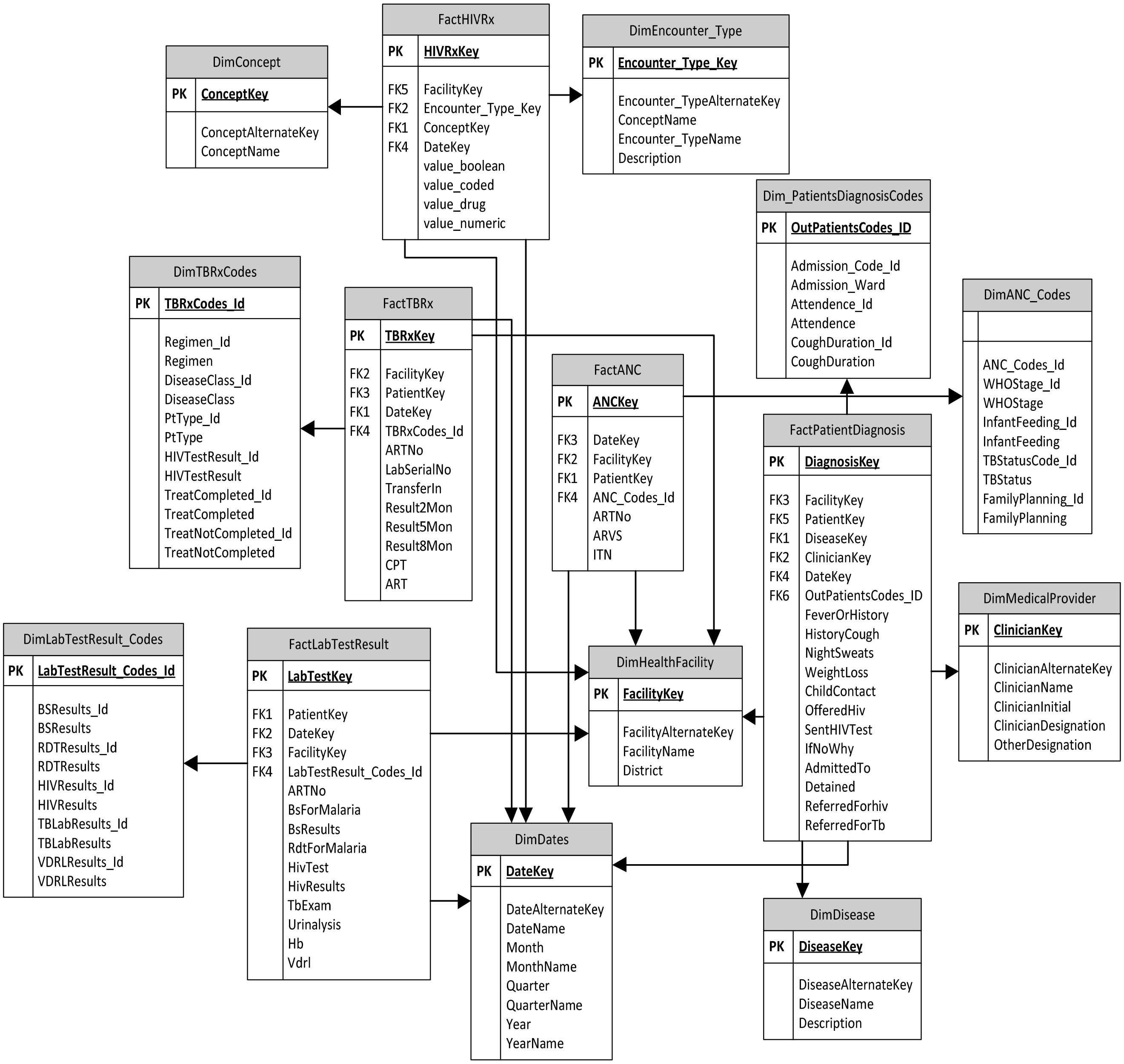


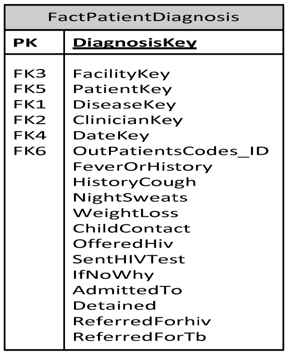
Figure 5.9: Logical Dimensional Data Model

5.3.2 Facts

Facts are based on the user's requirements as specified in the functional requirements of the system analysis section. The fact tables are patient diagnosis, antenatal, TB treatment, HIV treatment, Laboratory test.

Patient Diagnosis Fact

The patient diagnosis fact table shown in figure 5.10 stores facts about each patient visit, diagnosis, admissions and referrals. This table enable the business to analyze outpatient service utilization, disease incidences, admission rates, referral rates and patient linkage to care and treatment. A given patient visit was taken as the grain for the patient diagnosis fact table.

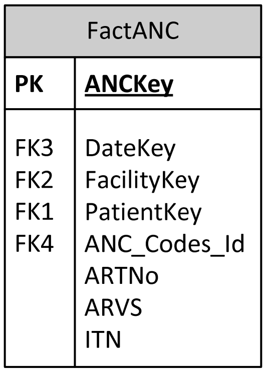


FactPatientDiagnosis	
PK	**DiagnosisKey**
FK3	FacilityKey
FK5	PatientKey
FK1	DiseaseKey
FK2	ClinicianKey
FK4	DateKey
FK6	OutPatientsCodes_ID
	FeverOrHistory
	HistoryCough
	NightSweats
	WeightLoss
	ChildContact
	OfferedHiv
	SentHIVTest
	IfNoWhy
	AdmittedTo
	Detained
	ReferredForhiv
	ReferredForTb

Figure 5.10: Patient Diagnosis Fact

Antenatal Fact Table

The antenatal fact table shown in figure 5.11 stores facts about HIV antenatal attendances and related clinical assessments which include WHO clinical staging, ARV drugs, infant feeding and counseling, TB status of antenatal patients and provision of free insecticide treated nets. This table enable the project to analyze the performance of the health care units in regard to the maternal care and services provided to the HIV antenatal patients. A given patient antenatal visit was taken as the grain for the antenatal fact table.

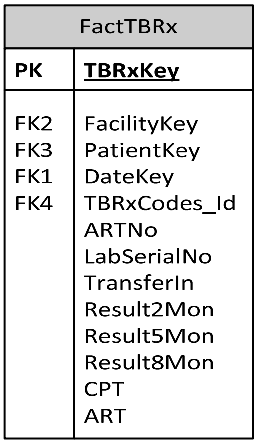


FactANC	
PK	**ANCKey**
FK3 FK2 FK1 FK4	DateKey FacilityKey PatientKey ANC_Codes_Id ARTNo ARVS ITN

Figure 5.11: Antenatal Fact Table

TB Treatment Fact Table

The TB treatment fact table shown in figure 5.12 stores facts about TB patient care and treatment, follow-up progress of treatment and outcomes of treatment for patients registered. The table enable the project to analyze how the health centers are performing in relation to TB, TB/HIV control and TB treatment success rate. A given TB patient treatment record was taken as the grain for the TB Treatment fact table.

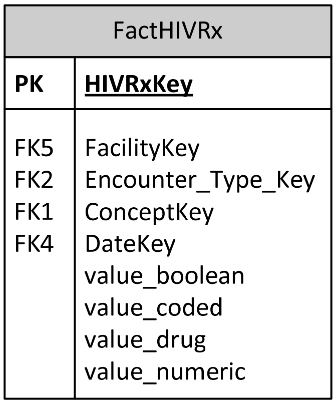


FactTBRx	
PK	**TBRxKey**
FK2 FK3 FK1 FK4	FacilityKey PatientKey DateKey TBRxCodes_Id ARTNo LabSerialNo TransferIn Result2Mon Result5Mon Result8Mon CPT ART

Figure 5.12: TB Treatment Fact Table

HIV Care Fact Table

The HIV care fact table shown in figure 5.13 stores facts about HIV patients accessing HIV care services, appointments of all HIV positive patients on HIV care services. The table enables the project to analyze the performance of the health center in regard to the HIV care and treatment. A given HIV patient appointment record was taken as the grain for the HIV care fact table.

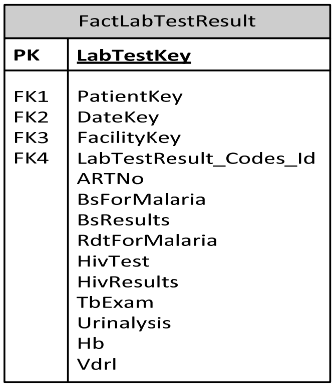


FactHIVRx	
PK	**HIVRxKey**
FK5 FK2 FK1 FK4	FacilityKey Encounter_Type_Key ConceptKey DateKey value_boolean value_coded value_drug value_numeric

Figure 5.13: HIV Care Fact Table

Laboratory Test Fact Table

The Laboratory fact table shown in figure 5.14 stores facts about at the different laboratory test done at the health facilities. The table enables the project to analyze laboratory services utilization at the different health centers. A given laboratory test record was taken as the grain for the Laboratory fact table.



FactLabTestResult	
PK	LabTestKey
FK1 FK2 FK3 FK4	PatientKey DateKey FacilityKey LabTestResult_Codes_Id ARTNo BsForMalaria BsResults RdtForMalaria HivTest HivResults TbExam Urinalysis Hb Vdrl

Figure 5.14: Laboratory Test Fact Table

5.3.3 Dimensions

Eight key dimensions are identified among the five fact tables. The dimensions are; Date, Patient, Health Facility, Disease, Encounter Type, Medical Provider, Patient Diagnosis codes, concept. The confirmed dimensions are date, patient and healthy facility. The dimension tables are described in Table 5.4 below.

Table Name	Description
Patient Dimension	This table contains patient information such as PatientAlternateKey, PatientName, Gender, Age. The data is used to show patient bio data for the integrated clinical data.
Data Dimension	This table stores all dates the when the clinical data was collected.
Health Facility Dimension	This table stores all the health facility and districts where they are located.

Concept Dimension	Table stores all the clinical observation related to HIV care and treatment
Encounter Type Dimension	Categorizes patient visit under HIV care.
Disease Dimension	Table stores the different diseases and the categories to which they belong.
Medical Provider Dimension	Stores all medical provider information which includes ClinicianAlternateKey, ClinicianName and ClinicianDesignation
Patient diagnosis codes dimension	Stores information on all codes in the patient diagnosis fact table

Table 5.4: Dimension Tables

5.4 Design of the Physical Database

Having designed the logical model of the data mart, this model was converted into a description of the physical database including tables and constraints. Given the nature and types of queries the data mart users usually execute, the data mart database was optimized to perform well for those types of queries. The data mart tables and constraints were designed as shown in figure (5.15). The data mart physical table description are in Appendix C.

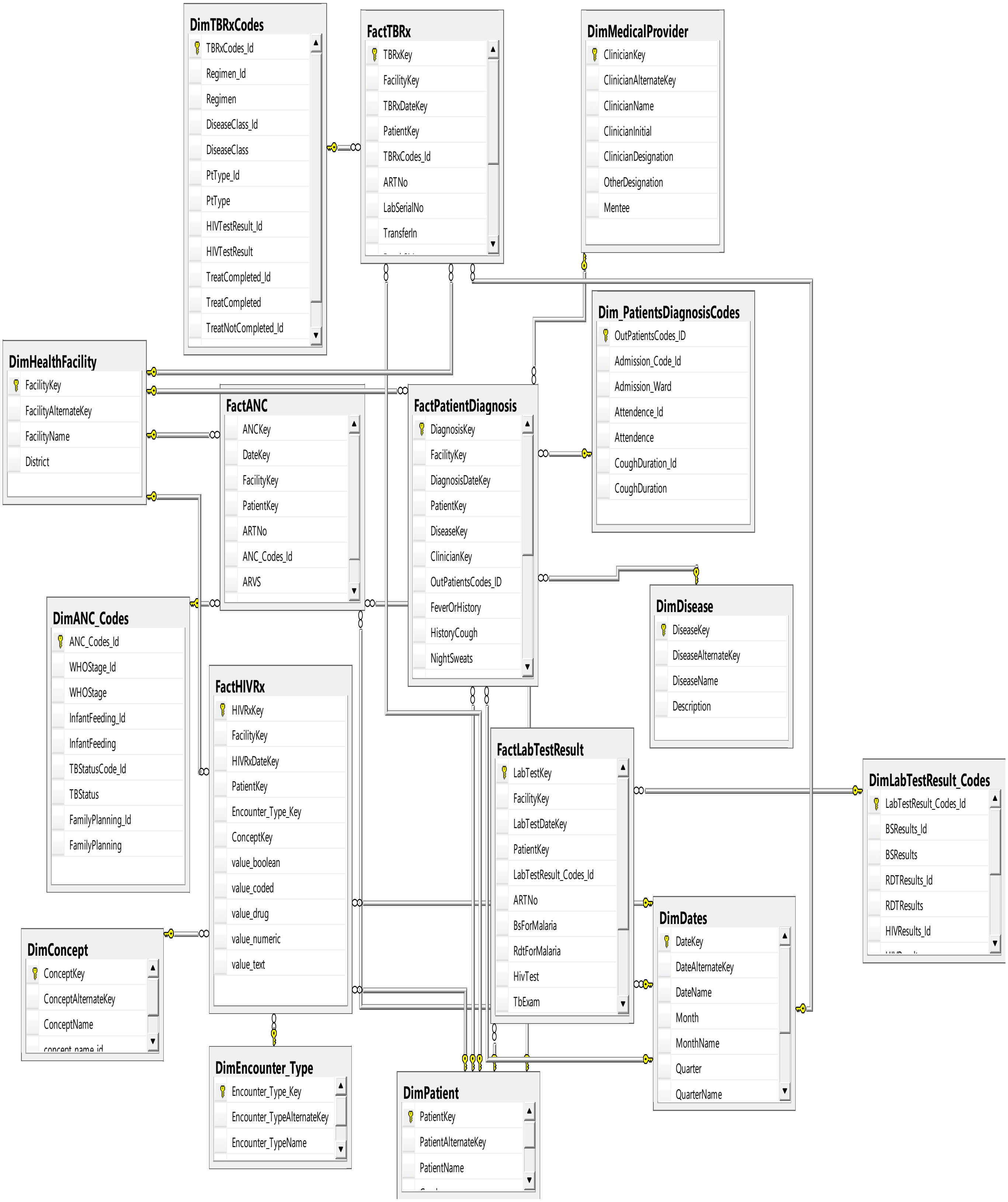


Figure 5.15: Physical Design of the Data Mart

CHAPTER SIX

SYSTEM IMPLEMENTATION

6.1 Introduction

This chapter describes the implementation of the design described in the preceding chapter to meet the requirements of the users in the MENTORS project of IDI. Microsoft technologies (Microsoft SQL Server 2012 and Microsoft Visual Studio 2010) were chosen as the technologies to build the data mart solution, because they provided all tools needed to support the building of the different components of the data mart system.

Microsoft SQL Server 2012 covers not only relational database management service, but also integrated service, analysis service and reporting service. Among them, integration service helped to integrate clinical data from different the multiple clinical data sources, by providing function of data extraction, transformation and load; analysis services provided the function of OLAP to enable analysis of the prevailing situation and predict the future trend; reporting services provided the function of creating various forms of data report and graphical display of the analysis result. The code which was used to create the data mart solution was written using the mentioned Microsoft technologies. (See Code as attachment in appendix D).

6.2 Database Development

The staging database and the physical data mart storage, for use as the central storage for the clinical data from the multiple clinical data sources, were implemented in Microsoft SQL server database engine. The T-SQL script for the implementation of the data mart database are in appendix D.

6.3 Data Extraction, Transformation, and Load (ETL)

This was one of the critical processes of the data mart implementation, where clinical data extracted from the multiple clinical data sources was loaded into the staging database and the data mart for centralized storage. As mentioned in the earlier chapters of this report, the clinical data systems gathered by the MENTORS project of IDI from its various implementation sites were used as the source systems from which data was extracted. These data sources included; MySQL database storing HIV care encounter data and Ms Access database storing outpatients medical encounters, Antenatal, TB treatment and laboratory data. The ETL design for data load from source to the staging database and data mart was designed using Microsoft SQL Server Integration Services (SSIS). The ETL process is explained in the following steps:

6.3.1 Step 1 (Data is extracted from the multiple sources)

The first step in the process of consolidating clinical data centrally into the data mart environment, involved the extraction of the desired clinical data from the multiple clinical data sources into the staging area for further manipulation. The data loading process to the staging database is done without much transformation to ensure data is copied at higher speed. This is shown in figure 6.16 and figure 6.17

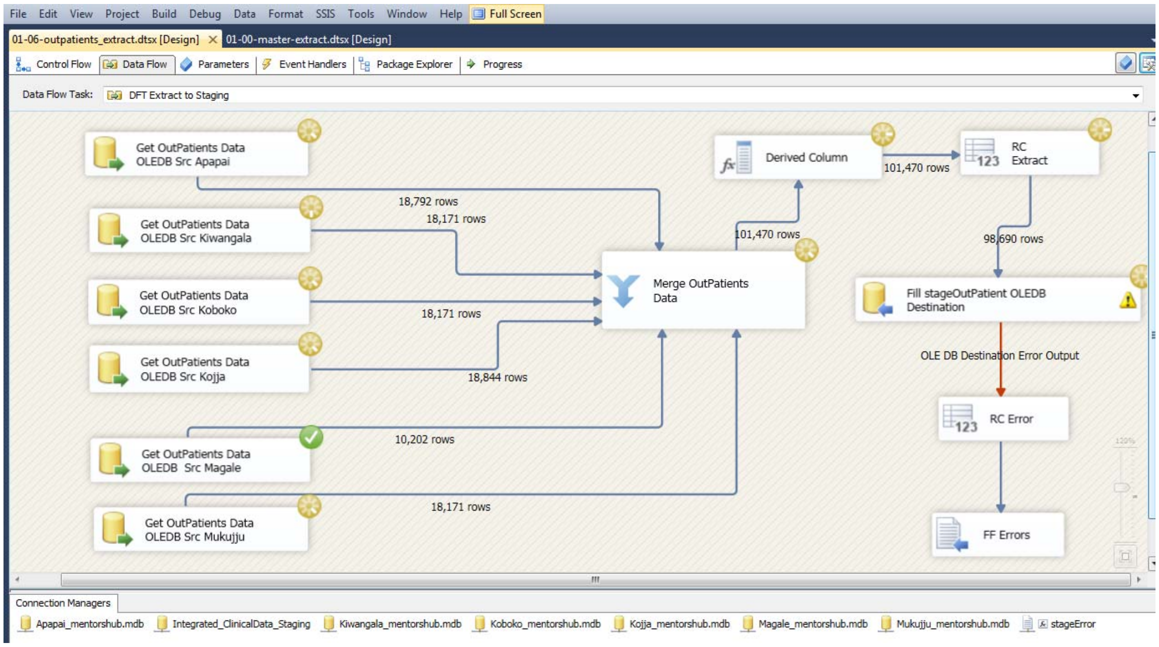
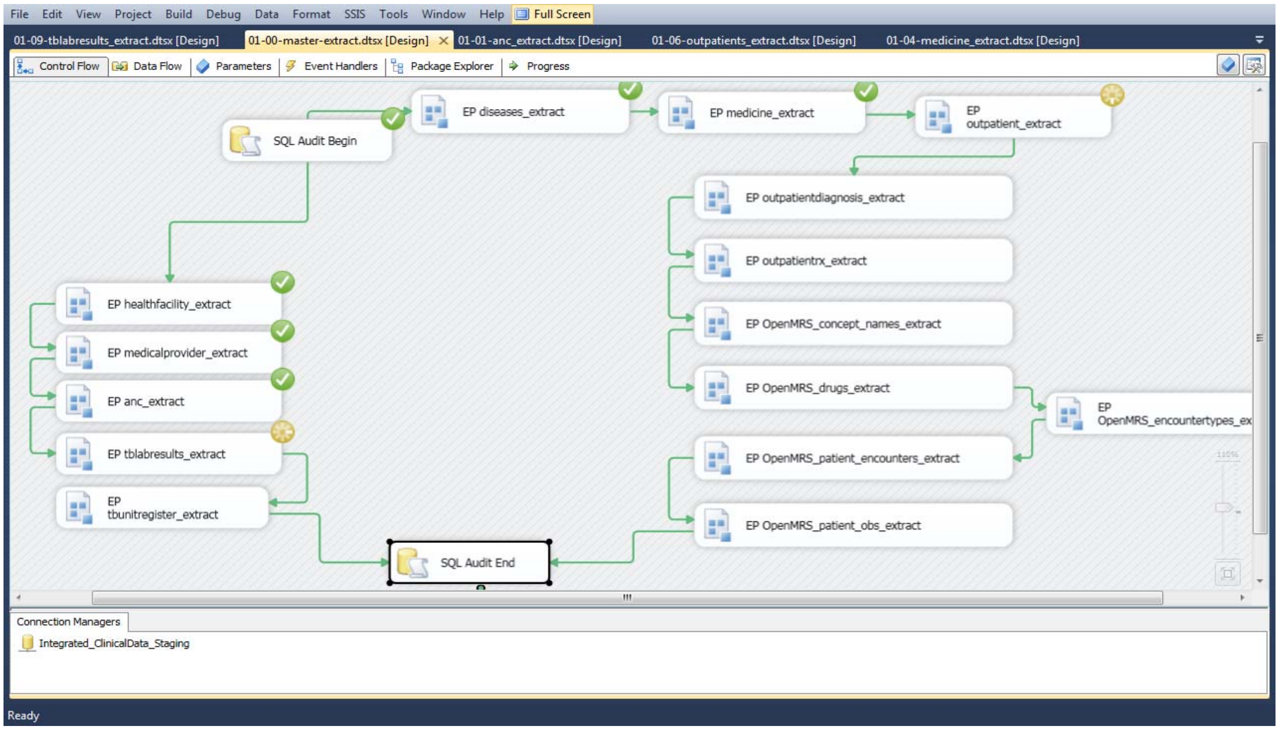

File Edit View Project Build Debug Data Format SSIS Tools Window Help Full Screen
01-06-outpatients_extract.dtsx [Design]
01-00-master-extract.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explorer
Progress
Data Flow Task: DFT Extract to Staging
Get OutPatients Data OLEDB Src Apapai
Get OutPatients Data OLEDB Src Kiwangala
Get OutPatients Data OLEDB Src Koboko
Get OutPatients Data OLEDB Src Kojja
Get OutPatients Data OLEDB Src Magale
Get OutPatients Data OLEDB Src Mukujju
18,792 rows
18,171 rows
18,171 rows
18,844 rows
10,202 rows
18,171 rows
Merge OutPatients Data
101,470 rows
Derived Column
101,470 rows
RC Extract
98,690 rows
Fill stageOutPatient OLEDB Destination
OLE DB Destination Error Output
RC Error
FF Errors
Connection Managers
Apapai_mentorshub.mdb
Integrated_ClinicalData_Staging
Kiwangala_mentorshub.mdb
Koboko_mentorshub.mdb
Kojja_mentorshub.mdb
Magale_mentorshub.mdb
Mukujju_mentorshub.mdb
stageError

Figure 6.16: ETL Package in SSIS extracting Outpatients Encounters source tables from various similar data sources into a single staging data table.


File Edit View Project Build Debug Data Format SSIS Tools Window Help Full Screen
01-09-tblabresults_extract.dtsx [Design]
01-00-master-extract.dtsx [Design]
01-01-anc_extract.dtsx [Design]
01-06-outpatients_extract.dtsx [Design]
01-04-medicine_extract.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explorer
Progress
SQL Audit Begin
EP diseases_extract
EP medicine_extract
EP outpatient_extract
EP outpatientdiagnosis_extract
EP outpatientrx_extract
EP OpenMRS_concept_names_extract
EP OpenMRS_drugs_extract
EP OpenMRS_encountertypes_ex
EP OpenMRS_patient_encounters_extract
EP OpenMRS_patient_obs_extract
EP healthfacility_extract
EP medicalprovider_extract
EP anc_extract
EP tblabresults_extract
EP tbunitregister_extract
SQL Audit End
Connection Managers
Integrated_ClinicalData_Staging
Ready

. Figure 6.17: Master ETL package in SSIS extracting all source tables to staging database.

6.3.2 Step 2 (Data is transformed and cleaned before being loaded)

Before extracted data was loaded to the target system (data mart), it had to be transformed and cleaned in order to add value and also improve the quality. The data transformation was done using built-in transformations contained in SSIS. The transformation and cleaning involved, joining source database tables, reformatting some of the columns to confirm to constraints in the target data mart schema, generating surrogate keys , combining data from multiple data sources, correcting misspelling, dealing with missing elements or parsing into standard formats and de-duplicating data.

The following ETL design figure 6.18 conforms the patient data tables from different sources into a single confirmed patient dimension in the data mart.

.

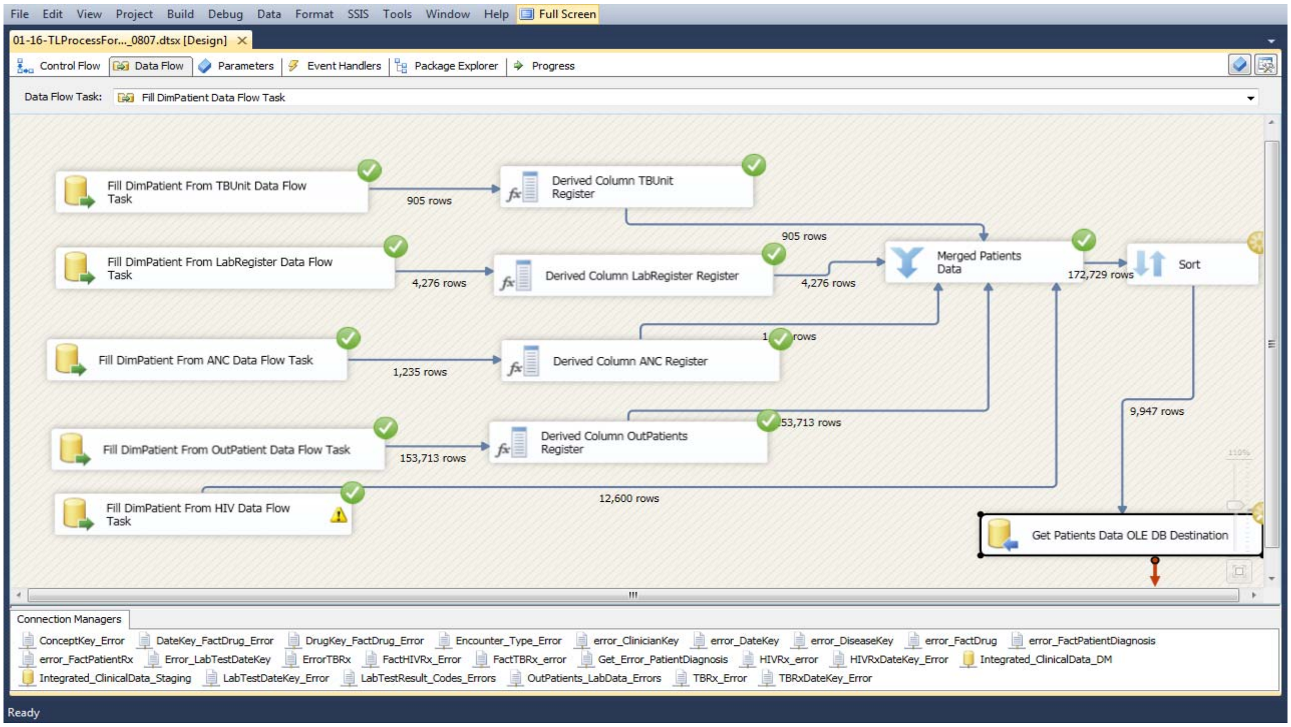
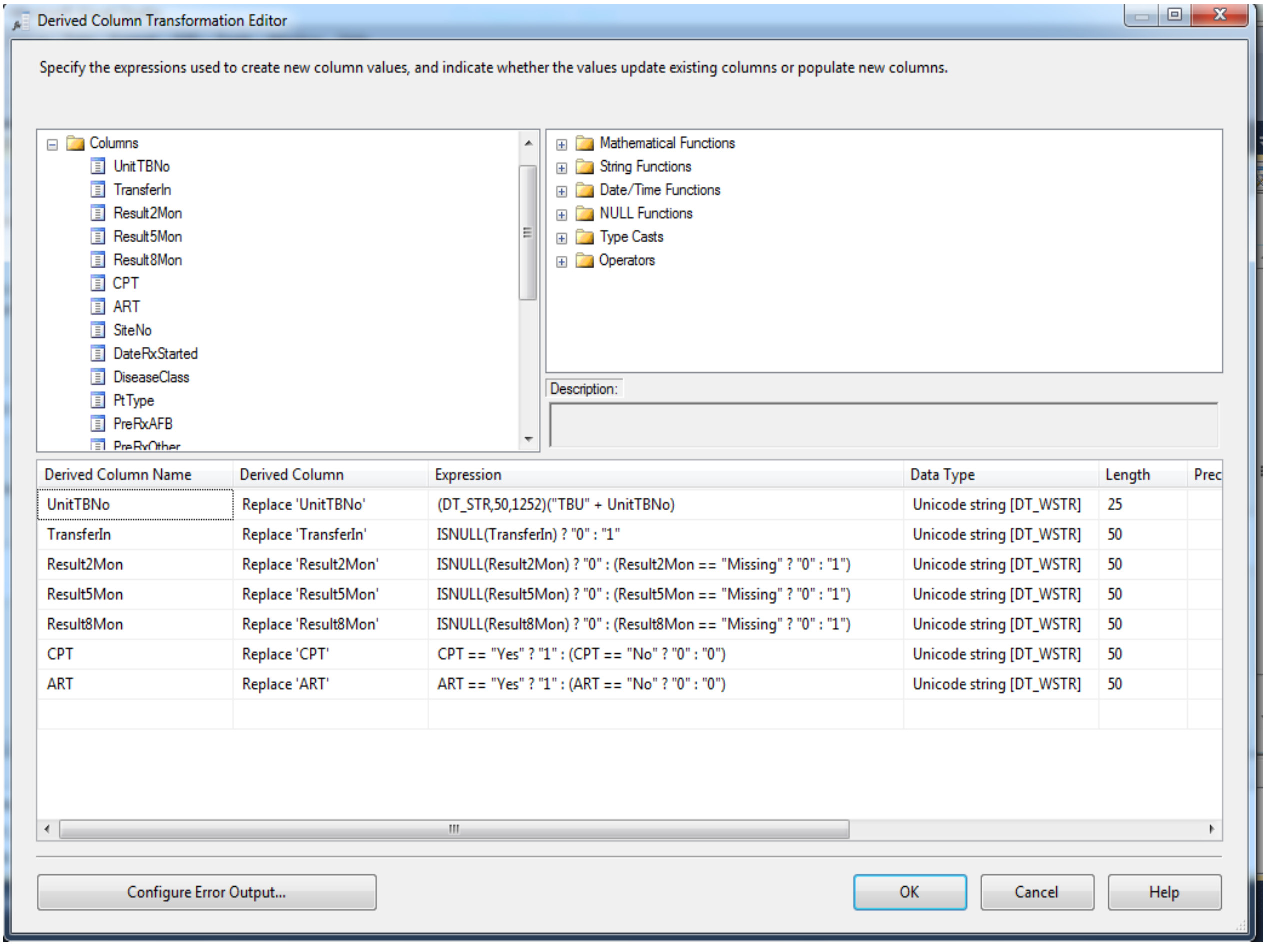

File
Edit
View
Project
Build
Debug
Data
Format
SSIS
Tools
Window
Help
Full Screen
01-16-TLProcessFor..._0807.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explorer
Progress
Data Flow Task:
Fill DimPatient Data Flow Task
Fill DimPatient From TBUnit Data Flow Task
905 rows
Derived Column TBUnit Register
905 rows
Fill DimPatient From LabRegister Data Flow Task
4,276 rows
Derived Column LabRegister Register
4,276 rows
Merged Patients Data
172,729 rows
Sort
Fill DimPatient From ANC Data Flow Task
1,235 rows
Derived Column ANC Register
1 rows
9,947 rows
Fill DimPatient From OutPatient Data Flow Task
153,713 rows
Derived Column OutPatients Register
53,713 rows
Fill DimPatient From HIV Data Flow Task
12,600 rows
Get Patients Data OLE DB Destination
Connection Managers
ConceptKey_Error
DateKey_FactDrug_Error
DrugKey_FactDrug_Error
Encounter_Type_Error
error_ClinicianKey
error_DateKey
error_DiseaseKey
error_FactDrug
error_FactPatientDiagnosis
error_FactPatientRx
Error_LabTestDateKey
ErrorTBRx
FactHIVRx_Error
FactTBRx_error
Get_Error_PatientDiagnosis
HIVRx_error
HIVRxDateKey_Error
Integrated_ClinicalData_DM
Integrated_ClinicalData_Staging
LabTestDateKey_Error
LabTestResult_Codes_Errors
OutPatients_LabData_Errors
TBRx_Error
TBRxDateKey_Error
Ready

Figure 6.18: ETL design confirming dimensions

The following ETL design figure 6.19 uses the derived column transformation to reformat column values (UnitNo -> “TBU + UnitNo”) and checks for null values which are substituted with value “0”..



Specify the expressions used to create new column values, and indicate whether the values update existing columns or populate new columns.

Derived Column Name	Derived Column	Expression	Data Type	Length	Prec
UnitTBNo	Replace 'UnitTBNo'	(DT_STR,50,1252)("TBU" + UnitTBNo)	Unicode string [DT_WSTR]	25	
TransferIn	Replace 'TransferIn'	ISNULL(TransferIn) ? "0" : "1"	Unicode string [DT_WSTR]	50	
Result2Mon	Replace 'Result2Mon'	ISNULL(Result2Mon) ? "0" : (Result2Mon == "Missing" ? "0" : "1")	Unicode string [DT_WSTR]	50	
Result5Mon	Replace 'Result5Mon'	ISNULL(Result5Mon) ? "0" : (Result5Mon == "Missing" ? "0" : "1")	Unicode string [DT_WSTR]	50	
Result8Mon	Replace 'Result8Mon'	ISNULL(Result8Mon) ? "0" : (Result8Mon == "Missing" ? "0" : "1")	Unicode string [DT_WSTR]	50	
CPT	Replace 'CPT'	CPT == "Yes" ? "1" : (CPT == "No" ? "0" : "0")	Unicode string [DT_WSTR]	50	
ART	Replace 'ART'	ART == "Yes" ? "1" : (ART == "No" ? "0" : "0")	Unicode string [DT_WSTR]	50	

Figure 6.19: ETL design derived column SSIS tool to check for Null

The following ETL design figure 6.20 flags the erroneous dates as shown in figure 6.21 to error destination file for manual intervention.

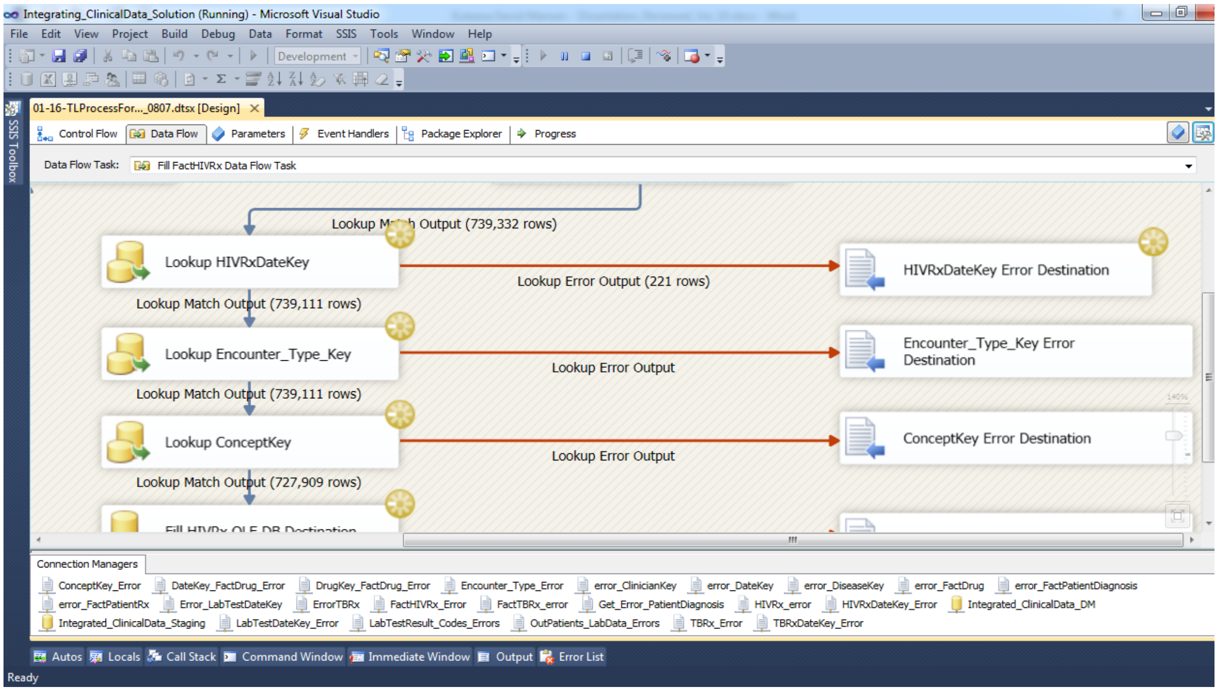
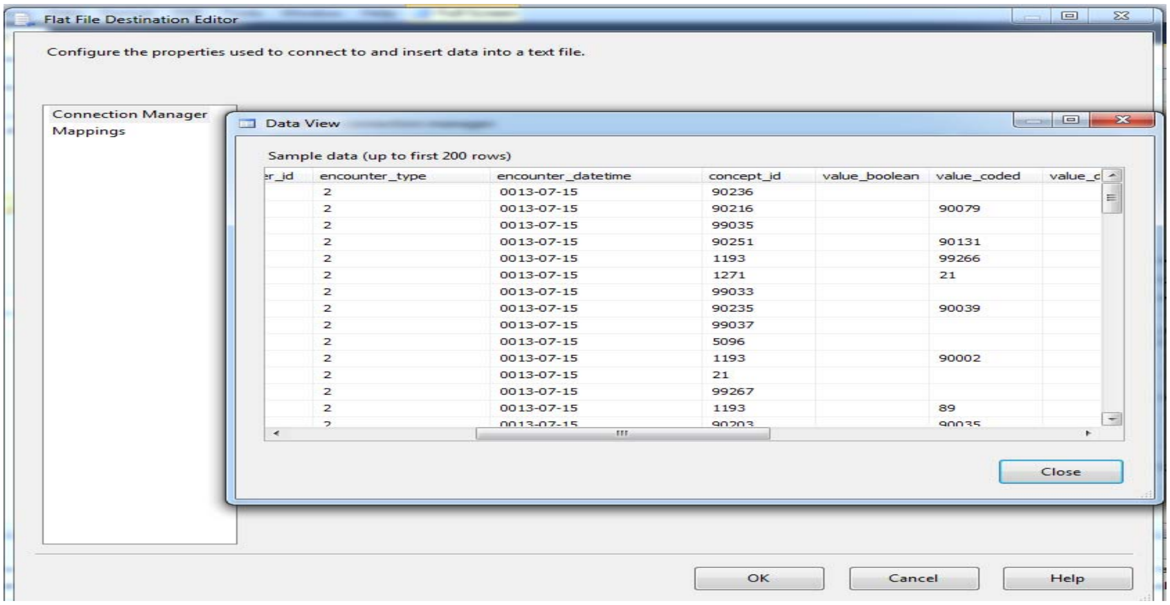

Integrating_ClinicalData_Solution (Running) - Microsoft Visual Studio
File Edit View Project Build Debug Data Format SSIS Tools Window Help
Development
SSIS Toolbox
01-16-TLProcessFor..._0807.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explorer
Progress
Data Flow Task: Fill FactHIVRx Data Flow Task
Lookup Match Output (739,332 rows)
Lookup HIVRxDateKey
Lookup Error Output (221 rows)
HIVRxDateKey Error Destination
Lookup Match Output (739,111 rows)
Lookup Encounter_Type_Key
Lookup Error Output
Encounter_Type_Key Error Destination
Lookup Match Output (739,111 rows)
Lookup ConceptKey
Lookup Error Output
ConceptKey Error Destination
Lookup Match Output (727,909 rows)
Fill HIVRx OLE DB Destination
Connection Managers
ConceptKey_Error
DateKey_FactDrug_Error
DrugKey_FactDrug_Error
Encounter_Type_Error
error_ClinicianKey
error_DateKey
error_DiseaseKey
error_FactDrug
error_FactPatientDiagnosis
error_FactPatientRx
Error_LabTestDateKey
ErrorTBRx
FactHIVRx_Error
FactTBRx_error
Get_Error_PatientDiagnosis
HIVRx_error
HIVRxDateKey_Error
Integrated_ClinicalData_DM
Integrated_ClinicalData_Staging
LabTestDateKey_Error
LabTestResult_Codes_Errors
OutPatients_LabData_Errors
TBRx_Error
TBRxDateKey_Error
Autos
Locals
Call Stack
Command Window
Immediate Window
Output
Error List
Ready

Figure 6.20: ETL design erroneous HIVRxDatekey flagged to Error Destination file



Configure the properties used to connect to and insert data into a text file.

Sample data (up to first 200 rows)

er_id	encounter_type	encounter_datetime	concept_id	value_boolean	value_coded	value_c
	2	0013-07-15	90236			
	2	0013-07-15	90216		90079	
	2	0013-07-15	99035			
	2	0013-07-15	90251		90131	
	2	0013-07-15	1193		99266	
	2	0013-07-15	1271		21	
	2	0013-07-15	99033			
	2	0013-07-15	90235		90039	
	2	0013-07-15	99037			
	2	0013-07-15	5096			
	2	0013-07-15	1193		90002	
	2	0013-07-15	21			
	2	0013-07-15	99267			
	2	0013-07-15	1193		89	
	2	0013-07-15	90203		90035	

Figure 6.21: ETL design Flat file destination editor showing erroneous encounter_dates

6.3.3 Step 3 (Data is loaded into the data mart)

Once the data quality and business rules have been applied to the extracted clinical data in the staging area, both the dimensions and fact tables can be loaded as required by the target system. The different task involved in the loading process mainly focused on dimension table processing such as surrogate key assignment, code lookup to provide appropriate definitions. Loading of the dimensions and fact tables was sequential as shown in figure 6.22, where the dimension tables are loaded before loading the fact tables. Loading the fact tables was the last step in the process of data mart loading.

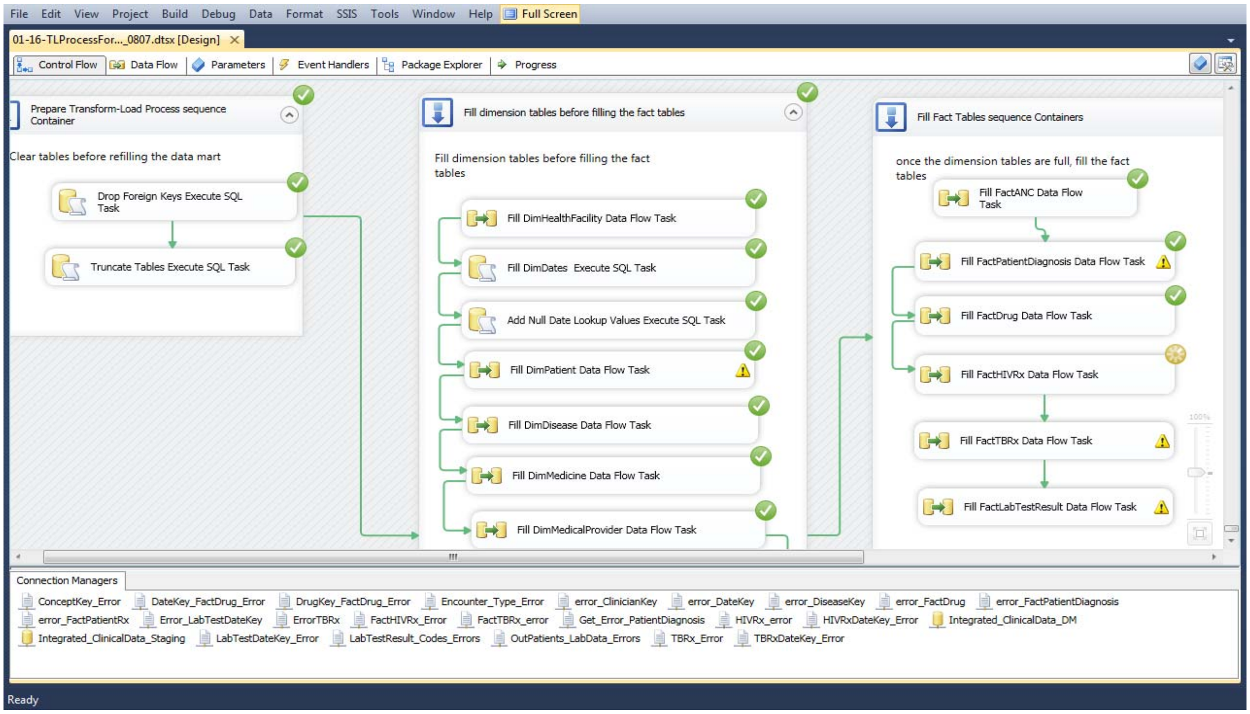

File Edit View Project Build Debug Data Format SSIS Tools Window Help Full Screen
01-16-TLProcessFor..._0807.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explorer
Progress
Prepare Transform-Load Process sequence Container
Clear tables before refilling the data mart
Drop Foreign Keys Execute SQL Task
Truncate Tables Execute SQL Task
Fill dimension tables before filling the fact tables
Fill dimension tables before filling the fact tables
Fill DimHealthFacility Data Flow Task
Fill DimDates Execute SQL Task
Add Null Date Lookup Values Execute SQL Task
Fill DimPatient Data Flow Task
Fill DimDisease Data Flow Task
Fill DimMedicine Data Flow Task
Fill DimMedicalProvider Data Flow Task
Fill Fact Tables sequence Containers
once the dimension tables are full, fill the fact tables
Fill FactANC Data Flow Task
Fill FactPatientDiagnosis Data Flow Task
Fill FactDrug Data Flow Task
Fill FactHIVRx Data Flow Task
Fill FactTBRx Data Flow Task
Fill FactLabTestResult Data Flow Task
Connection Managers
ConceptKey_Error
DateKey_FactDrug_Error
DrugKey_FactDrug_Error
Encounter_Type_Error
error_ClinicianKey
error_DateKey
error_DiseaseKey
error_FactDrug
error_FactPatientDiagnosis
error_FactPatientRx
Error_LabTestDateKey
ErrorTBRx
FactHIVRx_Error
FactTBRx_error
Get_Error_PatientDiagnosis
HIVRx_error
HIVRxDateKey_Error
Integrated_ClinicalData_DM
Integrated_ClinicalData_Staging
LabTestDateKey_Error
LabTestResult_Codes_Errors
OutPatients_LabData_Errors
TBRx_Error
TBRxDateKey_Error
Ready

Figure 6.22: SSIS ETL package loading all the dimensions and fact tables of the data mart

After loading clinical data into the data mart storage system, there was need for end users to access the data. Reports could be created directly from the data mart, or pull data analysis results to your reports through OLAP cubes. For this project, reporting and analysis services for SQL server were used to develop different types of reports to enable the end users explore data stored in the data mart.

6.4 OLAP Cubes

OLAP cubes were developed using Microsoft’s SQL server analysis services (SSAS). SSAS is multidimensional database server in which data takes the form of measures, dimensions, hierarchies and cubes.

OLAP cubes were processed to provide end users with a mechanism for viewing and analysing data mart information very easily and quickly. The structure of the OLAP cube developed showing the resulting multidimensional model and corresponding measures and dimensions is shown in figure 6.23. After processing and deploying the cube, cube data is viewable on the browser tab in cube designer and dimension data is viewable on the browser tab in dimensional designer as shown in figure 6.24. Alternatively Ms Excel shortcut which can be started from within SQL Server Data Tools or SQL Server Analysis Server can also be used to browse the cube. Excel opens with a pivot table already in the worksheet and predefined connection to the model workspace database shown in figure 6.25.

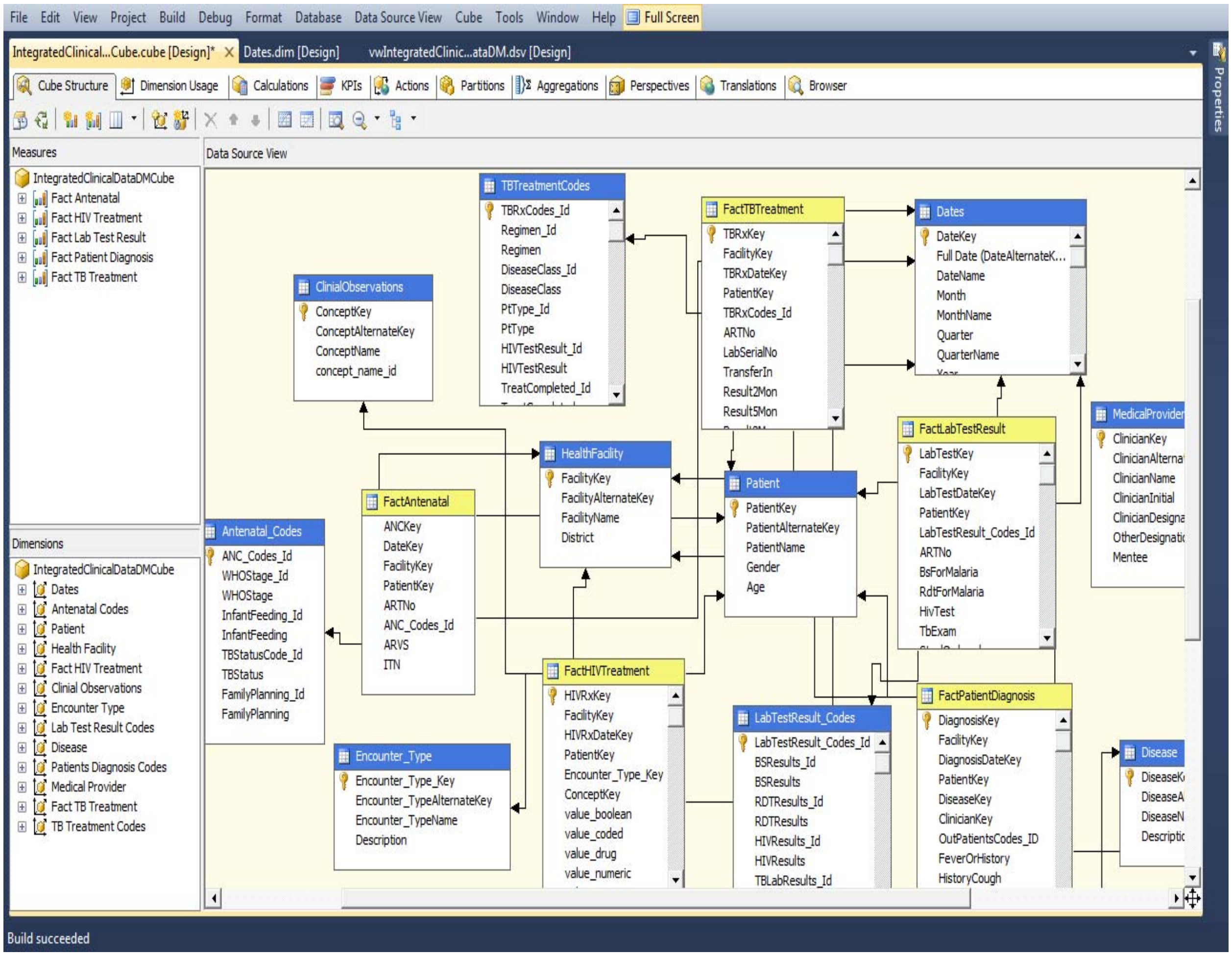
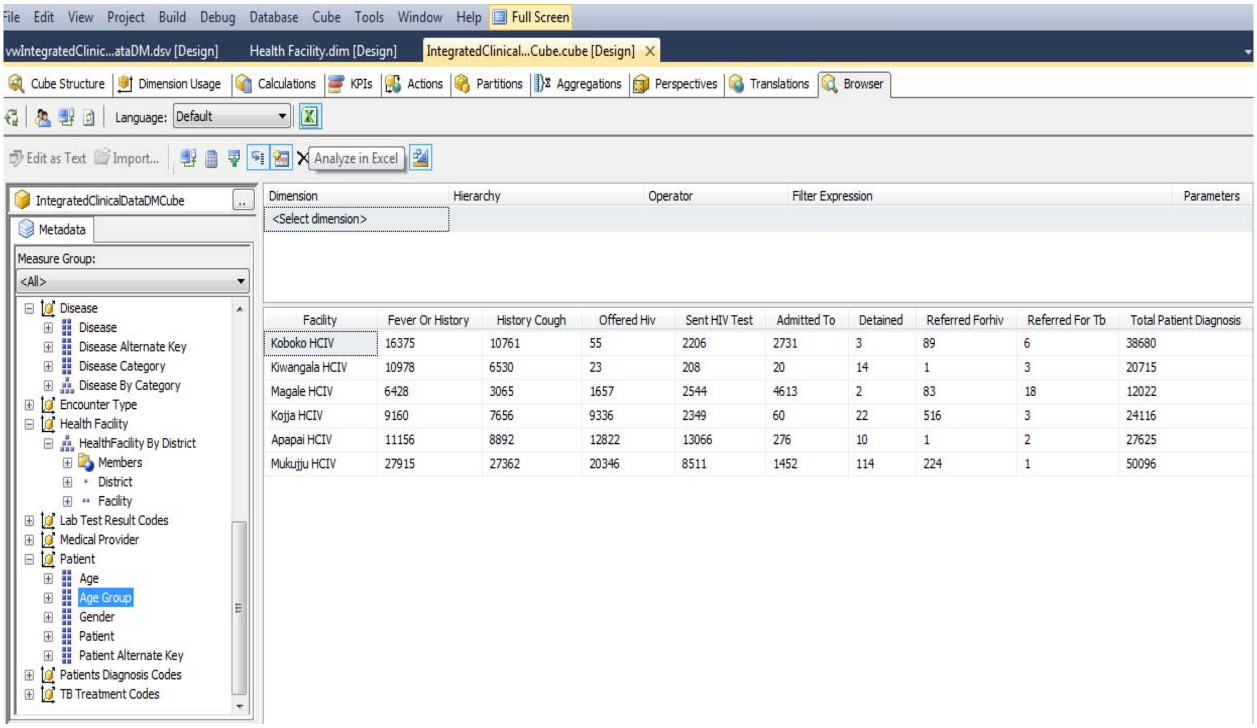


Figure 6.23: Multidimensional model

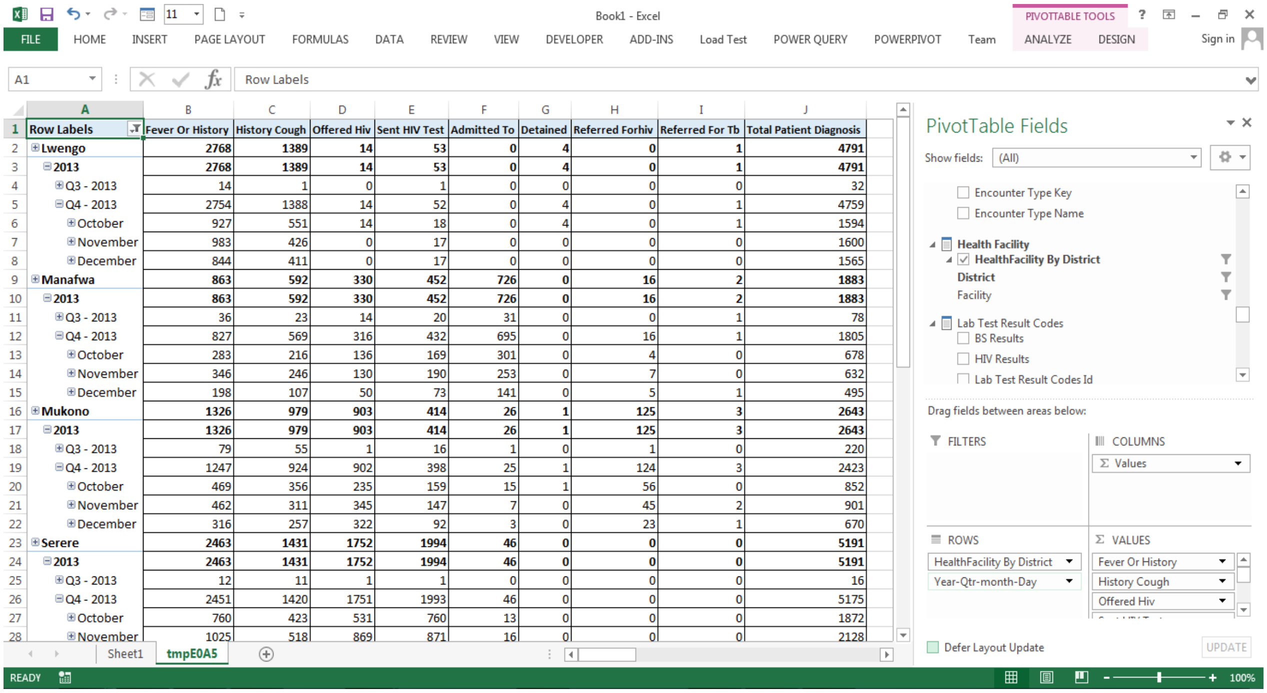


IntegratedClinicalDataDMCube

Measure Group: <All>

Facility	Fever Or History	History Cough	Offered Hiv	Sent HIV Test	Admitted To	Detained	Referred Forhiv	Referred For Tb	Total Patient Diagnosis
Koboko HCIV	16375	10761	55	2206	2731	3	89	6	38680
Kiwangala HCIV	10978	6530	23	208	20	14	1	3	20715
Magale HCIV	6428	3065	1657	2544	4613	2	83	18	12022
Kojja HCIV	9160	7656	9336	2349	60	22	516	3	24116
Apapai HCIV	11156	8892	12822	13066	276	10	1	2	27625
Mukujju HCIV	27915	27362	20346	8511	1452	114	224	1	50096

Figure 6.24: Analysis of Outpatients Exam History, Admissions and Diagnosis by Health Facility



Row Labels	Fever Or History	History Cough	Offered Hiv	Sent HIV Test	Admitted To	Detained	Referred Forhiv	Referred For Tb	Total Patient Diagnosis
Lwengo	**2768**	**1389**	**14**	**53**	**0**	**4**	**0**	**1**	**4791**
2013	**2768**	**1389**	**14**	**53**	**0**	**4**	**0**	**1**	**4791**
Q3 - 2013	14	1	0	1	0	0	0	0	32
Q4 - 2013	2754	1388	14	52	0	4	0	1	4759
October	927	551	14	18	0	4	0	1	1594
November	983	426	0	17	0	0	0	0	1600
December	844	411	0	17	0	0	0	0	1565
Manafwa	**863**	**592**	**330**	**452**	**726**	**0**	**16**	**2**	**1883**
2013	**863**	**592**	**330**	**452**	**726**	**0**	**16**	**2**	**1883**
Q3 - 2013	36	23	14	20	31	0	0	1	78
Q4 - 2013	827	569	316	432	695	0	16	1	1805
October	283	216	136	169	301	0	4	0	678
November	346	246	130	190	253	0	7	0	632
December	198	107	50	73	141	0	5	1	495
Mukono	**1326**	**979**	**903**	**414**	**26**	**1**	**125**	**3**	**2643**
2013	**1326**	**979**	**903**	**414**	**26**	**1**	**125**	**3**	**2643**
Q3 - 2013	79	55	1	16	1	0	1	0	220
Q4 - 2013	1247	924	902	398	25	1	124	3	2423
October	469	356	235	159	15	1	56	0	852
November	462	311	345	147	7	0	45	2	901
December	316	257	322	92	3	0	23	1	670
Serere	**2463**	**1431**	**1752**	**1994**	**46**	**0**	**0**	**0**	**5191**
2013	**2463**	**1431**	**1752**	**1994**	**46**	**0**	**0**	**0**	**5191**
Q3 - 2013	12	11	1	1	0	0	0	0	16
Q4 - 2013	2451	1420	1751	1993	46	0	0	0	5175
October	760	423	531	760	13	0	0	0	1872
November	1025	518	869	871	16	0	0	0	2128

PivotTable Fields

Show fields: (All)

ROWS: HealthFacility By District, Year-Qtr-month-Day

COLUMNS: Σ Values

VALUES: Fever Or History, History Cough, Offered Hiv

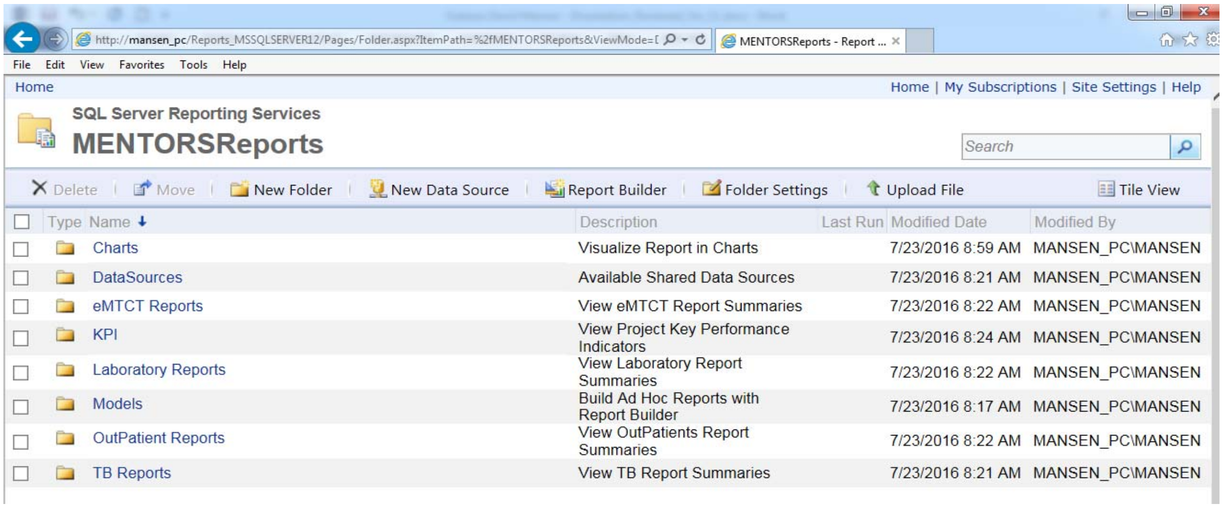
Figure 6.25: Excel Pivot Table Analysis of Outpatients Exam History, Admissions and Diagnosis by Health Facility by Quarter by Month

Excel offered a better browsing experience where users could explore cube data interactively as shown in figure 6.25 using horizontal and vertical axes to analyse the relationships in the data. Because of the drill down and interactivity that excel pivot tables provided for the user, it was the preferred solution for browsing cube data in this project.

6.5 Web-Based Reporting Interface

A web-based user reporting interface was implemented using Microsoft SQL server reporting services (SSRS). This tool delivered pre-packaged standard report formats that were easy to understand and easy to use.

Given that the reports produced in this project contained sensitive data, there was a need to secure the data so that only authorized users were able to access the reports. Reporting services offered the tools for accomplishing user security through security roles and item level security which offered control over who had access to each report and resources. Figure 6.26 shows SSRS report manager web interface which can be used to setup security to control access to the reports and also organize reports to assist users in finding what they need.



http://mansen_pc/Reports_MSSQLSERVER12/Pages/Folder.aspx?ItemPath=%2fMENTORSReports&ViewMode=[
MENTORSReports - Report ...
File Edit View Favorites Tools Help
Home
Home | My Subscriptions | Site Settings | Help
Search
Delete | Move | New Folder | New Data Source | Report Builder | Folder Settings | Upload File
Tile View

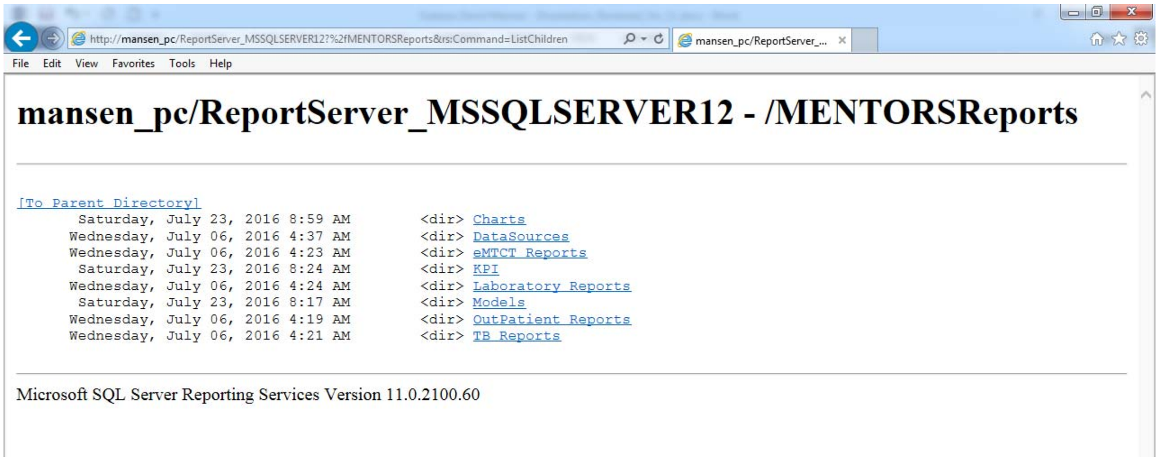
SQL Server Reporting Services

MENTORSReports

□	Type	Name ↓	Description	Last Run	Modified Date	Modified By
□		Charts	Visualize Report in Charts		7/23/2016 8:59 AM	MANSEN_PC\MANSEN
□		DataSources	Available Shared Data Sources		7/23/2016 8:21 AM	MANSEN_PC\MANSEN
□		eMTCT Reports	View eMTCT Report Summaries		7/23/2016 8:22 AM	MANSEN_PC\MANSEN
□		KPI	View Project Key Performance Indicators		7/23/2016 8:24 AM	MANSEN_PC\MANSEN
□		Laboratory Reports	View Laboratory Report Summaries		7/23/2016 8:22 AM	MANSEN_PC\MANSEN
□		Models	Build Ad Hoc Reports with Report Builder		7/23/2016 8:17 AM	MANSEN_PC\MANSEN
□		OutPatient Reports	View OutPatients Report Summaries		7/23/2016 8:22 AM	MANSEN_PC\MANSEN
□		TB Reports	View TB Report Summaries		7/23/2016 8:21 AM	MANSEN_PC\MANSEN

Figure 6.26: Report folders on the Report Manager Web application

Report manager enables users to easily navigate the report folders to access the various summarized clinical information that would aids them in the decision making process. Besides the report manager, users can also obtain access to reports via the report server web service user interface which can be accesses via the URL: http://localhost/ReportServer as shown in figure 6.27.



http://mansen_pc/ReportServer_MSSQLSERVER12?%2fMENTORSReports&rs:Command=ListChildren
mansen_pc/ReportServer_...
File Edit View Favorites Tools Help

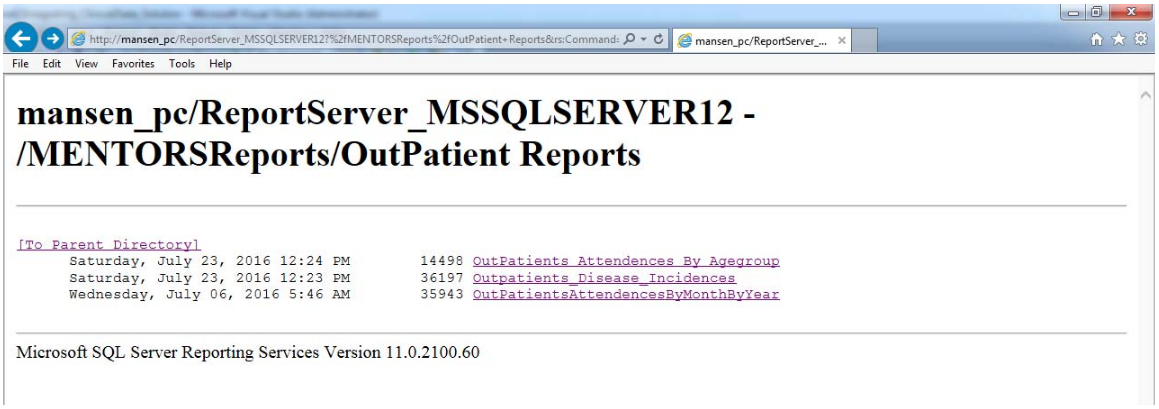
mansen_pc/ReportServer_MSSQLSERVER12 - /MENTORSReports

```
[To Parent Directory]
   Saturday, July 23, 2016 8:59 AM        <dir> Charts
  Wednesday, July 06, 2016 4:37 AM        <dir> DataSources
  Wednesday, July 06, 2016 4:23 AM        <dir> eMTCT Reports
   Saturday, July 23, 2016 8:24 AM        <dir> KPI
  Wednesday, July 06, 2016 4:24 AM        <dir> Laboratory Reports
   Saturday, July 23, 2016 8:17 AM        <dir> Models
  Wednesday, July 06, 2016 4:19 AM        <dir> OutPatient Reports
  Wednesday, July 06, 2016 4:21 AM        <dir> TB Reports
```

Microsoft SQL Server Reporting Services Version 11.0.2100.60

Figure 6.27: Report folders on the Report Server web service

In figure 6.27 clicking one of these links such as the Outpatient Reports directory displays the objects within the subfolder as shown figure 6.28 below.



http://mansen_pc/ReportServer_MSSQLSERVER12?%2fMENTORSReports%2fOutPatient+Reports&rs:Command:
mansen_pc/ReportServer_...
File
Edit
View
Favorites
Tools
Help

mansen_pc/ReportServer_MSSQLSERVER12 - /MENTORSReports/OutPatient Reports

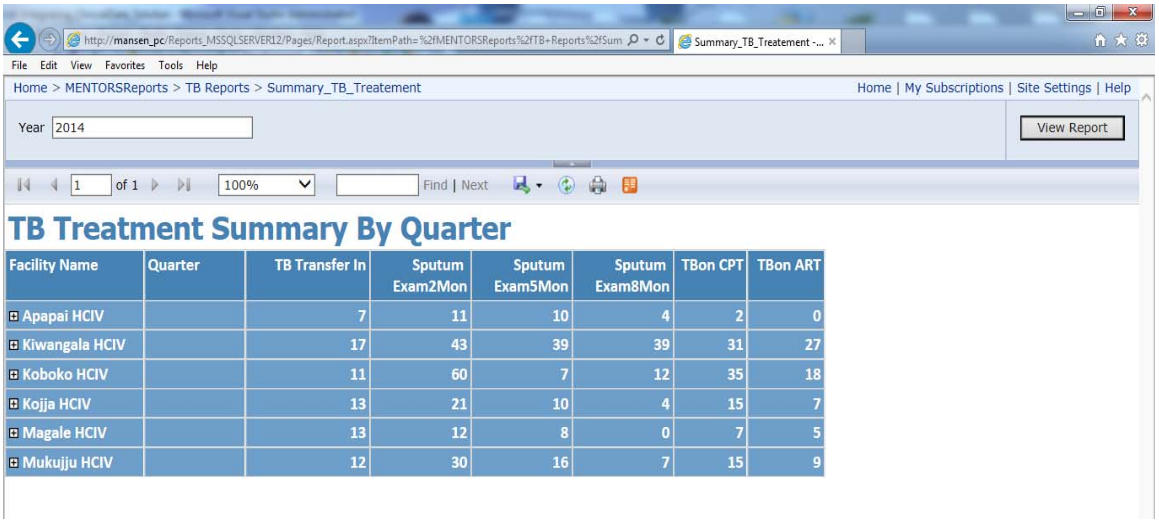
```
[To Parent Directory]
      Saturday, July 23, 2016 12:24 PM        14498 OutPatients Attendences By Agegroup
      Saturday, July 23, 2016 12:23 PM        36197 Outpatients_Disease_Incidences
      Wednesday, July 06, 2016 5:46 AM        35943 OutPatientsAttendencesByMonthByYear
```

Microsoft SQL Server Reporting Services Version 11.0.2100.60

Figure 6.28: Report in the OutPatients Reports Subfolder

Some of the sample reports generated using SSRS are shown in figure 6.29, figure 6.30 and figure 6.31. These reports can be used as a basis for generating a more complete set of other reports in the future. The reports are described below;

Figure 6.29 gives summary information about TB treatment by site by quarter, more especially sputum examination follow-up and TB/HIV co-infected started on co-trimoxazole (CPT) and ARV treatment (ART).



http://mansen_pc/Reports_MSSQLSERVER12/Pages/Report.aspx?ItemPath=%2fMENTORSReports%2fTB+Reports%2fSum
Summary_TB_Treatement -...
File Edit View Favorites Tools Help
Home > MENTORSReports > TB Reports > Summary_TB_Treatement
Home | My Subscriptions | Site Settings | Help
View Report
of 1
100%
Find | Next

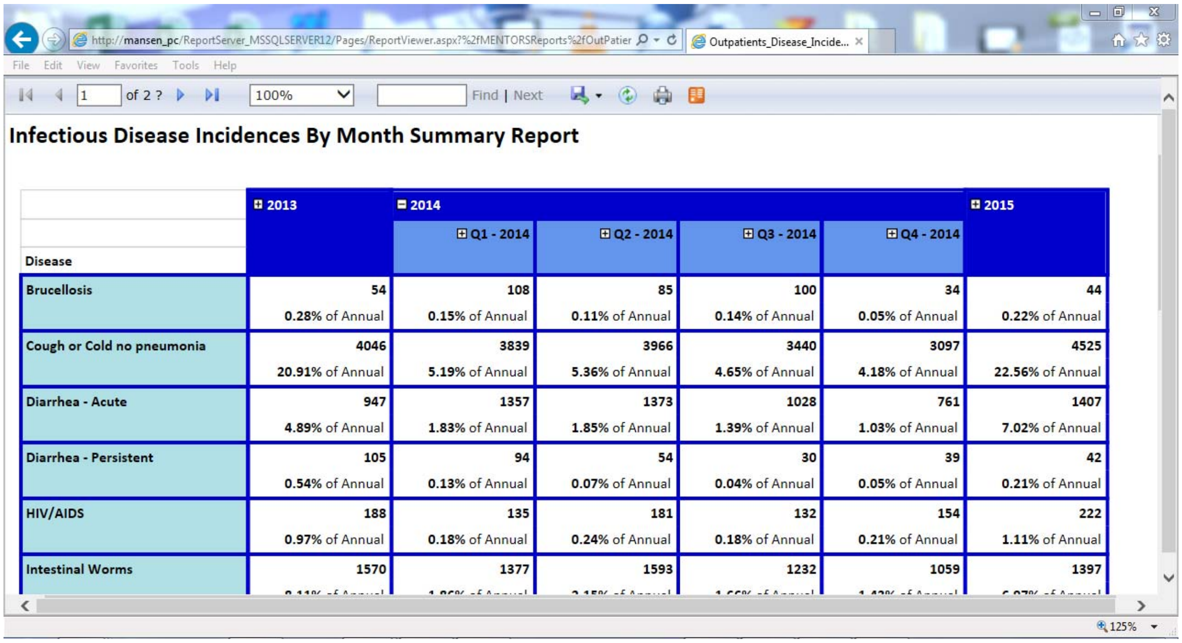
Year 2014

TB Treatment Summary By Quarter

Facility Name	Quarter	TB Transfer In	Sputum Exam2Mon	Sputum Exam5Mon	Sputum Exam8Mon	TBon CPT	TBon ART
⊞ Apapai HCIV		7	11	10	4	2	0
⊞ Kiwangala HCIV		17	43	39	39	31	27
⊞ Koboko HCIV		11	60	7	12	35	18
⊞ Kojja HCIV		13	21	10	4	15	7
⊞ Magale HCIV		13	12	8	0	7	5
⊞ Mukujju HCIV		12	30	16	7	15	9

Figure 6.29: TB Treatment Summary by Site by Quarter

Figure 6.30 represents a drill down summary report of infectious disease incidences by month. With this report users can begin analysis at the year level and drill to the quarter and month level, enabling the users analyse disease occurrences at the different times of the year.



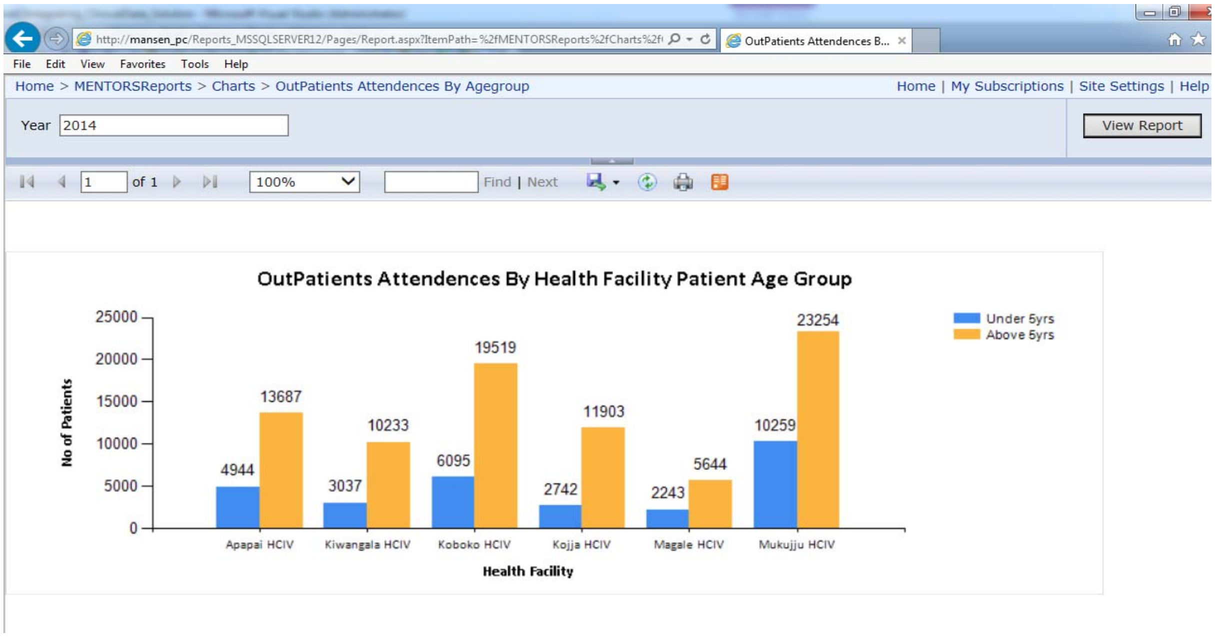
http://mansen_pc/ReportServer_MSSQLSERVER12/Pages/ReportViewer.aspx?%2fMENTORSReports%2fOutPatier
Outpatients_Disease_Incide...
File Edit View Favorites Tools Help
of 2 ?
100%
Find | Next
125%

Infectious Disease Incidences By Month Summary Report

	⊞ 2013	⊟ 2014				⊞ 2015
		⊞ Q1 - 2014	⊞ Q2 - 2014	⊞ Q3 - 2014	⊞ Q4 - 2014	
Disease						
Brucellosis	54 0.28% of Annual	108 0.15% of Annual	85 0.11% of Annual	100 0.14% of Annual	34 0.05% of Annual	44 0.22% of Annual
Cough or Cold no pneumonia	4046 20.91% of Annual	3839 5.19% of Annual	3966 5.36% of Annual	3440 4.65% of Annual	3097 4.18% of Annual	4525 22.56% of Annual
Diarrhea - Acute	947 4.89% of Annual	1357 1.83% of Annual	1373 1.85% of Annual	1028 1.39% of Annual	761 1.03% of Annual	1407 7.02% of Annual
Diarrhea - Persistent	105 0.54% of Annual	94 0.13% of Annual	54 0.07% of Annual	30 0.04% of Annual	39 0.05% of Annual	42 0.21% of Annual
HIV/AIDS	188 0.97% of Annual	135 0.18% of Annual	181 0.24% of Annual	132 0.18% of Annual	154 0.21% of Annual	222 1.11% of Annual
Intestinal Worms	1570	1377	1593	1232	1059	1397

Figure 6.30: Infectious Disease Incidences by Month Summary Report

Figure 6.31 represents a chart report displaying outpatients attendances by health facility by age group. This graph shows patient utilization of the facility outpatients section by age group



Year 2014

http://mansen_pc/Reports_MSSQLSERVER12/Pages/Report.aspx?ItemPath=%2fMENTORSReports%2fCharts%2fi
OutPatients Attendences B...
File Edit View Favorites Tools Help
Home > MENTORSReports > Charts > OutPatients Attendences By Agegroup
Home | My Subscriptions | Site Settings | Help
View Report
1 of 1
100%
Find | Next
OutPatients Attendences By Health Facility Patient Age Group
No of Patients
25000
20000
15000
10000
5000
0
4944
13687
3037
10233
6095
19519
2742
11903
2243
5644
10259
23254
Apapai HCIV
Kiwangala HCIV
Koboko HCIV
Kojja HCIV
Magale HCIV
Mukujju HCIV
Health Facility
Under 5yrs
Above 5yrs

Figure 6.31: Outpatients Attendances by Health Facility by Age group

6.6 System Testing

The objective of testing was to find out whether the system satisfies user specifications and hardware requirements. Iterative testing was used right from extracting and transforming the raw data to loading it into the data mart.

6.7 System Validation

Does the system meet the aims and objectives for which it was designed? The systems capability to automate the extraction of clinical data from multiple sources and centralize it into a single data repository (data mart) was evidenced by the availability of the data extracted from all the source systems being found in the data mart. A few reports generated from the data mart were used as a good illustration of how the consolidated system could readily avail reports for supporting decision making, very easily and quickly.

The system was made available to end users who run the reports from the data mart and compared the results against the clinical data in the various clinical data source systems. Reports were also run from different clinical data source systems and compared against the single data repository (data mart). This allowed for the accuracy and availability of the reports generated from the data mart to be verified.

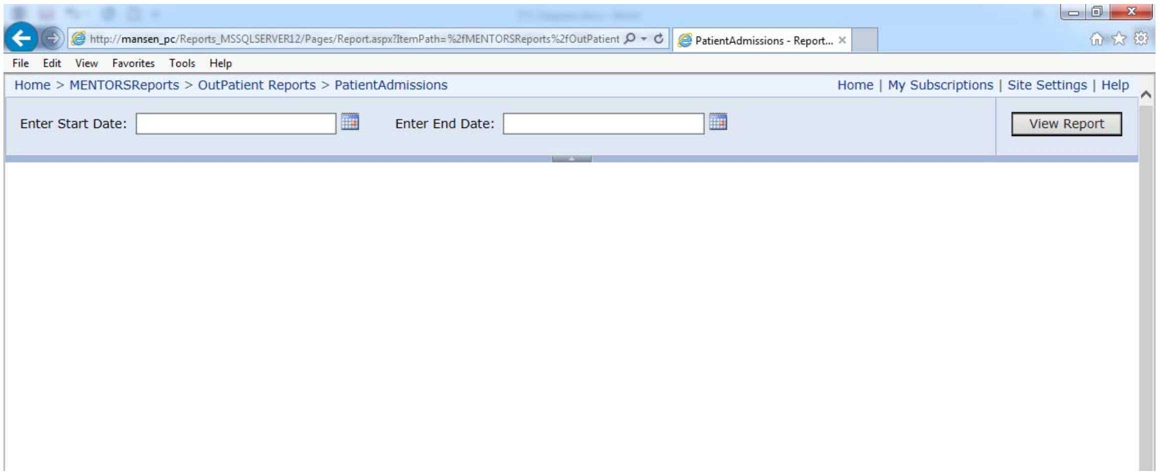
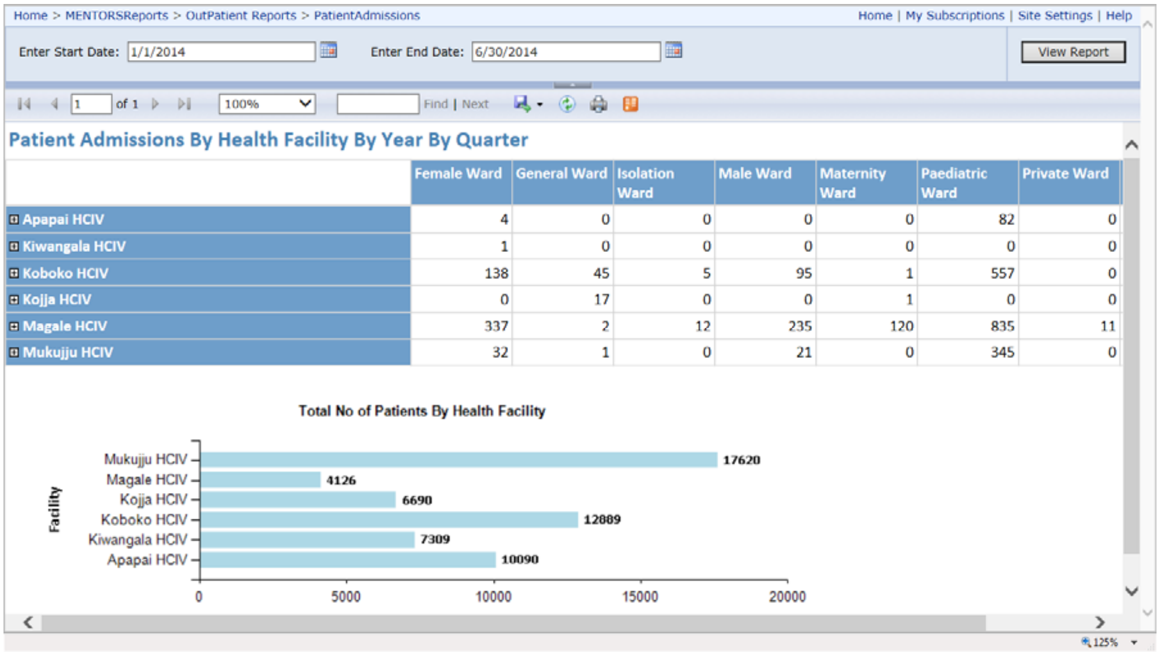

http://mansen_pc/Reports_MSSQLSERVER12/Pages/Report.aspx?ItemPath=%2fMENTORSReports%2fOutPatient
PatientAdmissions - Report...
File Edit View Favorites Tools Help
Home > MENTORSReports > OutPatient Reports > PatientAdmissions
Home | My Subscriptions | Site Settings | Help
Enter Start Date:
Enter End Date:
View Report

Figure 6.32: Sample report executed from the web browser

Figure 6.32 shows an interface of the report to be executed from the web browser. Data parameters are provided for the user to be able to input different date values as may be dictated by the reporting requirements. Once satisfied with the values, a user can then press on the view report button to get the output as shown in figure 6.33 below.



Home > MENTORSReports > OutPatient Reports > PatientAdmissions

Enter Start Date: 1/1/2014

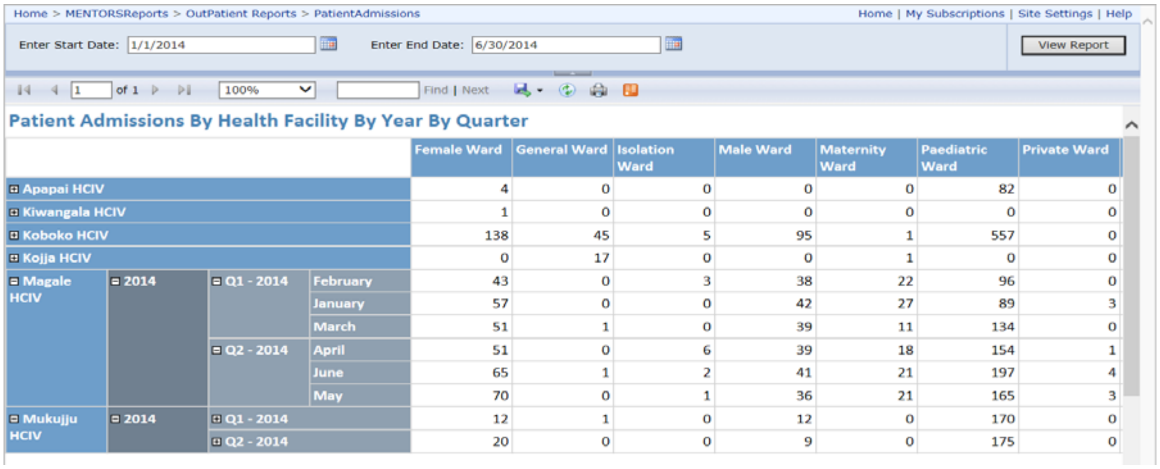
Enter End Date: 6/30/2014

Patient Admissions By Health Facility By Year By Quarter

	Female Ward	General Ward	Isolation Ward	Male Ward	Maternity Ward	Paediatric Ward	Private Ward
⊞ Apapai HCIV	4	0	0	0	0	82	0
⊞ Kiwangala HCIV	1	0	0	0	0	0	0
⊞ Koboko HCIV	138	45	5	95	1	557	0
⊞ Kojja HCIV	0	17	0	0	1	0	0
⊞ Magale HCIV	337	2	12	235	120	835	11
⊞ Mukujju HCIV	32	1	0	21	0	345	0

Figure 6.34: Report run from the browser showing Patient Admission by Facility

The report in figure 6.34 above gives a tabular and graphical presentation of the data in the data mart for ease of interpretation by the users.



Home > MENTORSReports > OutPatient Reports > PatientAdmissions

Enter Start Date: 1/1/2014

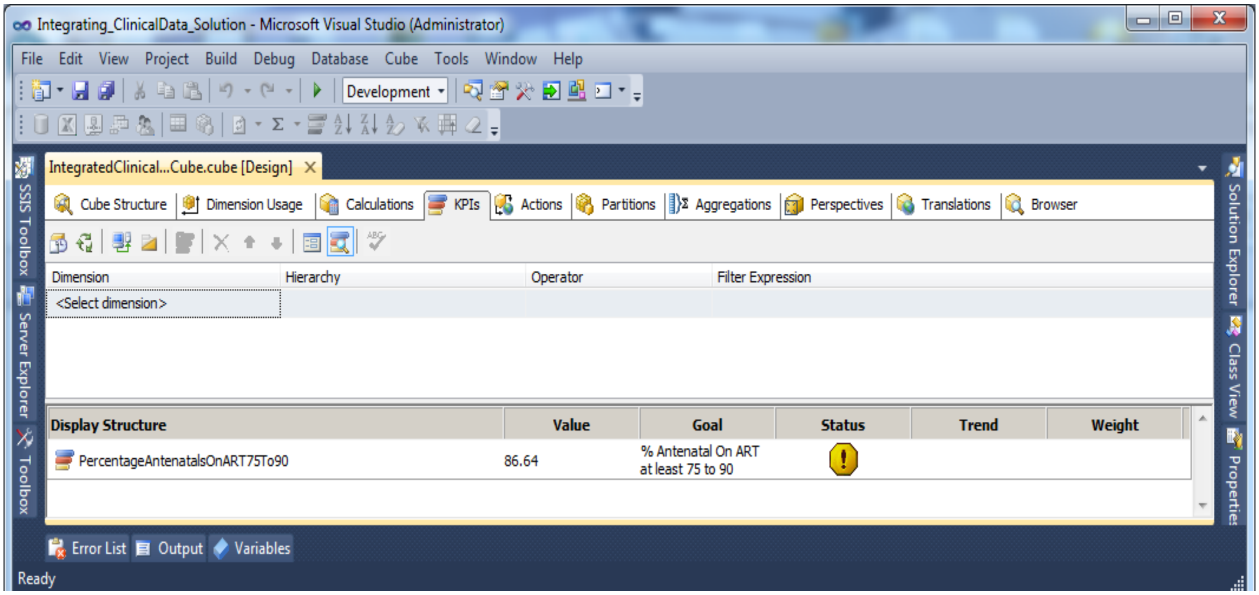
Enter End Date: 6/30/2014

Patient Admissions By Health Facility By Year By Quarter

				Female Ward	General Ward	Isolation Ward	Male Ward	Maternity Ward	Paediatric Ward	Private Ward
⊞ Apapai HCIV				4	0	0	0	0	82	0
⊞ Kiwangala HCIV				1	0	0	0	0	0	0
⊞ Koboko HCIV				138	45	5	95	1	557	0
⊞ Kojja HCIV				0	17	0	0	1	0	0
⊟ Magale HCIV	⊟ 2014	⊟ Q1 - 2014	February	43	0	3	38	22	96	0
			January	57	0	0	42	27	89	3
			March	51	1	0	39	11	134	0
		⊟ Q2 - 2014	April	51	0	6	39	18	154	1
			June	65	1	2	41	21	197	4
			May	70	0	1	36	21	165	3
⊟ Mukujju HCIV	⊟ 2014	⊞ Q1 - 2014		12	1	0	12	0	170	0
		⊞ Q2 - 2014		20	0	0	9	0	175	0

Figure 6.35: Report showing Patient Admission by Facility with drill down capability

The report in figure 6.35, shows patient admissions by different health facilities drilled down by year by quarter and month. The report has a drill down capability encouraging the user to interact more with the report by visualizing data from summary to its detailed format.

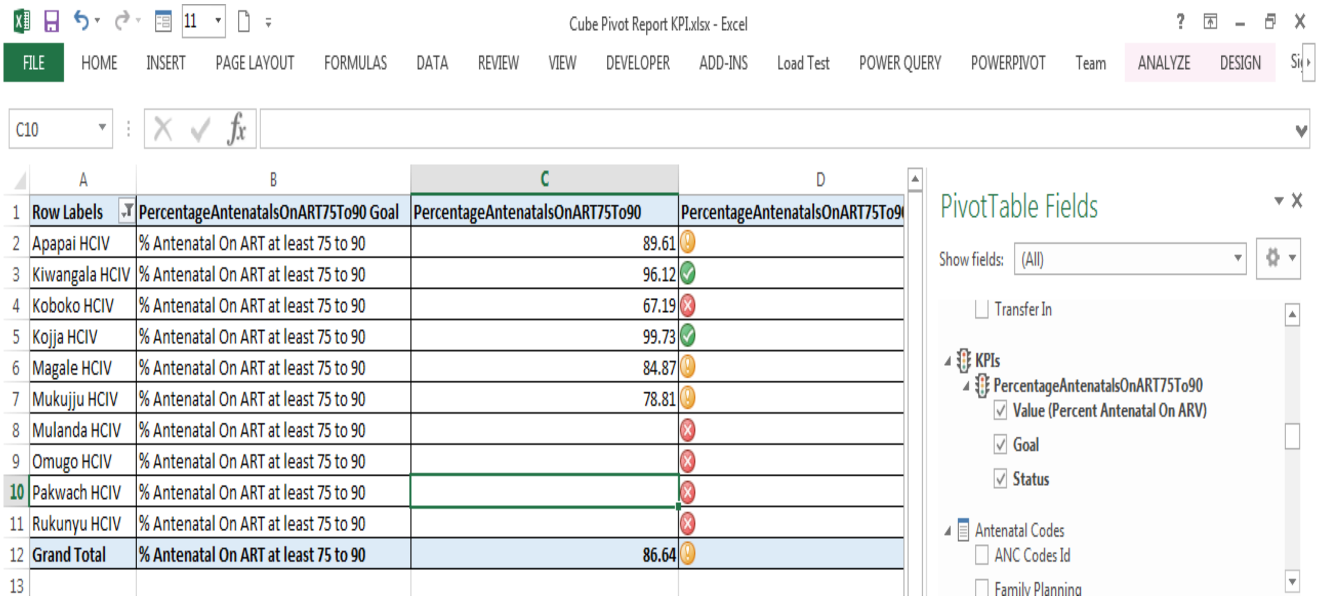


Dimension	Hierarchy	Operator	Filter Expression
<Select dimension>			

Display Structure	Value	Goal	Status	Trend	Weight
PercentageAntenatalsOnART75To90	86.64	% Antenatal On ART at least 75 to 90	!		

Figure 6.36: Report showing KPIs in SSAS

The KPI report in figure 6.36 shows the status of the percentage of HIV antenatal patients on antiretroviral treatment (ART) displayed in SSRS and Ms Excel. The KPI report with visual elements alert the users to the deviations from the expected results. This a great tool to support top level managers in decision making whereby they can be able to visualize the progress towards a defined organisation objective without delving into the details of the data.



	Row Labels	PercentageAntenatalsOnART75To90 Goal	PercentageAntenatalsOnART75To90	PercentageAntenatalsOnART75To9
2	Apapai HCIV	% Antenatal On ART at least 75 to 90	89.61	!
3	Kiwangala HCIV	% Antenatal On ART at least 75 to 90	96.12	✓
4	Koboko HCIV	% Antenatal On ART at least 75 to 90	67.19	✗
5	Kojja HCIV	% Antenatal On ART at least 75 to 90	99.73	✓
6	Magale HCIV	% Antenatal On ART at least 75 to 90	84.87	!
7	Mukujju HCIV	% Antenatal On ART at least 75 to 90	78.81	!
8	Mulanda HCIV	% Antenatal On ART at least 75 to 90		✗
9	Omugo HCIV	% Antenatal On ART at least 75 to 90		✗
10	Pakwach HCIV	% Antenatal On ART at least 75 to 90		✗
11	Rukunyu HCIV	% Antenatal On ART at least 75 to 90		✗
12	**Grand Total**	**% Antenatal On ART at least 75 to 90**	**86.64**	!
13				

PivotTable Fields

Show fields: (All)

- □ Transfer In
- KPIs
  - PercentageAntenatalsOnART75To90
    - ☑ Value (Percent Antenatal On ARV)
    - ☑ Goal
    - ☑ Status
- Antenatal Codes
  - □ ANC Codes Id
  - □ Family Planning

Figure 6.37: KPI results displayed in Ms Excel

The report in figure 6.36 shows only a single report value at a time. However exporting this report from SSAS to Microsoft Excel allows the end users to view all the KPI values of all the health facilities at once as shown in figure 6.37.

6.8 Data Mart Prototype Evaluation

At the end of the design it was deemed necessary to evaluate the functionality and success of the data mart system by addressing questions as to whether the developed system worked technically as designed from the end user's perspective and whether it produced the desired results. The evaluation was done using a Likert scale attitude statement as shown in table 6.5, with additional interview with key project staff (unit managers, data manager and clinicians), to obtain their perception towards the developed data mart system. The general perception of measuring the success of the data mart was such that the end users notice the ease of use offered by the data mart system and are happy to use the data mart system in supporting their decision making process, including also the research process. Furthermore the evaluation was to also prove the ease of accessing the information stored in the data mart system, in addition to the quality of the data stored in the system. In order to achieve the goal of evaluating the data mart system, five project staff were chosen from the different units of the project and the result are illustrated in table 6.5

Table 6.5: Data Mart Evaluation

S/N	Factors/Bench Marks	SD	D	D or A	A	SA
1	The Data Mart system implemented is a success				3	2
2	The Data Mart system does not satisfy my requirements		3		2	
3	The Data Mart system is easy to use				5	
4	It is easy to retrieve data from the system and understand such data.				2	3
5	Using the system make my job easier for me.				4	1
6	The report/output in line with the business requirements				3	2
7	You have access to timely information when you need it				5	
8	The system provide sufficient information			1	4	
9	You are satisfied with the system accuracy				4	1

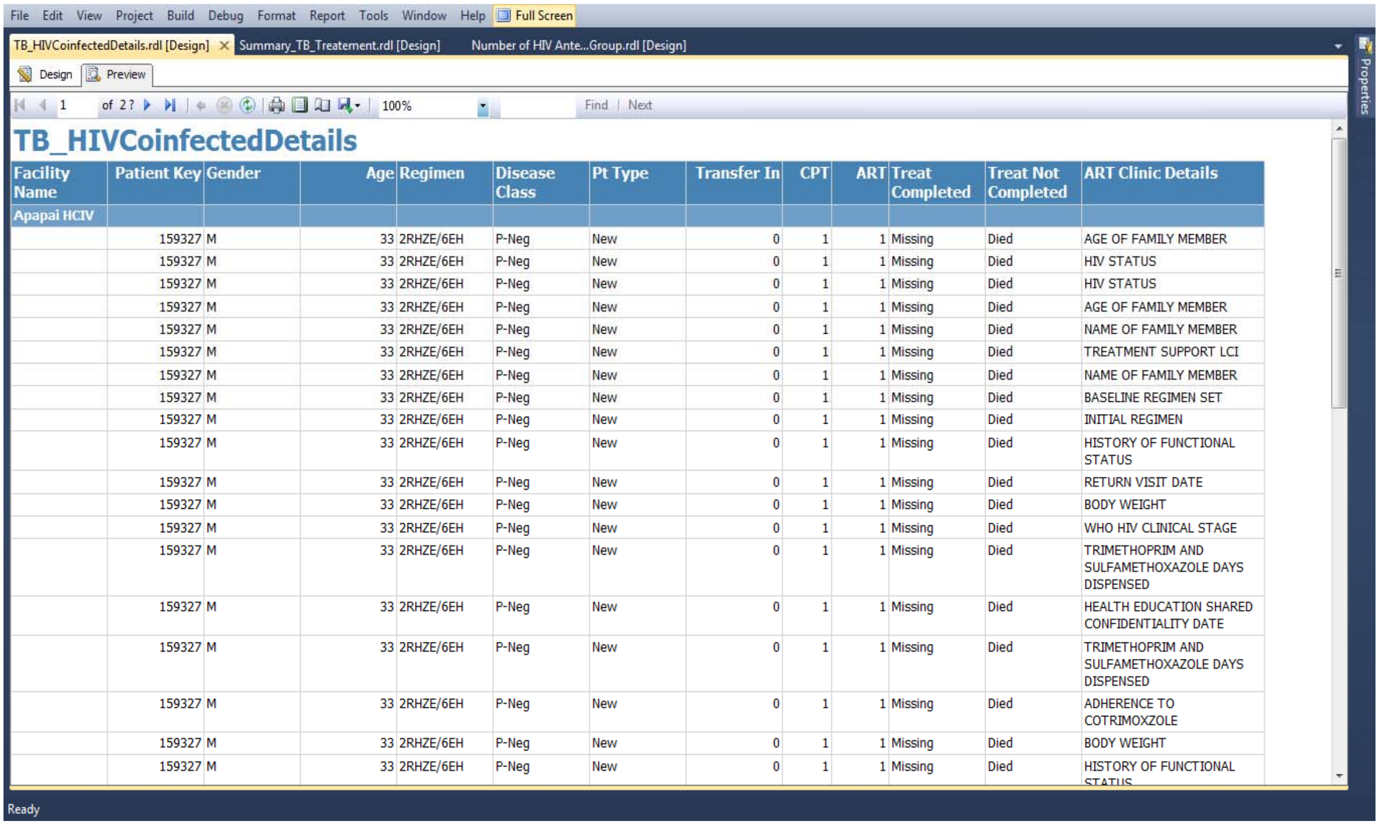
SD- Strongly Disagree, D-Disagree, D or A-Disagree or Agree, A-Agree and SA – Strongly Agree

The evaluation results in table 6.5 above show that the majority of the participants in the evaluation process revealed positive results for most of the evaluation questions, hence this shows that the data mart prototype has been fairly appreciated by the end users.

As earlier mentioned interview were also conducted to further evaluate the data mart prototype and feedback was collected from the data manager, the unit managers, and clinicians via short structured interview based question “How does the developed data mart system improve availability of clinical data for analysis, enhance data quality and also support the decision making and research process in the MENTORS project?” According to the data manager’s view point, “Implementation of the data mart will greatly reduce the complexity and time taken to prepare the gathered clinical data from the various project supported health facilities for analysis purposes, given that the integration process would now be automated. “The developed data mart provides a single data repository from which it is easy to construct queries for exporting clinical data to other statistical softwares like STATA and SPSS for more advanced statistical analysis to support research outcomes”. Also “The data mart system provides a simplified mechanism for systematically processing data errors which are logged and later used to improve the quality of data in the source OLPT systems. The data manager also raised some issues anticipated with the actual implementation of the data mart system which included; access rights and confidentiality of the clinical data in the data mart system.

According to the monitoring and evaluation manager’s (unit manager) view “information generated from the data mart system would be very valuable by enabling timely acquisition of data to measure the project progress towards attaining its objectives. The data mart would also enable the unit managers make informed decisions regarding the effectiveness of the different intervention activities carried out in the supported project health facilities. Citing an example given by the unit manager is the KPI report in figure 6.37 which would be very useful in determining how the project implementation vary from site to site in regard to improving antenatal care for HIV positive pregnant mothers. The unit manager also mentioned that in comparison to the previous process of acquiring data for decision making, the excel pivot table report in figure 6.25 extracted from the OLAP cube made data readily available for filtering according to the different clinical dimensions held in the data mart.

The feedback from the clinicians from the MENTORS project agreed that developing the data mart system would be very valuable for the clinician. Furthermore they mentioned that the output in figure 6.38 which shows treatment records of patients on both HIV and TB treatment would enable the clinicians to have a holistic view of patient records on both TB and HIV treatment in order to improve patient care for TB/HIV co-infected patients across the different supported project sites.


File Edit View Project Build Debug Format Report Tools Window Help Full Screen
TB_HIVCoinfectedDetails.rdl [Design]
Summary_TB_Treatement.rdl [Design]
Number of HIV Ante...Group.rdl [Design]
Properties
Design
Preview
1 of 2 ?
100%
Find | Next
Ready

TB_HIVCoinfectedDetails

Facility Name	Patient Key	Gender	Age	Regimen	Disease Class	Pt Type	Transfer In	CPT	ART	Treat Completed	Treat Not Completed	ART Clinic Details
Apapai HCIV												
	159327	M	33	2RHZE/6EH	P-Neg	New	0	1	1	Missing	Died	AGE OF FAMILY MEMBER
	159327	M	33	2RHZE/6EH	P-Neg	New	0	1	1	Missing	Died	HIV STATUS
	159327	M	33	2RHZE/6EH	P-Neg	New	0	1	1	Missing	Died	HIV STATUS
	159327	M	33	2RHZE/6EH	P-Neg	New	0	1	1	Missing	Died	AGE OF FAMILY MEMBER
	159327	M	33	2RHZE/6EH	P-Neg	New	0	1	1	Missing	Died	NAME OF FAMILY MEMBER
	159327	M	33	2RHZE/6EH	P-Neg	New	0	1	1	Missing	Died	TREATMENT SUPPORT LCI
	159327	M	33	2RHZE/6EH	P-Neg	New	0	1	1	Missing	Died	NAME OF FAMILY MEMBER
	159327	M	33	2RHZE/6EH	P-Neg	New	0	1	1	Missing	Died	BASELINE REGIMEN SET
	159327	M	33	2RHZE/6EH	P-Neg	New	0	1	1	Missing	Died	INITIAL REGIMEN
	159327	M	33	2RHZE/6EH	P-Neg	New	0	1	1	Missing	Died	HISTORY OF FUNCTIONAL STATUS
	159327	M	33	2RHZE/6EH	P-Neg	New	0	1	1	Missing	Died	RETURN VISIT DATE
	159327	M	33	2RHZE/6EH	P-Neg	New	0	1	1	Missing	Died	BODY WEIGHT
	159327	M	33	2RHZE/6EH	P-Neg	New	0	1	1	Missing	Died	WHO HIV CLINICAL STAGE
	159327	M	33	2RHZE/6EH	P-Neg	New	0	1	1	Missing	Died	TRIMETHOPRIM AND SULFAMETHOXAZOLE DAYS DISPENSED
	159327	M	33	2RHZE/6EH	P-Neg	New	0	1	1	Missing	Died	HEALTH EDUCATION SHARED CONFIDENTIALITY DATE
	159327	M	33	2RHZE/6EH	P-Neg	New	0	1	1	Missing	Died	TRIMETHOPRIM AND SULFAMETHOXAZOLE DAYS DISPENSED
	159327	M	33	2RHZE/6EH	P-Neg	New	0	1	1	Missing	Died	ADHERENCE TO COTRIMOXZOLE
	159327	M	33	2RHZE/6EH	P-Neg	New	0	1	1	Missing	Died	BODY WEIGHT
	159327	M	33	2RHZE/6EH	P-Neg	New	0	1	1	Missing	Died	HISTORY OF FUNCTIONAL STATUS

Figure 6.38: Treatment record of TB/HIV Co-infected Patient

Figure 6.38 show treatment records of TB/HIV co-infected patient which a combination of data coming from two different databases. (TB specific data extracted from the an Access database and HIV specific data extracted from OpenMRS HIV database built on MySQL database platform)

CHAPTER SEVEN

DISCUSION, CONCLUSION AND RECOMMENDATIONS

7.1 Introduction

This chapter concludes what was achieved by the project, what wasn't achieved, an evaluation of the project, a summary of the project and recommendations. The chapter discusses the issues experienced and addressed while planning, designing and developing the data mart system for the MENTORS project at IDI, and sums up with future improvements on the data mart system

7.2 Discussion

Data integration is critical for healthcare organisations that wish to improve the quality of healthcare service delivery, but have huge volumes of healthcare data stored in many different data sources. The concept of data warehousing is deemed as the most appropriate solution for integrating and accessing data from multiple data sources. The implementation of the data mart system to consolidate data from the multiple data sources gathered by the MENTORS project of IDI was a great success. For the first time project staff were able to readily access information they required to improve the quality of healthcare service delivery at the various supported health facilities of the MENTORS project. The research also provided a demonstration of the development and implementation of an otherwise costly project in a low resource setting. The developed data mart could be used to solve the data integration challenges in organisations/projects with the similar settings.

Given the poor quality of data stored in the various clinical data sources a lot of effort was devoted to clean the source systems before extracting data from them. Hence emphasis of improving the quality of data in the source systems needs to be highlighted. Accessing clinical data was a challenge given the confidentiality attached to patient information.

7.3 Conclusion

The main objective of the research was to build a consolidated view of clinical data from multiple data sources, using a data mart, to ease analysis of data, improve the availability and quality of information required for decision making and clinical research within the MENTORS project of IDI. The findings show that through a data mart implementation, information availability for decision making and clinical research within the MENTORS project of IDI is greatly improved. In addition, given that the data mart system can be used to support in-depth data analysis, efficient reporting and querying of information, one cannot underestimate its importance in supporting the MENTORS project to achieve its objectives. Its implementation provides evidence that centralized data storage, information retrieval and reporting in the MENTORS project is possible and attainable and this can be extended to other health organization/projects with similar challenges.

7.4 Recommendations

Healthcare organization with large volumes of healthcare data in multiple data sources should adopt the data warehousing technology as a way of improving availability of critical information needed by healthcare providers for their decision making. The infectious Disease Institute Ltd (IDI) should use the data mart approach as a starting point to scale-up to a fully-fledged clinical data warehouse storing clinical data from the other projects that will support the entire organization in decision making and research undertakings.

REFERENCES

Abello, A., Samos, F and Saltor, F., 2002. On Relationships Offering New Drill Across Possibilities: International Workshop on Data Warehousing and OLAP, ACM.

Ado. A., Aliyu, A., Bello, S.A., Garba. A Sharifai, G. A and Gezawa, A. S., 2014. Building a Diabetes Data Warehouse to Support Decision Making in Healthcare Industry. Journal of Computer Engineering: 16(2), 138-143.

Ariyachandra, T. and Watson, H. 2005. Data warehouse architectures: factors in the selection, decision and success of the architectures. [Online] Available at: http://www.terry.uga.edu/~hwatson/DW_Architecture_Report.pdf [Accessed 20 Feb 2015].

Ariyachandra, T. and Watson, H. 2010. Key organizational factors in data warehouse architecture selection. Decision Support Systems. [Online] Available at: https://cours.etsmtl.ca/mti820/public_docs/lectures/keyorganizationalfactorsindwarchitectureselection.pdf [Accessed 20 Feb 2015].

Arnrich. B., Walter. J., Alexander Albert, A., Ennker, J. and Ritter, H., 2004. Data Mart Based Research in Heart Surgery: Challenges and Benefits. Proc Int Conf MedInfo, San Francisco CA, AMIA [online] Available at: http://arxiv.org/ftp/arxiv/papers/0812/0812.2874.pdf [Accessed 15 June 2015].

Ballard, C., Davies, N., Gavazzi, M., Lurie, M. and Stephani, J., 2003. IBM Informix: Integration Through Data Federation, IBM, Redbooks. [Online] Available at: http://www.iiug.org/library/ids/technical/sg247032.pdf [Accessed 23 Apr 2015].

Blankenship, K., 2013. Clinical Data Integration: From Challenges to Opportunities. Lumeris, Inc. [online] Available at: http://www.healthcareitnews.com/sites/default/files/resource-media/pdf/lumeris_clinical_data_integration_whp.cdi_.08-13.v2_single_page.pdf

Bobak, A.R., 2012. Connecting the data, Westfield, NJ 07090 U.S.A.: Technics Publications.

Bonafati, et al., 2001. Designing Data Marts for Data Warehouses. ACM Transactions on Software Engineering and Methodology: 10(4).

Boterenbrood, F., Krediet, I. and Goossen, W., 2014. Building a high quality medical data architecture for multiple uses in an integrated health care environment. Journal of Hospital Administration, 3(5). [Online] Available at: www.sciedu.ca/journal/index.php/jha/article/download/3904/2672 [Accessed 22 Mar 2015].

Branson. A., Tamas Hauer. T., McClatchey, R., Rogulin. D and Shamdasani. J., 2008. A Data Model for Integating Heterogenous Medical Data in the Health-e-Project: CCS Research Centre, University of the West of England, Bristol, UK. [Online] Available at: http://arxiv.org/ftp/arxiv/papers/0812/0812.2874.pdf [Accessed 23 Apr 2015].

Cali, A., Lembo, D. and Rosati, R., 2005. A Comprehensive Semantic Framework for Data Integration Systems. Journal of Applied Logic. 3(2): 308–328. [Online] Available at: http://www.sciencedirect.com/science/article/pii/S157086830400062X [Accessed 22 Jul 2015].

Chenhui, Z., Huilong, D. and Xudong, Lu. 2008. An integration approach of healthcare information system. In: In: S. Nuwagela, 2013. Data Warehousing Model for Integrating Fragmented Electronic Health Records From Disparate and Heterogenous Clinical Data Stores. [Online] Available at: http://eprints.qut.edu.au/60880/1/Saliya_Nugawela_Thesis.pdf [Accessed 22 Feb 2015].

Chuck Ballard. C., Hamid, A. A., Frankus, R., Hasegawa. F., Larrechart, J.,Pietro Leo, P. and Ramos, J, 2006. Improving Business Performance Insight with Business Intelligence and Business Process Management, IBM, Redbooks.

Colesca, S. E and Dobrica, R. L., 2009. Information Management in Healthcare Organization: The Ninth International Conference, “Investment and Economic Recovery. Economia Seria Management, 12(1) [Online] Available at:

http://www.management.ase.ro/reveconomia/2009-1s/22.pdf [Accessed 20 May 2015].

Connolly, T. M. and Begg, C. E., 2005. Database Systems: A Practical Approach to Design, Implementaion, and Management. Fourth Edition. Addison Wesley.

Daft, R.L. (1978). A Dual-Core Model of Organizational Innovation. Academy of Management Journal, 21(2), 193-210. In: A.Y. Akbulut. 2003. An Investigation of the Factors that Influence Electronic Information Sharing Between State and Local Agencies. Ph.D. Louisiana State University. [Online] Available at: http://etd.lsu.edu/docs/available/etd-0619103-214616/unrestricted/Akbulut_dis.pdf [Accessed 28 Jul 2016].

Doan , A., Halevy, A. and Ives Z., 2012. Principles of Data Integration, Waltham, MA, USA: Elsevier.

Evans, C., 2013. Data Warehousing in Health Care [online] Available at: http://www.christopherevans.org/wp-content/uploads/2013/02/Data-Warehousing-in-Health-Care.pdf

Gray, P., G. and Watson, H., J., 1998. Decision Support in the Data Warehouse, Prentice-Hall, New Jersey.

Gonzales, M., L and Bagchi, K., 2011. Diffusion of Business Intelligence and Data Warehousing: An Exploratory Investigation of Research and Practice. Proceedings of the 44th Hawaii International Conference on System Sciences. [Online] Available at: https://www.computer.org/csdl/proceedings/hicss/2011/4282/00/09-01-04.pdf [Accessed 28 Jul 2016].

HIMSS, 2013. Clinical and Business Analytics: Data Management - A Foundation for Analytics. [Online] Available at:

http://www.himss.org/files/himssorg/content/files/201304_data_integration_final.pdf [Accessed 22 Feb 2015].

Hoffer, J. A., Ramesh, V., and Topi, H., 2011. Modern database management. 10th ed: Prentice Hall

Hofmann, F. and Lehner, H.F., 2001. Requirements engineering as a success factor in software projects Software, IEEE, 18(4). [Online] Available at:

http://www.ics.uci.edu/~wscacchi/Software-Process/Readings/Req-Engr-SuccessFactors-Software-July01.pdf [Accessed 15 May 2015].

HSCC Clinical Data Warehouse, 2013. The South Carolina Biomedical Research Tool of the Future. Health Sciences South Carolina.
Available at: http://www.healthsciencessc.org/upload/index_164_1262775643.pdf

IL-Yeol, S., 2009. Data Warehousing Systems: Foundations and Architectures. In: L. Liu and M.T Özsu, ed. 2009. Encyclopedia of Database Systems. Springer US.

Inmon, W. H., 1993. Building the Data Warehouse In: T.Connoly and C. Begg, ed. 2005. Database Systems: A practical Approach to Design, Implementation, and Management. Pearson Education Limited. International Journal of Engineering and Computer Science: Volume 3, issue 10, October 2014 Page No. 8510-8518

Inmon, W. H., 2005. Building the Data Warehouse. Fourth Edition. Wiley. Indianapolis.

Inmon, W. H., 2007. Data Warehousing in the HealthCare Environment, Inmon Data Systems, Available at: www.inmondatasystems.com. [Accessed 27 Mar 2015].

Kerkri, E. M., et al., 2001. An approach for integrating heterogeneous information sources in a medical data warehouse. Journal of Medical Systems, 25(3), 167-176. doi: 10.1023/A:1010728915998

Kimball, R. and Ross, M., 2002. The Data Warehouse Toolkit. Second Edition. Wiley Computer Publishing. New York.

Kimball, R. and Ross, M., 2013. The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling. Third Edition. John Wiley & Sons, Inc., Indianapolis, Indiana.

Koeller, P., 2006. Integration of Data Sources Through Data Mining. In: J. Wang, ed. 2006. Encyclopedia of Data Warehousing and Mining. Idea Group Inc.

Koh, H.C and Tan, G., 2005. Data Mining Applications in Healthcare. Journal of Healthcare Information Management. 19(2) [Online] Available at: https://www.himss.org/files/HIMSSorg/content/files/jhim/19-2/datamining.pdf [Accessed 22 Feb 2015].

Kontio, J., 2005. Data Warehousing Solutions for Reporting Problems In: J. Wang, ed. 2006. Encyclopedia of Data Warehousing and Mining. Idea Group Inc.

Kossi, K., et al 2009. Comparing Strategies to Integrate Health Information Systems Following a Data Warehouse Approach in Four Countries. Proceeding of the 10th International Conference on Social Implicationof Computers in Developing Countries, Dubai, UAE. Available at: http://www.uio.no/studier/emner/matnat/ifi/INF3290/h10/undervisningsmateriale/ComparingStrategiesForHISintegration.pdf [Accessed 25 Apr 2015].

Laudon. C. K. and Laudon, J. P., 2012. Management Information Systems : Managing the Digital Firm. Twelfth Edition. Pearson Education, Inc.

Leitheiser, R. L., 2001. Data Quality in Health Care Data Warehouse Environments. Proceedings of the 34th Hawaii International Conference on System Sciences [online] Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.110.9908&rep=rep1&type=pdf .

LeSueur, D., 2014. Five Reasons HealthCare Data is Unique and Difficult to Measure. Health Catalyst. [online] Available at: https://www.healthcatalyst.com/wp-content/uploads/2014/08/5-Reasons-Healthcare-Data-Is-Unique-and-Difficult-to-Measure.pdf.

Loper. D., Klettke, M., Bruder, I. and Heuer. A., 2013. Enabling Flexible Integration of Healthcare Information Using the Entity-Attribute-Value Storage Model. BioMed Central Ltd. [Online] Available at: http://www.hissjournal.com/content/pdf/2047-2501-1-9.pdf [Accessed 22 May 2015].

Marco, D. (2000). Independent Data Marts - Part 1. The Data Administration Newsletter. [online] Available at: http://www.tdan.com/view-articles/4881 [Accessed 25 Feb 2015].

McCann. R., Doan, A., 2003. Building Data Integration Systems: A Mass Collaboration Approach, International Workshop on the Web and Databases. [online] Available at: http://www.ics.forth.gr/webdb03/webdb03-proceedings.pdf#page=30 [Accessed 30 Feb 2015].

Mussa, B., Yonah, Z. and Tarimo, C., 2014. Design and Implementation of Livestock Data Marts for a Web and Mobile-Based Decision Support System for Smallholder Livestock Keepers: Case Study of Tanzania. International Journal of Engineering and Computer Science, 3(10), 8510-8518.

Nanette, B. S., 2013. Health Information Management Technology. Twelfth Edition, American Information Management Technology.

Ogbuji, O., 2009. Clinical Data Acquisition Storage and Management. In: L. Liu and M.T Özsu, ed. 2009. Encyclopedia of Database Systems. Springer US.

ORACLE, 2011. ORACLE Healthcare Analytics Data Integration. [Online] Available at: http://www.oracle.com/us/industries/healthcare/healthcare-analytics-integration-ds-1360701.pdf [Accessed 20 Mar 2015].

Palmer, J., 2013. The clinical Data Warehouse- A New Mission – Critical Hub, Oracle Health Sciences [Online] Available at: http://www.oracle.com/us/industries/health-sciences/clinical-data-warehouse-hub-2272366.pdf [Accessed 20 Mar 2015].

Patel, C. and Weng, C., 2009. Clinical Data and Information Models. In: L. Liu and M.T Özsu, ed. 2009. Encyclopedia of Database Systems. Springer US.

Pedersen, T. B., & Jensen, C. S. (1998). Research issues in clinical data warehousing. IEEE. Available at: http://people.cs.aau.dk/~csj/Papers/Files/1998_pedersenISS.pdf. [Accessed 20 Mar 2015].

Ponniah, P., 2010. Data Warehousing Fundamentals for IT Professionals. John Wiley & Sons, Inc.

Rahm, E. and Bernstein, P., A., (2001). A survey of approaches to automatic schema matching. VLDB Journal, 10(4), 334-350.

Raghupathi. W. and Raghupathi, V., 2014. Big Data Analytics in Healthcare: Promise and Potential, BioMed Central Ltd. [online] Available at:

http://www.hissjournal.com/content/pdf/2047-2501-2-3.pdf [Accessed 25 May 2015].

Rashmi, C., Pahwa, P., 2014. Data Mart Designing and Integration Approaches, International Journal of Computer Science and Mobile Computing, 3(4), 74-79

Riazati, D., 2012. Integration of Multidimensional Data in Heterogeneous Data Marts RMIT University, Melbourne, Victoria, Australia. Ph. D. RMIT University, Australia. [Online] Available at: http://researchbank.rmit.edu.au/eserv/rmit:160370/Riazati.pdf [Accessed 20 Feb 2015].

Rivest, S., Bédard. Y., Proulx, M. J., Nadeau, M., Hubert, F. and Pasto, J., 2005. SOLAP technology: Merging business intelligence with geospatial Technology for interactive spatio-temporal exploration and analysis of data. ISPRS Journal of Photogrammetry & Remote Sensing, issue 60(1), 17–33.

Rogers, E., M., 1995. Diffusion of innovations. Fourth Edition. New York: Free Press.

Rotem-Gal-Oz, A., 2012. SOA Patterns, Shelter Island: Manning Publications.

Rusincovitch, S. A. and Shang, H. C., 2012. Conceptual Model for Research-Driven Data Marts. Duke Health Technology Solutions, Duke University Health System

Sahama, T. R and Croll, P. R., 2007. A Data Warehouse Architecture for Clinical Data Warehousing. American Computer Society. Inc. [online] Available at: http://crpit.com/confpapers/CRPITV68Sahama.pdf [Accessed 20 May 2015].

Scerbo., M., 2009. A Health Care Claims Data Mart: Construction and Exploitation. SAS Institute Inc. [Online] Available at: http://www2.sas.com/proceedings/sugi24/Dataware/p113-24.pdf

Sekaran., U and Bourgie., R., 2010. Research Methods for Business: A Skill Building Approach. John Wiley & Sons, Inc., U.K Shepard, R. J., 2002. Ethics in exercise science research. Sports Med, 32(3), 169–183. [Online] Available at: http://link.springer.com/article/10.2165%2F00007256-200232030-00002 [Accessed 20 Apr 2015].

Sheta, O. E. and Eldeen, A. N., 2013. Evaluating a Healthcare Data Warehouse For Cancer Diseases. International of Computer Science and Information Technology and Security 3(3) [online] Available at: http://arxiv.org/ftp/arxiv/papers/1307/1307.3448.pdf [Accessed 5 May 2015].

Silverman, D., (1999). Doing qualitative research: a practical handbook. London, UK: Sage Publications Ltd.

Soderlund, J., 2011. Building a Business Intelligence System with the Pentaho BI Suite. Msc. Royal Institute of Technology, Stockholm, Sweden. [Online] Available at:

http://www.nada.kth.se/utbildning/grukth/exjobb/rapportlistor/2011/rapporter11/soderlund_johan_11106.pdf [Accessed 20 May 2015].

Stolba., N and Schanner., A., 2007. eHEALTH Integrator – Clinical data integration in Lower Austria. Institute of Software and Interactive Systems, Vienna University of Technology, Austria. [Online] Available at: http://publik.tuwien.ac.at/files/pub-inf_4712.pdf. [Accessed 20 June 2015].

Stolba, N., 2007. Towards a Sustainable Data Warehouse Approach for Evidence-Based Healthcare. Ph.D. Vienna University of Technology. Austria

Teh Ying Wah, Ong Suan Sim (2009). Development of a Data Warehouse for Lymphoma Cancer Diagnosis and Treatment Decision Support 3(6). Information Science and Applications. Available at: http://www.wseas.us/e-library/transactions/information/2009/28-906.pdf.

Teodoro., D., et al, 2009. Integration of Biomedical Data Using Federated Databases, Swiss Medical Informatics. [Online] Available at: http://www.medical-informatics.ch/index.php/smiojs/article/download/209/191. [Accessed 20 May 2015].

Teste., O., 2010. Towards Conceptual Multidimensional Design in Decision Support Systems. In: Mussa, B., Yonah, Z. and Tarimo, C., 2014. Design and Implementation of Livestock Data Marts for a Web and Mobile-Based Decision Support System for Smallholder Livestock Keepers: Case Study of Tanzania. International Journal of Engineering and Computer Science, 3(10), 8510-8518.

Thong, J., Y., L., 1999. An integrated Model of Information Systems Adoption in Small Business. Journal of Management Information Systems, 15(4). 187-214. In: A.Y. Akbulut. 2003. An Investigation of the Factors that Influence Electronic Information Sharing Between State and Local Agencies. Ph.D. Louisiana State University. [Online] Available at:http://etd.lsu.edu/docs/available/etd-0619103-214616/unrestricted/Akbulut_dis.pdf [Accessed 28 Jul 2016].

Tooker. R., N., 2010. The Case for Enterprise Data Integration in HealthCare. KBM GROUP. [Online] Available at: http://www.kbmg.com/wp-content/uploads/2010/09/KBMG_WP_DataIntegration_HC.pdf [Accessed 20 Mar 2015].
Verhulst. S., 2006. Backgorund Issues on Data Quality. Connecting for Health Common Framework, Markle Foundation. [Online] Available at:
http://www.connectingforhealth.org [Accessed 20 May 2015].

Verma, R. and Harper, J., 2001. Life Cycle of a Data Warehousing Project in Healthcare. Journal of Healthcare Information Management, 15(2). [Online] Available at:
http://webdocs.cs.ualberta.ca/~zaiane/courses/cmput605/2008/pdf/Life_Cycle.pdf
[Accessed 7 June 2015].

Viangteeravat., T., et al., 2011, Clinical Data Integration of Distributed Data Sources Using Health Level Seven (HL7) v3-RIM Mapping. Journal of Clinical Bioinformatics 2011, 1(32) [Online] Available at:
http://jclinbioinformatics.biomedcentral.com/articles/10.1186/2043-9113-1-32 [Accessed 20 May 2015].

Wager, K. A., Lee, F.W and Glaser, J.P., 2005. Managing Healthcare Information Systems: A Practical Approach for healthcare Executives. John Wiley & Sons, Inc.
[Online] Available at:
http://samples.sainsburysebooks.co.uk/9780787979515_sample_418719.pdf [Accessed 26 May 2015].

White, C., 2005. Data Integration: Using ETL, EAI, and EII Tools to Create an Integrated Enterprise. [Online] Available at:
http://download.101com.com/tdwi/research_report/DIRR_Report.pdfIntegration.pdf
[Accessed 23 Apr 2015].

White, C., 2006. A Roadmap to Enterprise Data Integration. Information Integration Solutions, IBM. [Online] Available at:
ftp://public.dhe.ibm.com/software/emea/de/db2/A-Roadmap-To-Enterprise-Data-Integration.pdf [Accessed 23 Apr 2015].

Williams, P., and Gunter, B., 2005. Triangulating qualitative research and computer transaction logs in health information study. New Information Perspectives, 58(1/2), 129-139.

Umer, S., Afzal, M., Hussain, M., Ahmad, H.F and Khalid Latif, K., 2010. Towards Automatic HL7-RIM and Relational, NUST School of Electrical Engineering and Computer Science (SEECS), Pakistan [Online] Available at:
http://www.ringholm.de/persist/20100514_IHIC_mapping_RIM_db_schema.pdf
[Accessed 23 Jul 2015].

APPENDIX A: PROJECT IMPLEMENTATION PLAN

TASK	ACTIVITIES	DELIVERABLES	DURATION
PLANNING AND DATA COLLECTION	- Determining the scope - Developing data collection tools - Make appointments with interviewees	Questionnaires. Questionnaire answers. Interview results.	20 Days
SYSTEM ANALYSIS AND DESIGN	- Determining Transactions, Actors and Interactions. - Defining all functions of the proposed system. - Identifying inputs and outputs. - Determine the logical design - Database Designs	UML Diagrams (Use case, class, activity diagrams). User Interface Diagrams. Dimensional Diagrams.	30 Days
SYSTEMS DEVELOPMENT	Construction of Information System - System Databases - System Interfaces	Presentation of Functioning Information System Prototype	40 Days
SYSTEMS TESTING	- Writing, testing and documenting application programs - Unit, Integration and Systems testing	Test cases	5 Days
SYSTEM DOCUMENTATION	- Developing User Manual - Developing Technical manual	User Manual Technical Manual	30 Days

Table 3.1: List of activities in the project schedule to be followed during system implementation

APPENDIX B: INTERVIEW GUIDE

Part One – Explain to the respondent

Explain the purpose of this research to the correspondent by highlighting the area of focus.

Part Two - Correspondent's details

Name and Mobile No: ____

Date of interview______________________ Start and End time ________________

1. Which section are you associated with in this project?

2. What is your designation in this project?

Part Three - Interview Questions

Current data source & Decision-making process:

3. Which kind of data sources do you use to assist decision-making and research?

4. How do you collect or access data from the mentioned data sources to support decision making and research?

5. Identify example management problems/decisions you address or would like to address by using the data sources mentioned?

Data Sources	Problems/Decisions/Analysis I would like to address	Which routine analysis do you conduct or would like to conduct

Current Issues:

6. Are you satisfied with the support provided for decision-making processes by the current Information Systems?

7. What are the main information related problems you have identified in the decision-making process supporting clinical service management in your area?

8. What are the main data quality issues impacting the trust in clinical data used for the decision-making processes in your area?

Data Storage/ data analysis:

9. What operating system does the system run on?

10. Do these various data sources store sufficient data fields for your decision-making processes?

11. According to your knowledge, how long is data kept in the data sources?

12. What kind of reports are generated and how?

13. What difficulties are experienced when generating reports?

14. According to your knowledge, what analysis tools do you use to analyze the clinical data?

15. Do you have any concerns regarding data security and information privacy that should be incorporated in the integrated clinical data system development?

APPENDIX C: DATA MART TABLES

FactPatientDiagnosis

Name	DataType	Constraint	Nullable	Description
DiagnosisKey	int	PK	No	
FacilityKey	int	FK_FactPatientDiagnosis_DimHealthFacility	No	
DiagnosisDateKey	int	FK_FactPatientDiagnosis_DimDates	No	
PatientKey	int	FK_FactPatientDiagnosis_DimPatient	No	
DiseaseKey	int	FK_FactPatientDiagnosis_DimDisease	No	
ClinicianKey	int	FK_FactPatientDiagnosis_DimMedicalProvider	No	
OutPatientsCodes_ID	int	FK_FactPatientDiagnosis_Dim_PatientsDiagnosisCodes	No	
FeverOrHistory	smallint		No	
HistoryCough	smallint		No	
NightSweats	smallint		No	
WeightLoss	smallint		No	
ChildContact	smallint		No	
OfferedHiv	smallint		No	
SentHIVTest	smallint		No	
AdmittedTo	smallint		No	
Detained	smallint		No	
ReferredForhiv	smallint		No	
ReferredForTb	smallint		No	

FactHIVRx

Name	DataType	Constraint	Nullable	Description
HIVRxKey	int	PK	No	
FacilityKey	int	FK_FactHIVRx_DimHealthFacility	No	
HIVRxDateKey	int	FK_FactHIVRx_DateKey	No	
PatientKey	int	FK_FactHIVRx_DimPatient	No	
Encounter_Type_Key	int	FK_FactHIVRx_DimEncounter_Type	No	
ConceptKey	int	FK_FactHIVRx_DimConcept	No	
value_boolean	tinyint		No	
value_coded	int		No	
value_drug	int		No	
value_numeric	int		No	
value_text	nvarchar(255)		Yes	

DimPatient

Name	DataType	Constraint	Nullable	Description
PatientKey	int	PK	No	
PatientAlternateKey	nvarchar(50)		No	
PatientName	nvarchar(120)		Yes	
Gender	nvarchar(50)		No	
Age	smallint		No	

DimHealthFacility

Name	DataType	Constraint	Nullable	Description
FacilityKey	int	PK	No	
FacilityAlternateKey	nvarchar(50)		No	
FacilityName	nvarchar(50)		No	
District	nvarchar(50)		No	

DimDates

Name	DataType	Constraint	Nullable	Description
DateKey	int	PK	No	
DateAlternateKey	date		No	
DateName	nvarchar(50)		No	
Month	int		No	
MonthName	nvarchar(50)		No	
Quarter	int		No	
QuarterName	nvarchar(50)		No	
Year	int		No	
YearName	nvarchar(50)		No	

DimDisease

Name	DataType	Constraint	Nullable	Description
DiseaseKey	int	PK	No	
DiseaseAlternateKey	nvarchar(20)		No	
DiseaseName	nvarchar(50)		No	
Description	nvarchar(50)		No	

DimMedicalProvider

Name	DataType	Constraint	Nullable	Description
ClinicianKey	int	PK	No	
ClinicianAlternateKey	nvarchar(50)		No	
ClinicianName	nvarchar(50)		No	
ClinicianInitial	nvarchar(50)		No	
ClinicianDesignation	nvarchar(50)		No	
OtherDesignation	nvarchar(50)		Yes	
Mentee	nvarchar(50)		Yes	

DimConcept

Name	DataType	Constraint	Nullable	Description
ConceptKey	int	PK	No	
ConceptAlternateKey	int		No	
ConceptName	nvarchar(255)		Yes	
concept_name_id	int		Yes	

DimEncounter_Type

Name	DataType	Constraint	Nullable	Description
Encounter_Type_Key	int	PK	No	
Encounter_TypeAlternateKey	int		No	
Encounter_TypeName	nvarchar(50)		No	
Description	nvarchar(250)		Yes	

APPENDIX D: IMPLEMENTATION CODE

SQL statements used to create database

```
USE [master]

GO

CREATE DATABASE [Integrated_ClinicalData_DM]

 ( NAME = N'Integrated_ClinicalData_DM', FILENAME = N'C:\Program Files
(x86)\Microsoft SQL
Server\MSSQL11.MSSQLSERVER12\MSSQL\DATA\Integrated_ClinicalData_DM.md
f' , SIZE = 141312KB , MAXSIZE = UNLIMITED, FILEGROWTH = 1024KB )

 LOG ON

( NAME = N'Integrated_ClinicalData_DM_log', FILENAME = N'C:\Program Files
(x86)\Microsoft SQL
Server\MSSQL11.MSSQLSERVER12\MSSQL\DATA\Integrated_ClinicalData_DM_lo
g.ldf' , SIZE = 149696KB , MAXSIZE = 2048GB , FILEGROWTH = 10%)

GO

USE [Integrated_ClinicalData_DM]

GO

CREATE TABLE [dbo].[DimPatient](

                    [PatientKey] [int] IDENTITY(1,1) NOT NULL,

                    [PatientAlternateKey] [nvarchar](50) NULL,

                    [PatientName] [nvarchar](120) NULL,

                    [Gender] [nvarchar](50) NULL,

                    [Age] [float] NULL,

PRIMARY KEY CLUSTERED

(

                    [PatientKey] ASC

)

) ON [PRIMARY]

GO
```

```
CREATE TABLE [dbo].[DimHealthFacility](
			[FacilityKey] [int] IDENTITY(1,1) NOT NULL,
			[FacilityAlternateKey] [nvarchar](50) NULL,
			[FacilityName] [nvarchar](50) NULL,
			[District] [nvarchar](50) NULL,
PRIMARY KEY CLUSTERED
(
			[FacilityKey] ASC
)
) ON [PRIMARY]
GO
CREATE TABLE [dbo].[DimDates](
			[DateKey] [int] IDENTITY(1,1) NOT NULL,
			[DateAlternateKey] [date] NOT NULL,
			[DateName] [nvarchar](50) NULL,
			[Month] [int] NOT NULL,
			[MonthName] [nvarchar](50) NOT NULL,
			[Quarter] [int] NOT NULL,
			[QuarterName] [nvarchar](50) NOT NULL,
			[Year] [int] NULL,
			[YearName] [nvarchar](50) NOT NULL,
 (
			[DateKey] ASC
)
) ON [PRIMARY]
GO
```

```
CREATE TABLE [dbo].[DimConcept](
                [ConceptKey] [int] IDENTITY(1,1) NOT NULL,
                [ConceptAlternateKey] [int] NULL,
                [ConceptName] [nvarchar](255) NULL,
                [concept_name_id] [int] NULL,
PRIMARY KEY CLUSTERED
(
                [ConceptKey] ASC
)
) ON [PRIMARY]
Go

CREATE TABLE [dbo].[DimDisease](
                [DiseaseKey] [int] IDENTITY(1,1) NOT NULL,
                [DiseaseAlternateKey] [nvarchar](20) NULL,
                [DiseaseName] [nvarchar](50) NULL,
                [Description] [nvarchar](50) NULL,
PRIMARY KEY CLUSTERED
(
                [DiseaseKey] ASC
)
) ON [PRIMARY]
GO
```

```
CREATE TABLE [dbo].[DimConcept](
                    [ConceptKey] [int] IDENTITY(1,1) NOT NULL,
                    [ConceptAlternateKey] [int] NULL,
                    [ConceptName] [nvarchar](255) NULL,
                    [concept_name_id] [int] NULL,
PRIMARY KEY CLUSTERED
(
                    [ConceptKey] ASC
)
) ON [PRIMARY]
GO
CREATE TABLE [dbo].[DimEncounter_Type](
                    [Encounter_Type_Key] [int] IDENTITY(1,1) NOT NULL,
                    [Encounter_TypeAlternateKey] [int] NULL,
                    [Encounter_TypeName] [nvarchar](50) NULL,
                    [Description] [nvarchar](250) NULL,
PRIMARY KEY CLUSTERED
(
                    [Encounter_Type_Key] ASC
)
) ON [PRIMARY]
GO
```

```
CREATE TABLE [dbo].[DimMedicalProvider](
                [ClinicianKey] [int] IDENTITY(1,1) NOT NULL,
                [ClinicianAlternateKey] [nvarchar](50) NULL,
                [ClinicianName] [nvarchar](50) NULL,
                [ClinicianInitial] [nvarchar](50) NULL,
                [ClinicianDesignation] [nvarchar](50) NULL,
                [OtherDesignation] [nvarchar](50) NULL,
                [Mentee] [nvarchar](50) NULL,
PRIMARY KEY CLUSTERED
(
                [ClinicianKey] ASC
)
) ON [PRIMARY]
GO

CREATE TABLE [dbo].[DimTBRxCodes](
                [TBRxCodes_Id] [int] IDENTITY(1,1) NOT NULL,
                [Regimen_Id] [int] NULL,
                [Regimen] [nvarchar](50) NULL,
                [DiseaseClass_Id] [int] NULL,
                [DiseaseClass] [nvarchar](50) NULL,
                [PtType_Id] [int] NULL,
                [PtType] [nvarchar](50) NULL,
                [HIVTestResult_Id] [int] NULL,
                [HIVTestResult] [nvarchar](50) NULL,
                [TreatCompleted_Id] [int] NULL,
                    [TreatCompleted] [nvarchar](50) NULL,
                    [TreatNotCompleted_Id] [int] NULL,
                    [TreatNotCompleted] [nvarchar](50) NULL,
 CONSTRAINT [PK_DimTBRxCodes] PRIMARY KEY CLUSTERED
(
) ON [PRIMARY]
GO
CREATE TABLE [dbo].[FactANC](
                    [ANCKey] [int] IDENTITY(1,1) NOT NULL,
                    [DateKey] [int] NULL,
                    [FacilityKey] [int] NULL,
                    [PatientKey] [int] NULL,
                    [ARTNo] [nvarchar](50) NULL,
                    [ANC_Codes_Id] [int] NOT NULL,
                    [ARVS] [smallint] NULL,
                    [ITN] [smallint] NULL
) ON [PRIMARY]
GO
CREATE TABLE [dbo].[FactHIVRx](
                    [HIVRxKey] [int] IDENTITY(1,1) NOT NULL,
                    [FacilityKey] [int] NULL,
                    [HIVRxDateKey] [int] NULL,
                    [PatientKey] [int] NULL,
                    [Encounter_Type_Key] [int] NULL,
                    [ConceptKey] [int] NULL,
                    [value_boolean] [tinyint] NULL,
                    [value_coded] [int] NULL,
            [value_drug] [int] NULL,
            [value_numeric] [float] NULL,
            [value_text] [nvarchar](255) NULL,
 CONSTRAINT [PK__FactHIVR__3AF5F158698E744F] PRIMARY KEY CLUSTERED
(
            [HIVRxKey] ASC
)
) ON [PRIMARY]
GO

CREATE TABLE [dbo].[FactLabTestResult](
            [LabTestKey] [int] IDENTITY(1,1) NOT NULL,
            [FacilityKey] [int] NULL,
            [LabTestDateKey] [int] NULL,
            [PatientKey] [int] NULL,
            [LabTestResult_Codes_Id] [int] NULL,
            [ARTNo] [nvarchar](50) NULL,
            [BsForMalaria] [smallint] NULL,
            [RdtForMalaria] [smallint] NULL,
            [HivTest] [smallint] NULL,
            [TbExam] [smallint] NULL,
            [StoolOrdered] [smallint] NULL,
            [Urinalysis] [smallint] NULL,
            [Hb] [smallint] NULL,
            [Vdrl] [smallint] NULL,
 CONSTRAINT [PK__FactLabT__2836F3258C08BC2E] PRIMARY KEY CLUSTERED
```

```
(
                    [LabTestKey] ASC
)
) ON [PRIMARY]
GO
CREATE TABLE [dbo].[FactPatientDiagnosis](
                    [DiagnosisKey] [int] IDENTITY(1,1) NOT NULL,
                    [FacilityKey] [int] NULL,
                    [DiagnosisDateKey] [int] NULL,
                    [PatientKey] [int] NULL,
                    [DiseaseKey] [int] NULL,
                    [ClinicianKey] [int] NULL,
                    [OutPatientsCodes_ID] [int] NULL,
                    [FeverOrHistory] [int] NULL,
                    [HistoryCough] [smallint] NULL,
                    [NightSweats] [smallint] NULL,
                    [WeightLoss] [smallint] NULL,
                    [ChildContact] [smallint] NULL,
                    [OfferedHiv] [smallint] NULL,
                    [SentHIVTest] [smallint] NULL,
                    [AdmittedTo] [smallint] NULL,
                    [Detained] [smallint] NULL,
                    [ReferredForhiv] [smallint] NULL,
                    [ReferredForTb] [smallint] NULL,
 CONSTRAINT [PK__FactPati__3FD995525831BA76] PRIMARY KEY CLUSTERED
(
                    [DiagnosisKey] ASC
)
) ON [PRIMARY]
GO
CREATE TABLE [dbo].[FactTBRx](
                    [TBRxKey] [int] IDENTITY(1,1) NOT NULL,
                    [FacilityKey] [int] NULL,
                    [TBRxDateKey] [int] NULL,
                    [PatientKey] [int] NULL,
                    [TBRxCodes_Id] [int] NULL,
                    [ARTNo] [nvarchar](50) NULL,
                    [LabSerialNo] [nvarchar](50) NULL,
                    [TransferIn] [smallint] NULL,
                    [Result2Mon] [smallint] NULL,
                    [Result5Mon] [smallint] NULL,
                    [Result8Mon] [smallint] NULL,
                    [CPT] [smallint] NULL,
                    [ART] [smallint] NULL,
 CONSTRAINT [PK__FactTBRx__4EE56E2B2361341D] PRIMARY KEY CLUSTERED
(
                    [TBRxKey] ASC
)
) ON [PRIMARY]
GO
```

ETL Code to fill Date Dimension

-- Create values for DimDates as needed.

```
Declare @StartDate date = '01/01/2012'
Declare @EndDate date = '01/01/2016'

-- Use a while loop to add dates to the table
Declare @DateInProcess date
Set @DateInProcess = @StartDate

While @DateInProcess <= @EndDate
 Begin
 -- Add a row into the date dimension table for this date
 Insert Into DimDates
 ( [DateAlternateKey], [DateName], [Month], [MonthName], [Quarter], [QuarterName], [Year], [YearName] )
 Values (
  @DateInProcess -- [Date]
  , DateName( weekday, @DateInProcess )  -- [DateName]
  , Month( @DateInProcess ) -- [Month]
  , DateName( month, @DateInProcess ) -- [MonthName]
  , DateName( quarter, @DateInProcess ) -- [Quarter]
  , 'Q' + DateName( quarter, @DateInProcess ) + ' - ' + Cast( Year(@DateInProcess) as nVarchar(50) ) -- [QuarterName]
  , Year( @DateInProcess )
  , Cast( Year(@DateInProcess ) as nVarchar(50) ) -- [YearName]
  )
 -- Add a day and loop again
 Set @DateInProcess = DateAdd(d, 1, @DateInProcess)
 End

-- Add additional lookup values to DimDates

Set Identity_Insert [Integrated_ClinicalData_DM].[dbo].[DimDates] On
Insert Into [Integrated_ClinicalData_DM].[dbo].[DimDates]
  ( [DateKey]
  , [DateAlternateKey]
  , [DateName]
  , [Month]
  , [MonthName]
  , [Quarter]
  , [QuarterName]
  , [Year], [YearName] )
  Select
   [DateKey] = -1
  , [DateAlternateKey] =  Cast('01/01/1900' as nVarchar(50) )
  , [DateName] = Cast('Unknown Day' as nVarchar(50) )
  , [Month] = -1
  , [MonthName] = Cast('Unknown Month' as nVarchar(50) )
  , [Quarter] =  -1
  , [QuarterName] = Cast('Unknown Quarter' as nVarchar(50) )
  , [Year] = -1
  , [YearName] = Cast('Unknown Year' as nVarchar(50) )
  Union
```

```
Select
  [DateAlternateKey] = -2
, [Date] = Cast('01/01/1900' as nVarchar(50) )
, [DateName] = Cast('Corrupt Day' as nVarchar(50) )
, [Month] = -2
, [MonthName] = Cast('Corrupt Month' as nVarchar(50) )
, [Quarter] =  -2
, [QuarterName] = Cast('Corrupt Quarter' as nVarchar(50) )
, [Year] = -2
, [YearName] = Cast('Corrupt Year' as nVarchar(50) )

Set Identity_Insert [Integrated_ClinicalData_DM].[dbo].[DimDates] Off
```