Participatory Approach to Data Warehousing in Health Care: Uganda's Perspective

Charles Otine

Blekinge Insitute of Technology Licentiate Dissertation Series

No 2011:04
ISSN 1650-2140
ISBN 978-91-7295-204-1

Participatory Approach to Data Warehousing in Health Care: Uganda's Perspective

Charles Otine

School of Planning and Media Design
Department of Technology and Aestetics
Blekinge Institute of Technology
Sweden

Blekinge Institute of Technology

Blekinge Institute of Technology, situated on the southeast coast of Sweden, started in 1989 and in 1999 gained the right to run Ph.D programmes in technology. Research programmes have been started in the following areas:

Applied Signal Processing
Computer Science
Computer Systems Technology
Development of Digital Games
Human Work Science with a special Focus on IT
Interaction Design
Mechanical Engineering
Software Engineering
Spatial Planning
Technosicence Studies
Telecommunication Systems

Research studies are carried out in faculties and about a third of the annual budget is dedicated to research.

Blekinge Institue of Technology
S-371 79 Karlskrona, Sweden
www.bth.se

Department of technology and Aestetics
School of Planning and Media Design
Graphic Design and Typesettning: Mixiprint, Olofstrom
Publisher: Blekinge Institute of Technology
Printed by Printfabriken, Karlskrona, Sweden 2011
ISBN 978-91-7295-204-1
um:nbn:se:bth-00491

Table of Contents

Abstract

This licentiate thesis presents the use of participatory approach to developing a data warehouse for data mining in health care.

Uganda is one of the countries that faced the largest brunt of the HIV/AIDS epidemic at its inception in the early 1980s with reports of close to a million deaths. Government and nongovernmental interventions over the years saw massive reductions in HIV prevalence rates over the years. This reduction in HIV prevalence rates led to great praises by the international community and a call for other countries to model Uganda's approach to battling the epidemic. In the last decade the reduction in HIV prevalence rates have stagnated and in some cases increased. This has lead to a call for reexamination of the HIV/AIDS fight with an emphasis on collective efforts of all approaches.

One of these collective efforts is the introduction of antiretroviral therapy (ART) for those already infected with the virus. Antiretroviral therapy has numerous challenges in Uganda not least of which is the cost of the therapy especially on a developing country with limited resources. It is estimated that of the close to 1 million infected in Uganda only 300,000 are on antiretroviral therapy (UNAIDS, 2009). Additional challenges of the therapy includes following through a treatment regimen that is prescribed. Given the costs of the therapy and the limited number of people able to access the therapy it is imperative that this effort be as effective as possible. This research hinges on using data mining techniques with monitoring HIV patient's therapy, most specifically their adherence to ART medication. This is crucial given that failure to adhere to therapy means treatment failure, virus mutation and huge losses in terms of costs incurred in administering the therapy to the patients.

A system was developed to monitor patient adherence to therapy, by using a participatory approach of gathering system specification and testing to ensure acceptance of the system by the stakeholders. Due to the cost implications of over the shelf software the development of the system was implemented using open source software with limited license costs. These can be implemented in resource constrained settings in Uganda and elsewhere to assist in monitoring patients in HIV therapy. A algorithm that is used to analyze the patient data warehouses for information on and quickly assists therapists in identifying potential risks such as non-adherence and treatment failure. Open source dimensional modeling tools power architect and DB designer were used to model the data warehouse using open source MYSQL database.

The thesis is organized in three parts with the first part presenting the background information, the problem, justification, objectives of the research and a justification for the use of participatory methodology. The second part presents the papers, on which this research is based and the final part contains the summary discussions, conclusions and areas for future research.

The research is sponsored by SIDA under the collaboration between Makerere University and Blekinge Institute of Technology (BTH) in Sweden.

My licentiate thesis is dedicated to all
who lost their lives due to the HIV/AIDS epidemic in Uganda.

Acknowledgment

I would like to thank the support of all my two supervisors Prof. Lena Trojer and Dr. Samuel Baker Kucel for guiding me through this process. I further give appreciation to my “half” supervisor Eng. Dr. Pater Lating for his fatherly and professional guidance. My parents, family and especially mother for all the support you have given me through all the periods of study away from home. I would also like to acknowledge my eldest sister Consolata Otine for such a strong support base. Lastly to a combination of support from SIDA to the Faculty of Technology and BTH, Campus Karlshamn for the home provided away from home. Thank you all.

List of Abbreviations

AIDS	Acquired Immune Deficiency Syndrome
AR	Action Research
ART	Antiretroviral therapy
ARV	Antiretroviral drugs
BTH	Blekinge Institute of Technology
CSV	Comma Separated Values
DW	Data Warehouse
DBMS	Database Management System
DSS	Decision Support System
DW	Data Warehouse
HIS	Health Information System
HIV	Human Immunodeficiency Virus
ICT	Information Communication Technology
KDD	Knowledge Discovery Databases
MAK	Makerere University
MOH	Ministry of Health Uganda
MySQL	MySQL Database Management System
NGOs	Non-Governmental Organizations
PAR	Participatory Action Research
PD	Participatory Design
PRA	Participatory Rural Appraisal
RRA	Rapid Rural Appraisal
UBOS	Uganda Bureau of Standards
UN	United Nations
WHO	World Health Organization

INTRODUCTION

This licentiate thesis is part of a study on using data mining in health care to facilitate HIV patient monitoring in Uganda. The research focuses on developing and implementing the integration of a data mining system in AIDS patient monitoring in a resource constrained setting by using cheap open source software to mitigate the otherwise expensive costs of third party software and data base management systems.

It emphasizes the importance of participatory approach to development of systems for health that ensures ownership by the stakeholders and continuity for the system implemented. The research is linked to the Millennium Development Goal (MDG) number 6 that seeks to halt and begins to reverse the spread of HIV/AIDS, malaria and other disease (UN, 2010). This is by introducing systems that facilitate the monitoring of HIV/AIDS adherence. The system uses analyses data in the patient treatment data warehouses to seek for patterns that indicate anomalies in therapy such as treatment failure and non-adherence.

The thesis is organized in 3 parts. Part I give the background of the study, the objectives, and scope, rational for the study, literature review and methodology. PartII gives an introduction to the papers followed by a presentation of the 4 papers. PartIII gives a brief summary of the papers, the concluding discussions and the statement of scientific contribution and originality. Finally a presentation of the further areas for research boding out of this research is given.

PART I – BACKGROUND

1.1. ICTs

Advances in Information and Communication technology (ICTs) are factors that should not be overlooked in establishing synergies to combat health care deficiency. This is particular important in the case of ICTs, that somewhat assist in the generation of new knowledge through analysis of data, which is collected through day-to-day operations of health care facilities. Large amounts of data are collected on a daily basis by health care facilities (Tan, 2003), but the form of this information and the way it is organized does not facilitate the generation of new knowledge (Siri Krishan Wasan, 2006). The use of ICTs in health care in the developing world has been minimal. This has been attributed to many reasons including the high costs of the ICTs (Watts, 2006). These costs include cost to infrastructure, hardware and software procurement, training needs among others. The costs incurred due to software in terms of license fees can be reduced somewhat by using open source software. Open source software is software that complies with a number of set criteria (Coar, 2006); one of the criteria being free distribution (Coar, 2006). Under the free distribution criteria, the license does not restrict a party from selling or giving away the software as a component of an aggregate software distribution containing programs from different sources. No fee or loyalty is required for the sale under this license. This arrangement significantly reduces the cost and access to software that can be used by the developing countries, which would not otherwise be able to afford the prohibitively expensive license fees involved.

In the developing world like Uganda, where the use of electronic medical records is still in its infancy, many health care providers still use the traditional paper methods of storing information. The resultant effect is difficulty in data retrieval, protection, backup, and more importantly analysis (Kriegel, 2007). The above weaknesses can be reduced through the incorporation of ICTs in heath care. The use of ICTs can help to reorganize data collected from health care providers into a form that can assist analysts in studying the information with the aim of generating new knowledge. This is referred to as data mining. The numerous open source solutions available are helpful in this endeavor, since the license options can reduce the software costs involved.

1.2. Data Mining

Data mining was named one of the ten emerging technologies to change the world by the MIT technology review (MIT, 2001). Data mining aims to identify unique patterns (Cabena P., 1998; Hand D., 2001; Larose, 2005) in huge quantities of data by methodically sifting through this data, which is generated by health care providers during their day-to-day activities. There is strong argument to confirm the fact that data mining is supported by categorization. Organization of the data into a format suitable for data mining involves categorization of the information into dimensions. Data mining is typically carried out on data that is stored in a data warehouse, which is a central repository of time variant data (Kimball and Ross, 2002;Connolly, 2002).

By employing the use of data mining on huge data sets, we can accomplish tasks such as prediction, forecasts, estimation, classification, clustering and association. These tasks when used in combination can be very useful to health care providers especially when dealing with patients that are facing chronic illness such as AIDS. The AIDS disease has adversely affected the nations of Africa. According to the United Nations Joint report on AIDS (2004), sub-Saharan Africa has close to two-thirds (2/3) of all the people infected with HIV in the world. With this scenario in mind, Africa has the largest percentage of AIDS patient's information in the world and this culminates into substantial quantities of data that can be used for study. This data if carefully stored could be used to generate a variety of information including the disease progression; number of people infected and gives an indication of the rate of infection at the general level. Specifically the data can be used in AIDS patient monitoring by giving information such as the disease progression, signs of relapse, treatment failure and so on based on similar patterns observed and analyzed in other patients.

Evidently with the large quantities of patient data available for data mining in Africa, the framework for data mining in open source is not yet in place. In part this is due to the format of data storage presently employed in most countries of sub-Saharan Africa, Uganda inclusive. Given the above one can argue that 'resource' in terms of data is wasted. *This research therefore aims to study, propose and develop a way forward in integrating data mining in health care to assist in monitoring AIDS patient treatment.* An open source health care data warehouse will be developed where the actual data mining shall be carried out, this will help define a standard or framework on which to develop future open source data warehouses.

1.3 Status of Information Systems in Health Care Uganda

The different information systems in Uganda have their different strengths and weaknesses, these present opportunities, where this research could be synergized with other interventions in addressing the Millennium Development Goal (MDG) number 6.

Strengths

- There is evidence and existence of basic functional HIS coordination mechanism present under the resource centre for Ministry of Health (MOH) Uganda.
- Appreciable demand for information from senior managers, policy makers, development partners for decision making.
- There is relative good usage of available information from Health Information System (HIS) for planning, budgeting and resource allocation at national level. This would have a great impact if it could be extrapolated to the local levels.
- There exists basic ICT infrastructure in many districts (77). This has been further improved by the fiber connectivity between certain districts in Uganda both by the government and other private communications companies.
- Basic routine system for data collection to the national level. This uses the local health facilities and centers feed into the hierarchy of District health systems until the ministry of health.
- Information from population surveys conducted by Uganda Bureau of Standards (UBOS) available.
- There is regular dissemination of information through meetings /workshops and other forums such as reports.

Weaknesses

- Private for profit facilities that report to HIS is poor. Most Private facilities maintain their own data and are reluctant to share with the others. This could work against the concept of a data warehouse.
- Lack of a training policy for health officers at all levels.
- Lack of a comprehensive strategic plan.
- No relational data warehouse for all HIS sources (attempts at a national health databank).
- Poor ICT infrastructure. The government policy of creating new districts, which means that many of these districts have very poor infrastructure let alone ICT infrastructure. This makes aggregation and collection of health information system from these districts extremely difficult.
- Poor integration of regional and national health facilities.
- Inadequate disaggregation of data by gender, socio economic status and geographical profiles (drill downs, dimensions data warehouse)

The above strengths and weaknesses present opportunities such as:

1. Capacity building for health information officers.

2. Improvements to ICT infrastructure country wide by the government.
3. Opportunities for research in HIV/AIDS given the progression of the disease, Uganda's experience.
4. Development of national and sub-national web based health data warehouses and repositories.
5. Trainings: The involvement of all stakeholders during the development ensures that the end users of the system will be knowledgeable about the system. These create user groups that are easier to train and champion the use of the system amongst their colleagues.

1.4. Definition of Key Theoretical Terms

AIDS: Acquired immunodeficiency syndrome. This is the final stage in an HIV virus infection whereby the human body can no longer effectively defend itself against infection from common and opportunistic infections.

Data Mining: Process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques (Fayyad, 1996).

Data mining dimension: This is a specific category of information about the subject understudy. They represent the tables in the data warehouse and are used to collect information about the process that is being studied (Kimball and Ross, 2002). When studying HIV Patient prescription information, dimensions can include patient information, drug information, test information, time when the prescription was given and so on. The time dimension is crucial because it helps to track the progression of the area being studied over a period this allows analysis of change to be undertaken.

Data Marts: These represent Small 'data warehouses' modeled around a specific business process. They are subsets of data warehouses that support the requirements of a particular department or business function. A fully fledged data warehouse can be made up of a set of data marts each with a specific business process that communicates with each other (Kimball and Ross, 2002).

Monitoring information system: A system that continually tracks down a dimension or dimensions and may give warnings in the face of an anomaly in the said dimension.

Data warehouse: This is the knowledge discovery database. This special database is characterized by data that is subject-oriented, time-variant, non-volatile and integrated and supports the management's decision-making process (Inmon 1996; Kimball and Ross, 2002).

HIV: Human immunodeficiency virus. This is the virus that causes the disease AIDS.

Health care: This refers to the entirety of care, services and supplies that are furnished to an individual and related to the health of the individual. This could include activities such as medical testing, diagnosis of the disease, treatment direction (prescriptions given and drugs), counseling where necessary, preventive care, discharge or end of treatment. The totality of these activities makes up health care.

1.5. Statement of the Problem

Patient monitoring during treatment is of the utmost importance. A patient's progress has to be consistently screened from the moment of diagnosis to the end of treatment. The degree of patient monitoring may vary depending on the medical condition the patient has been diagnosed with at the time. For the cases of chronic illness, for instance AIDS, a constant, absolute and accurate degree of patient monitoring is required. In Uganda, patient monitoring is a tedious task not only because of the manual techniques of patient monitoring but also because of the lack of incorporation of the broad advances in information systems.

While the large quantities of data involved and the mode of storage of the data makes manual analysis inefficient (Kriegel, 2007), the data can still be organized into databases that can be used to discover new knowledge from the past data using data mining techniques. This is especially important in countries like Uganda, which are heavily affected by chronic illness such as AIDS where medical cures are still being sort and all new knowledge about the disease are welcome.

1.6. Objectives

The main objective of the research project is to develop a HIV-data warehouse using participatory methods.

Specific Objectives

1. To review critically the literature on data mining in health care and patient monitoring focusing on AIDS patients.
2. To survey the status of information systems in health care in Uganda.
3. To develop the conditions for integrating data mining in AIDS patient monitoring.
4. To identify the constraints to incorporating data mining in monitoring of AIDS patients in Uganda.
5. To explore the deployment of data warehouses for use in data mining or knowledge discovery in AIDS.
6. To develop, simulate and run an open source health care data warehouse targeting AIDS patients in Uganda.

1.7. Justification

The last decade has seen ICTs playing an increasingly important role in our daily lives. This has been more evident in health care. Medical treatment faces challenges of discovering knowledge from the growing volume of data and sharing the generated knowledge amongst the people concerned. The research will take advantage of the

potential of huge quantities of information, generated from the daily activities during patient care for knowledge discovery through data mining by using ICTs. This will assist in putting the health care givers in a more informed position, when monitoring or administering patient treatment thus minimizing incidences of human error in treatment.

An important function of knowledge discovery databases is their use in predictions and forecasts (Fayyad et al. 1997; Siri et al. 2000). The outcome of this research can be used as a basis in predicting patient medical treatment direction and forecasting diseases relapses due to similar conditions in other patients with the same condition. This will help the healthcare 'learn using past data' which is a very important aspect when dealing with illness such as AIDS, where there are no known cures yet and everybody is trying to learn more about the disease while looking for cures.

Generation of new knowledge is arguably one of the most important steps in combating the AIDS pandemic in the world. The research will enable health care providers learn important commonalities in the disease among different patients thereby assisting in treatment monitoring. This will move towards addressing the MDG goal number 6 that deals with combating HIV/AIDS, Malaria and other diseases (UN, 2008).

1.8. Research Questions

The above scenario presents research questions that need to be answered as far as data mining in the healthcare of AIDS patients in Uganda is concerned.

- What is the status of information systems in health care in Uganda?
- What are the tools for modeling HIV Data in resource constrained settings?
- What are the constraints in developing a patient data warehouse using open source software?
- What is the importance of user engagement and participation in developing data warehouse systems in health care?
- What dimensional model in an HIV/AIDS Data warehouse can help in monitoring patient adherence to medication?

1.9 Scope

The research involves the development a data warehouse for health care in Uganda. The data mining operations are then implemented and simulated on the data warehouse. The data for the data warehouse is focusing on analysis of AIDS patient (adult women and men) records from ART centers & hospitals and other relevant HIV health centers such as Mild May Uganda. To enhance co-evolution of the system, the cooperating partners are identified including the researcher, faculty of technology Makerere, selected health care professionals, patient organizations, and Non-Governmental organizations (NGOs) dealing with AIDS. These cooperating partners are involved in

the development process thus ensuring that the outcome is a functional system. The development takes into consideration the information system used in healthcare in Sweden (Karlshamn and Lund Hospital) and in Muhibili hospital in Dar es Salaam.

1.10 Sample Screen Shots of the System

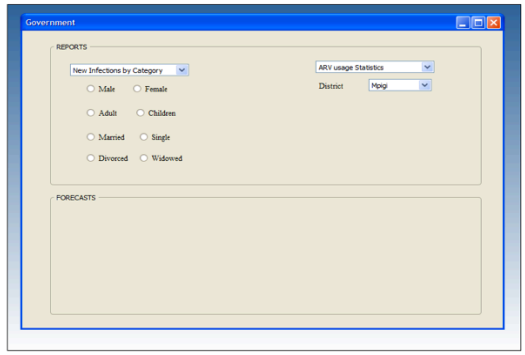


Figure 1: Patient Reporting Screen and ARV Statistics

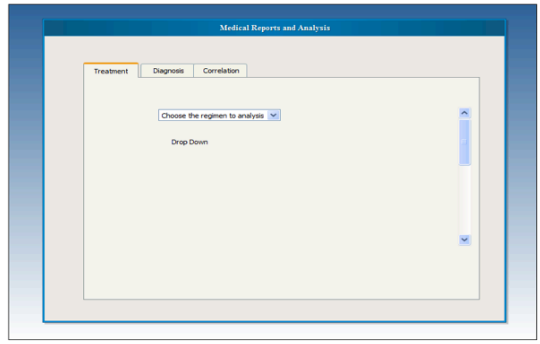


Figure 2: Patient Medical Report and Diagnosis and Regimen Options

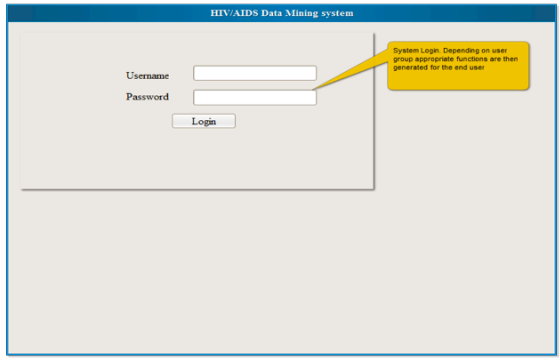


Figure 3: System Security and authentication

PART II – METHODOLOGICAL CONSIDERATIONS

2.1. Overview of Action Research

To achieve the objective of this research a branch of action research methodology is used. This is because this research requires complete end-user involvement in the process to ensure that the final system is functional and accepted by the community. Action research is conducted over different cycles with iterative and reflective steps that involve a process of taking an action and studying the action undertaken (O'Brain 1998 ; Riel, 2007). Ensuring that the research takes shape as it is being done as depicted in the diagram below.

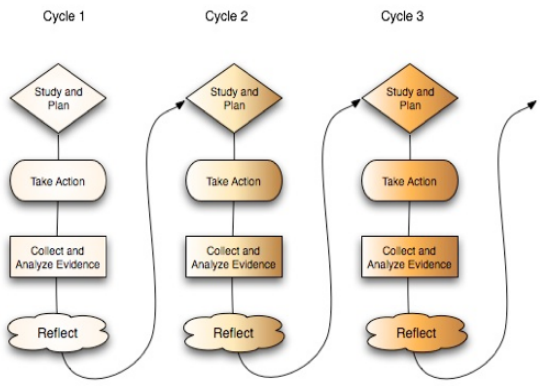


Figure 1: Progressive Problem solving with Action Research (Riel, 2007) Understanding Action Research)

2.2. Participatory Action Research

Action research has evolved over the years into Participatory Action Research (PRA) originating among the trade unions in Scandinavian countries. This resulted in research methodologies such as Participatory Design (PD), Participatory Rural Appraisal (PRA) (Chambers, 1990) and Rapid Rural Appraisal (RRA) that focus the research methodology on involvement of the entire community. PAR was used to define the requirements of the data warehouse system and to involve the users in define the goal of the project. It has been used to establish what the initial step for the project should be.

PAR research methodology is used in this research because it ensures that all the stakeholders in the research area participate in the research. The participants share experiences, learn from each other and determine the direction of the project. The researcher acts as a facilitator, guiding the collaborators through planning, taking action, observing, evaluation and critical reflection (Jean, 2002) with the collaborators/stakeholders being regarded more like co-researchers. Wadsworth (Wadsworth, 1998) likens participatory action research to learning by doing. Since all stakeholders concerned are involved in the entire research there is an element of democracy whereby those being helped determine the purposes and outcomes of the inquiry. Close involvement of users in the research enables users to bring innovation into the research as they share their different challenges, success and experiences (Hippel, 2005) and learn from each other.

To achieve the objectives of the research the good principles of PAR are applied including use of optimal ignorance, offsetting biases, learning from and with rural people, learning rapidly and progressively and triangulation which refers to using more that 2 sources to cross check facts.

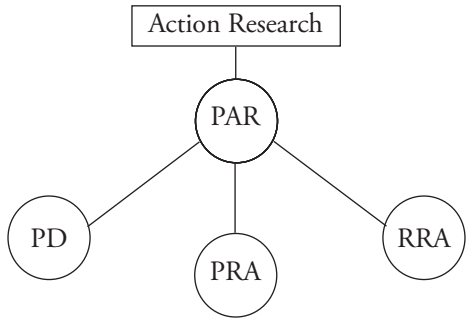


Figure 2: Action Research Linkage with Participatory Research

In this methodology PAR was used with seven major research techniques that were customized for this particular research. Additional Research techniques were also employed to ensure all aspects of the project were addressed. These are outlined below and are carried out iteratively through the course of the research:

Secondary Data Review

Content analysis was done on data gathered from selected relevant books, articles, journals, and reports. This involved conceptual and theoretical literature review on data mining and its impact on patient monitoring in Africa and Europe. This helped to indicate the status of information systems in health care in Uganda and helped in the identification of mitigating factors in the incorporation of data mining and data warehousing in monitoring of AIDS patients in Uganda.

Observation

Further information was gathered through direct participant observation. Operations of selected health care facilities providing health care services to AIDS patients were observed to gain understanding of how they carry out their operations. This gave an indication of the problems faced by the selected stakeholders in their normal activities and assisted in suggesting solutions to these problems. This technique was crucial in gathering further information on the status of health information systems in Africa as well as developing the conditions for integrating data mining in AIDS patients monitoring.

Semi structured interviews, Surveys and Questionnaires

Informal sessions with stakeholders were organized whereby only selected questions were prepared beforehand and new questions generated as the interviews were conducted from the answers received from those being interviewed. Surveys and questionnaires are used where the source of the information is extremely large and structured in form. These are effective in collecting qualitative information from the stakeholders involved. The stakeholders in this case include medical researchers, health care professionals, NGOs dealing with AIDS patients, faculty of technology Makerere and selected AIDS patients.

This technique assists in the generation of a report on the status of health information systems, identifying the constraints to incorporating data mining in monitoring AIDS patients, and assessing the training needs of personnel in Uganda to use health care information systems. This enables the definition of the requirements of the system proposed.

Workshop and Brainstorm seminars

Workshop involving all the local stakeholders as well as outsiders who are knowledgeable on the subject of the research was organized. Most crucially at the beginning of the research to enable the community identify and recognize the problem together and plan the way forward by coming up with a common vision acceptable to all. This also assisted in refining requirements and identifying the business processes that was common to the stakeholders and that would form the initial focus of the data warehouse.

Encouraging participation of all the people involved helped in avoiding situation where the researcher could 'imposes' their will on the community involved. The workshops

were used to update all concerned on the progress of the research, and enable all concerned to be involved in reviewing and providing evaluation of the actions taken as a group. This encourages learning from the different actions being taken in the research. This results in a forum where the community can share their different experiences, which can help the community to learn from one another as they evaluate the research, as well as to adapt and promote positive outcomes from the research.

Presentations

Communication is a crucial element of keeping the community in the loop as regards the progress of the research and thus ensures their continual involvement in the research, which is a core issue of participatory research. Communication of relevant information to the stakeholders is done using presentations in the form of portraits, posters, flow charts and diagrams. This is crucial because core to the success of PAR is the community's (all stakeholders) fair understanding of the issue at hand. Appropriate screen mockups were used to present the system design to the stakeholders to elicit feedback and quick demonstrations of prototypes.

2.3. System Design

The previous techniques are among other objectives helping to define the project scope and perform analysis. This is making clear the requirements, priorities, project plans and resource plans. These helped to inform the system design phase and assist in generating the logical data models, defining the extraction of data from source systems and the reporting and analytical functionality. As in the case with previous techniques, this is done in iteration with the involvement of the stakeholders using techniques shall, brainstorm sessions, one on one interviews, presentations, and observation. This phase assisted in developing the system dimensional model that formed the initial foundation for the data warehouse. Two main system processes were identified with the assistance of the health care providers and these became facts that were used as the initial data marts. This was the process related to prescriptions as patients were being assigned regimens of treatment and Medical Checkup as patients were continuously monitored during treatment by tests on a weekly, by monthly, monthly and quarterly basis.

2.4. Implementation and Testing

The infrastructure setup was done in this phase with setup of the physical databases using the open source database management system (MySQL). The extraction of the data from the source systems was done and loaded into a staging area database (called STG) where cleaning operations was done on the data. The cleaning involved quantitative adjustments as well as categorical mappings in the database. The quantitative adjustments was to have a common units for measured values entered into the data warehouse including parameters such as weight, height, CD4 counts and so on. The reporting and analytical functions of the system were then refined. Development of

the initial user interface was done at this stage using the information generated from the stakeholder from the screen mockups. Progressive testing of the system is carried out in this phase. For acceptance of the system by the end users, the testing is done in conjunction with all the stakeholders.

PART III – Introduction to the papers

Four papers were developed during the course of the research with 2 of these published in international conference journals and 1 submitted for publishing and the last in Manuscript.

Paper I: Otine, C.D., Kucel,S.B. and Trojer, L. (2007). *Knowledge Discovery in Health Care Using Data Mining.*

Published in the Proceeding of the 1st International Conference on Collaborative Research for Technological Development.ISBN:

The initial paper introduces the concept of knowledge discovery in health care using data mining and links this to the Ugandan as a developing country with limited resources. It provides an overview of knowledge discovery and introduces the knowledge discovery process and the concept of data warehouses and data marts. Different existing techniques of data mining are reviewed with selected applications to health care. The impact of the HIV/AIDS epidemic in Uganda is introduced and a linkage to the opportunity presented in terms of available information for data warehousing and data mining. It concludes by providing potential constraints to the implementation of such a system as well as the areas and opportunities for further research.

Paper II: Otine, C.D., Kucel, S.B. and Trojer, L. (2010). *Dimensional Modeling of HIV Data Using Open Source.*

Published in the Proceedings of World Academy of Science, Engineering and Technology Issue 63. March 2010.

The paper focuses on developing a dimensional model for an HIV/AIDS data warehouse using open source. Two open source data modeling tools are reviewed; both are used to generate a data warehouse model around two main business processes identified through interaction with stakeholders. These were prescription distribution and medical checkup. Source systems from selected health care providers were reviewed and used to validate the model generated. The model for the database is based on a star schema. Both models were then verified using the open source tools by linking up with two unique database management systems. The paper concludes by identifying certain weaknesses of using open source including the limited support in some cases and the limitations in handling complex data.

Paper III: Otine, C.D., Kucel, S.B. and Trojer, L. (2010). Implementation of an Open Source HIV/AIDS Data Warehouse in Uganda

Submitted for Publishing in the International Conference on Computer Science and Software Engineering Italy 27-29 April 2011

The paper focuses on the implementation of an open source data warehouse for AIDS patients basing on the dimensional model developed. The data warehouse was hosted on a MySQL database management system. SQL scripts were created for each of the dimensions identified conceptualization of the dimensional model. A second staging database STG was also created to hold the data from the different stakeholder source systems for cleaning and transformations before loading into the database. Initial and subsequent loading scripts were generated for loading each of the dimensions in the data warehouse. The time dimension was loaded using the one-date-everyday approach.

Paper IV: Otine, C.D., Kucel, S.B. and Trojer, L. (2011). Stakeholder Engagement and Participation in Determining Requirements for a Data Warehouse System for HIV/AIDS: Uganda's Experience

In Manuscript and not submitted.

This paper examines the process that was used to define the key requirements for the HIV/AIDS data warehouse developed in open source software. It highlights the key expectations of the different stakeholders involved in different aspects of HIV/AIDS service provision. This includes the care givers (doctors, nurses, pharmacists and counselors), the government, donors and other Non Governmental Organizations. These had different expectations and priorities of the system and in determining the direction and goal of the project required harmonizing their different expectations. The need for review and realignment is noted as especially in dealing with the different expectations. The success point here was the realization by the stakeholders that data warehouse systems should be built incrementally and their functionality gets more effective with time, data and increased collaboration. Therefore the initial unified goal is the most important, individual competing requirements that are not included originally can subsequently be added, as new processes, data marts and dimensions are defined.

3.1 Paper I

Knowledge Discovery in Health Care Using Data Mining

C.D Otine[1], S.B. Kucel[2] and L. Trojer[3]

ABSTRACT

The exponential growth of data banks in health care creates opportunities for knowledge generation using data mining. Advancements in information and communication technologies now mean that large quantities of data collected from different sources can be easily stored, secured and retrieved for analysis using databases. For data mining to be carried out the data sets from the different interacting source systems have to be organised in a data warehouse. Data mining offers the potential for exploring hidden patterns in data sets of a particular domain in this case health care. This can eventually be used to perform diagnosis and prognosis on different patient health care condition. Furthermore it places health care providers at a more informed point by enabling predictions hence through classification enhancing generation of new knowledge. This paper provides the state-of-art on data mining and its role on knowledge discovery in the health care sector.

Keywords: Database; Data mining; Data warehousing; Health care; Knowledge discovery.

INTRODUCTION

Continuous innovations in information technology means that ICTs will continue to play an increasing role in our daily lives. Take for instance the improvements made in terms of data storage; the move from paper storage to digital, the capability to store ever increasing quantity of data easily with development of improved digital storage drives with ever increasing capacity. This situation provides an opportunity for archival and analysis of data collected from different organisational operations over extensive periods of time. The size of these data banks necessitates the need to move away from the manual techniques of data analysis (Kriegel et al.2007) to the computerized techniques. Different database management systems provide the capability to archive and perform complex manipulation of the data that is collected; with the posibility of provision of extra functionality such as backup, encryption, security and complex quick analysis.

The health care sector can stand to gain from some of these advancements being made. Successful treatment of patients by health care providers is greatly determined by the ability of the health care provider to document treatment of patients. This results in the

1 Assistant Lecturer, Department of Electrical Engineering, Faculty of Technology, Makerere University, P.O. Box 7062, Kampala, Uganda. Email: hautine@tech.mak.ac.ug

2 Lecturer, Department of Mechanical Engineering, Faculty of Technology, Makerere University, P.O. Box 7062, Kampala, Uganda. Email: sbkucel@tech.mak.ac.ug

3 Professor, Department of Technoscience Studies, Blekinge Institute of Technology, P.O. Box 214, 374 24 Karlshamn, Sweden. Email: lena.trojer@bth.se

acquisition of considerable volumes of data from patients both past and present information, including patient bio information, treatement options, prescriptions, next of kin, diagnosis, and past illness to mention a few. The use of electronic medical records with powerful databases and data warehouses helps to improve the archiving and manipulation of patient information by the health care providers. Continued documentation of health care ensures the growth in the amount of information that is stored in these databases and data warehouses. This opens up the posibility for more detailed analysis of the information in these databases through data mining.

Analysis of large database sources leading to the identification of useful otherwise hidden pattern is what is referred to as data mining (Fayyad et al.1996). The identification of these hidden patterns can then be used to provide insight into new knowledge depending on the study area. Data mining is becoming a widely used branch of computer science (Kriegelet al. 2007) as is evident in its application in areas such as financial investment (Se-Hak&Steven, 2004), e-commerce, retail, manufacturing, telecommunications (Smith & Gupta, 2000), marketing and health.

This paper examines the use of data mining in health care in the process of knowledge discovery. It will also provide current applications of data mining in specific medical conditions. The paper has a methodology section indicating the identification, selection and analysis of the literature that was obtained. The results are then presented including an overview of knowledge discovery and data mining, specific application of data mining to health care and the concerns and requirements. The paper ends by providing a short conclusion.

METHODOLOGY

Identification of Publications

In order to identify selected publications in the area of knowledge discovery in health care systems, articles were selected from various databases and resources linked to the Electronic Library Information Navigator (ELIN) library system at Blekinge Tekniska Hogskola (BTH). This includes links to databases and resources such as Science Direct, Springer, BioMed Central, Blackwell Synergy, Emerald, Cambridge journals, and IEEE .Keywords such as database, data mining, data warehousing, health care, and knowledge discovery were used to facilitate the searches.

Selection of Publications

This literature review considered the papers and articles published between 1990 and 2007 in the areas of data mining, knowledge discovery and health care. The literature published in English were selected with specific emphasis placed on literature covering the relationship between knowledge discovery and data mining, applications of data mining to health care in different countries in the developed world and constraints and requirements for the setup of data mining in health care.

Analysis Strategy for Selected Publication

The selected articles were then analyzed for trends in data mining over the period of review started above. The selected literature was categorized according to areas of

emphasis including, framework for knowledge discovery and data mining, techniques, methods and algorithms for knowledge discovery and data mining, specific instances of application of knowledge discovery and data mining to health care.

KNOWLEDGE DISCOVERY AND DATA MINING OVERVIEW

Data mining is one of the key steps in the knowledge discovery process, (Fayyadet al. 1996) and (Wright, 2007). Knowledge discovery is often defined as "the non-trivial extraction of implicit, unknown, and potentially useful information from data" (Fayyad et al.1997). The data for use in the knowledge discovery process has to be prepared. Data preparation involves selecting the particular data to be targeted from the set of all data and performing some degree of 'data cleaning' after which the data adjustments can be made to the data before being stored in a central repository, the data warehouse (Fayyad et al. 1996; Brodley et al. 1999). Adjustments to the data are necessary in emphasizing the dimensions under study whereas cleaning is necessary due to errors that may be inherent in the target data. Erroneous target data may result in the discovery of misleading patterns during data mining.

The relationship between knowledge discovery process and data mining can best be summed up by the framework proposed by Fayyad (Fayyad et al.1996). This is shown in Figure 1, adapted from the same source. In this adaptation the transformed data is stored in a data warehouse.

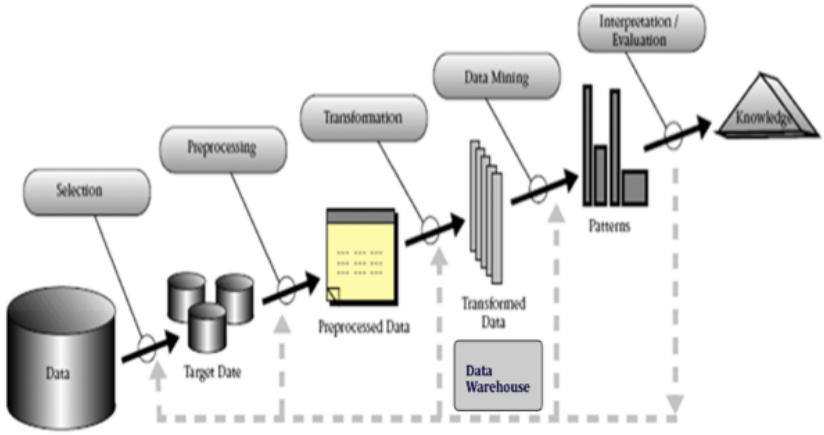


Figure 1: An adjustment of the Knowledge Discovery process (Adapted from Fayyad et al, 1996)

Data mining involves the application of different algorithms and other techniques of analysis to large data sources in an attempt to identify unique patterns in the data (Fayyad et al. 1996; Siriet al.2006; Wright 2007). The data warehouse provides a good location for the data mining activities. As such great care must be taken in its development, ensuring that the data in the warehouse as well as the structure of the warehouse correctly represents the area under study, in this case health care. In the former case;

the data can be enhanced for data mining by adding new attributes as well as by judicious aggregation of existing attributes. This has been shown by Balajo& Mark (2001) to result in higher quality knowledge discovery. In the latter case; the structure of the warehouse should be based on a correct dimensional model with careful considerations on the different dimensions or categorizations of the area under study that is to be included in the data warehouse. The warehouse development methodology should also be inclusive of all relevant stakeholders and have support from management since warehousing and data mining is a long term project requiring long term consistent commitment (refer to section on constraints and requirements).

Data warehouse growth is a gradual process involving sequential collection and aggregation of data from different relevant source systems (refer to Fig. 1). A common technique is therefore to develop small 'data warehouses' modeled around a specific business function these are known as data marts. The data marts should have conformed dimensions that enable communication in between the different data marts making up the data warehouse. This ensures that data mining and other analysis can be done across or between different business functions under which each data mart is modeled on. For instance in the case of health care we could analyze patterns between patient medications/prescription and symptoms indicating drug reactions.

There are different data mining techniques and methods in use with continuous introduction of new and refined algorithms for each of the different methods. However for the area of health care some of the more common data mining techniques and methods include the following:

Regression: One of the specific goals of data mining is prediction; the regression technique is paramount to achieving this goal. Using regression a data value can be mapped to a future prediction value (Fayyedet al. 1997; Siri et al. 2006). This property becomes essential in the establishment of patterns between different variables under study. For instance relationship between two different medications that the patient is on can be explored using this technique; furthermore indications on expected direction of treatment using a chosen treatment regimen can be determined.

Classification: This involves the development of a function that assists researchers in mapping a datum to one of the predefined classes. This could be used in patient healthcare by mapping the conditions being depicted by a patient to one of the known health care conditions. This method has a close relation to another method, clustering. Clustering involves the identification of clusters for previously unclassified data basing on their set up of similar attributes (Siri et al. 2006). For instance new diseases can fall into a similar cluster based on their set of similar symptoms, and these in turn could form a new class for study.

*Visualization:*This can be used to discover patterns in medical data sets. This technique uses scatter diagrams and Cartesian planes different attributes can be compared and analyzed.

*Summarization:*Data mining involves interaction with large data sets. Summarization provides compact descriptions of subsets of the data sets involved. For instance it could

be statistical summaries such as mean and standard deviation. The goal here is to derive summaries and rules of association between different data sets under study.

Change and Deviation Detection: This is crucial in discovering the most fundamental changes in the value of data from previously measured or normal values. This can be utilized as warning systems, to predict anomalies in the patient, or even a detection of a potential disease outbreak or epidemic. A sudden increase in the number of patients with a particular disease could also indicate the need for more preventive actions for the community involved. Closely related to this is *Time series analysis* where an attributes' value is examined over a period of time in equal time intervals.

SPECIFIC APPLICATIONS TO HEALTH CARE

Though data mining and knowledge discovery is slowly taking root in health care; its incorporation into business for instance retail stores, super markets, e-commerce and marketing has been faster with great advancements. One can argue that this is because of the demonstrated benefits from use of these technologies, especially the increase in sales. For instance a sales business may use data mining to know their most productive month of sales in a year and capitalize on maximizing their sales, not to mention their highest selling commodity.

Another possible reason for the faster incorporation of data mining into businesses as compared to health care is the fact that with health care data there are other concerns such as privacy, confidentiality and legal issues when it comes to analysis of such data. Even then there have been specific attempts at harnessing the benefits of this new technology in different areas of health care. Below are some of these attempts in brief:

Diagnostics Using Artificial Neural Networks

The use of artificial neural networks (ANN) in medical diagnostics is not uncommon. Artificial neural networks (ANN) attempt to model the cognitive system and neurological functions of the brain. They are thus in position to predict new observations from the past and present observations. ANN employs a type of machine learning algorithm that enables the system to learn new knowledge (Siri et al.2006) by making adjustments to the different mathematical formulas that make up the ANN. Specific applications have been to analyze patient blood and urine samples, study diabetes and to detect conditions like tuberculosis (Lundin , 1998). ANN has also been used in the development of drugs for the treatment of cancer patients.

ANN has been used in the development of diagnostic questionnaires for gastroephageal reflux disease (GERD) (Noya, 2005). This involved the use of a neural network model with one hidden layer to model the relationship between the input variable and the output variable.

Prediction of Patient Conditions

Jonsdottir et al. (2006) reports on the development of a tool, the predictive outcome model for breast cancer; this accurately predicts the 5 year outcome of an incidence of cancer. The tool employs the use of machine learning algorithms to enable the predic-

tion of the patient conditions. Patient information in the database is analyzed and used to determine which class the patient should belong to (classification). This enables the classification of the patient into the different survival groups that exist.

Advanced predictions have been carried out in kidney dialysis patient's survival. This has been reported by Kusiak *et al*. (2005) where a data mining approach is being used in the prediction of survival of patients with kidney dialysis. This information can be crucial in studying the impact of different treatment options on the predicted time of survival. It can also be used in determining the quality of care that should be accorded to a patient with a given time of survival.

Closely related to this is the use of data mining in the study of, maternal vaccination and preterm birth (Ivanka *et al.* 2006). This enabled researchers to study the relationship between the medicines used during pregnancy and their effect on preterm deliveries. This is a crucial issue since preterm deliveries have a huge impact on infant mortality.

CONSTRAINTS AND REQUIREMENTS

Data mining for knowledge discovery involves direct access to data that is under study. When this data happens to be patient medical records, then there are concerns regarding privacy and confidentiality of the patient data. This has generated a lot of debate. However for cases of data mining, where the need is for classification, clustering and other generic studies, the concern is about the relationship between the different patient data under study. Patient details such as names can therefore be encoded and or completely removed from the analysis phase. Even then there are still concerns when data mining goals such as prediction are to be achieved using methods such as regression or time series analysis.

Security is also a concern when dealing with medical data, however there are numerous strong encryption algorithms in use today, these provide sufficient security for the data stored in these data warehouses and databases. When using encryption algorithms to encrypt data in a database a balance needs to be struck between the encryption and decryption process required during analysis. A very strong encryption may result in slower analysis of the information stored in the warehouse (for data mining this can be several thousands to millions of rows) as resources are used in decrypting the information prior to analysis.

Effective data mining for knowledge discovery hinges on the use of electronic medical records. The use of electronic medical records means that the continued population of the warehouse can be done automatically as the health care provider's record the details of their daily interactions with patients. The systems are programmed to continually update the data warehouse with the recently changed or added information from the source systems. This is already being done in areas mentioned earlier where the incorporation of data mining and knowledge discovery has been effected earlier and faster than health care. Therefore those most likely to benefit from data mining are the health care providers who have introduced the use of electronic medical records

in their practices. This is most notable in the developed countries like Sweden with computerization in the health care sector standing at 87 %[4] country wide and at 100% in some areas. In the developing world, like in Africa the use of electronic medical records is still young, scarce and in some cases nonexistent. A gradual move to electronic medical record use will greatly impact the move towards knowledge discovery using data mining.

The development process of the medical data warehouses is an important factor in knowledge discovery since the validity of the identified patterns in data greatly depends on the correctness of the data warehouse and the data contained therein. As such the involvement of all the relevant stakeholders in health care and the intended end users of the data mining systems are very important. Methodologies such as participatory design (Keld et al. 2004) that ensure the participation of all the users should be employed to ensure that the end users are fully involved in the generation of the final system.

CONCLUSION

Data mining for knowledge discovery is a new field gaining a lot of interest in many areas let alone health care. Its application in health care has been mainly restricted to the western world with varying degree of use of electronic medical records in their health care systems. The specific applications are mostly for diseases which affect many people and the electronic medical records have helped accumulate substantial data banks on these diseases. The data banks offer an opportunity for data mining because of the large data sets involved. These diseases while affecting the developed world may not be as serious a problem as other diseases in the developing world such as AIDS, tuberculosis, and malaria. For these diseases, the developing world, especially Africa, have large data sets that they should make use of with data mining. This hinges on the gradual introduction of electronic medical records in health care and development of central data warehouses for collective data mining.

Acknowledgments: We would like to acknowledge the support of Sida/SAREC for this project and the contribution by BTH hospital in Karlskrona especially the IT strategist at the hospital, Thomas Phersson.

4 This data was directly obtained from the Information Technology (IT) Strategist at Blekinge Hospital in Karlskrona

REFERENCES

1. Balajo, R & Mark, W.I (2001). *Exploiting Data preparation to enhance Mining and Knowledge Discovery.* IEEE Transactions on Systems, MAN and Cybernetics.
2. Brodley, C.E., Lane, T., &Stough, T.M. (1999). *Knowledge discovery and data mining*. American Scientist 87(1):54-61
3. Fayyad, U., Gregory, P.S.&Padhraic Smith (1996). *From Data mining to Knowledge Discovery in Databases*. American Association for Artificial Intelligence.Pg 37-53.
4. Fayyad, U., Gregory, P.S. &Padhraic Smith (1997). *Knowledge Discovery and Data Mining: Towards a Unifying Framework*. American Association for Artificial Intelligence.
5. Ivanka, 0-B., Henrik, J., Lone, S. &Jorn, O. (2006). *Maternal vaccination and preterm birth: using data mining as a screening tool.* Pharm World Sci. (29):205-212
6. Jonsdottir, T., Hvannberg, E.T., Sigurdsson, H. & Sigurdsson S. (2008).*The feasibility of constructing a Predictive Outcome model for breast cancer using the tools of data mining*. Expert Systems with Applications 34:108-118
7. Keld, B. Finn, K &Jesper, S. (2004). *Participatory IT Design: Designing for Business and Workplace Realities.* MIT Press
8. Kriegel, H.P., Borgwardt, K.M., Kroger, P., Pryakhin, A., Schubert, M. &Zimek, A. (2007). *Future trends in data mining*. Data mining and Knowledge Discovery 15: 87-97
9. Kusiak, A. Dixon, B. & Shah, S. (2005). *Predicting survival time kidney dialysis patients: a data mining approach.*Computers in Biology and Medicine (8) 431-451.
10. Lundin, J. (1998). *Artificial Neural Networks in outcome prediction*. AnnsChirGynaecol 87: 128-130
11. Noya, H., Menachem, M., Zamir, H., &Moshe, L. (2005). *Applying Data mining Techniques in the Development of a Diagnostic Questionnaire for GERD*. Dig Dis Sci (52): 1871-1878
12. Se-Hak, C. &Steven, H.K. (2004). *Data mining for financial prediction and trading: application to single and multiple markets.* Expert Systems with Applications (26):131-139
13. Smith, K.A. & Gupta, J.N.D. (2000). *Neural networks in business: techniques and applications for the operations researcher.* Computers and Operations research. (27): 1023-1044
14. Siri, K.W., Vasuldha, B. &Harleen, K. (2006). *The impact of data mining techniques on Medical Diagnostics.* Data Science Journal 5:190-126.
15. Wright, P. (2007). *Knowledge discovery in Databases: Tools and Techniques.* Retrieved from-http://www.acm.org/crossroads/xrds5-2/kdd.htmlon 3/11/2007

3.2 Paper II

Dimensional Modeling of HIV Data using Open source

Charles D. Otine, Samuel B. Kucel, and Lena Trojer

Abstract— Selecting the data modeling technique for an information system is determined by the objective of the resultant data model. Dimensional modelling is the preferred modelling technique for data destined for data warehouses and data mining, presenting data models that ease analysis and queries which are in contrast with entity relationship modelling. The establishment of data warehouses as components of information system landscapes in many organizations has subsequently led to the development of dimensional modelling. This has been significantly more developed and reported for the commercial database management systems as compared to the open sources thereby making it less affordable for those in resource constrained settings. This paper presents dimensional modelling of HIV patient information using open source modeling tools. It aims to take advantage of the fact that the most affected regions by the HIV virus are also heavily resource constrained (sub-Saharan Africa) whereas having large quantities of HIV data. Two HIV data source systems were studied to identify appropriate dimensions and facts these were then modeled using two open source dimensional modeling tools. Use of open source would reduce the software costs for dimensional modeling and in turn make data warehousing and data mining more feasible even for those in resource constrained settings but with data available.

Keywords—About Database, Data Mining, Data warehouse, Dimensional Modeling, Open Source.

INTRODUCTION

Data models form the foundation of data warehousing and data mining systems since they help to describe how data is to be represented and acccessed. It is critical that the underlying data model correctly represent the data that is being studied [6], with accurate identification and representation of the required measures and variables. [2] notes that increasing development in the concept of information systems has resulted in interest in data models, since in essence data models form the blue print for the development of databases which is at the backbone of information systems. Database data models such as the flat model, hierarchical model, network model and the relational model have been suggested. The most common of these is the relational model with specific types such as the Entity relational model [1], the concept oriented model and the star and snow flake schemas for data warehouses. [17] refer to other variations of the Entity relationship (E-R) model such as the Multidimensional Entity relationship (ME/R) model, the EVER model and the StarER that combines the star model and the ER model.

A data warehouse is an all inclusive system that enables the extraction of data from different and often heterogeneous source systems [15] and their management in the 'warehouse' to provide user access and analysis. The accessed data can then be data

mined for new information. [16] reports that multi-dimensional data model have proved to be most suitable for data warehouse applications. Muti-dimensional models for data warehouses are generated by using the dimensional modeling technique which is in contrast with entity relationship modeling which aims to generate models that ensure efficiency of record insertion and upates not retrievals like in the case of data warehouses. This fundamental difference in the architecture renders the retrieval of large number of records from E-R model based systems resource intensive and therefore not suitable for data warehousing and data mining that deals with the retrieval of large volumes of data at a time.

[8] notes that detailed guidance for dimensional modeling during the complex data warehousing information systems projects is lacking. Also, [8] indicate that the large and complex nature of data warehousing projects result in difficulties during the design stage. The design stage is made more complicated by the little guidance available for dimensional modeling, with literature available suggesting instead the models suitable for particular situations. Furthermore the dimensional modeling and data warehousing tools are more common, more developed and more documented and reported for the case of 'over the counter[5]' commercial softwares [15]. These softwares prove to be prohibitively expensive for a majority of information system developers who may wish to engage in data warehousing and data mining. This leads to a loss of opportunity for establishements who may have abundant and continuously growing data from taking advantage of data warehousing due to the high costs inolved, especially software costs.

Take the case of sub-saharan Africa, a region most affected by the HIV-virus [20]. This culminates into large quantities of data on HIV infection but little is done to take advantage of this information with a bid to generate new knowledge using data warehousing and data mining. This is further hindered by the high cost of data warehousing and data mining tools availbale in the market and the little information on the cheap and free open source tools. This research paper looks at using open source data modeling tools in developing dimensional models for use in HIV patient data warehousing.

The use of open source is championed because of the high cost of 'off- the- shelf' data modeling and data mining tools and the limited literature on open source modeling tools.

DIMENSIONAL MODELING

Dimensional modeling is used to conceptualize data warehouses which are then implemented using star schemas or snow-flak schemas. It differs from Entity relationship (E-R) modeling that is used for ordinary transaction databases in that it aims to implement a database that eases user navigation [10], enhances performance [4] and interaction thereby improving analysis. Analysis of data in a data warehouse is key to data mining [6]; this is facilitated by the underlying warehouse data model. E-R modeling on the other hand aims to improve ease of understanding by users, enforce

5 Vendors such as Oracle, IBM and Microsoft have developed data warehousing and data mining in their Database Management system commercial tools

consistency and reduce redundancies in the data. With this architecture in mind E-R models are normalized to a large extent and therefore not suited for extensive and complex analysis of data.

Dimensional modeling helps to generate the star schemas. Star schemas are constituted by a fact-table in the centre surrounded by a range of dimensions (Figure 1). The fact table represents a concept of primary interest to the decision maker [5]. The fact table contains attributes known as measures that can be analyzed along different perspectives or dimensions. This assists in giving the data a multidimensional view [13]. Each of the dimensions that connect to the fact table in the centre of the model adds a primary key that acts as a foreign key and forms part of the composite primary key for a row of the fact table. One of the core dimensions of the star schema is the time dimension; this is used to give the information in the data warehouse a lifeline.

The data represented by star-schemas are extensively de-normalized with significant number of redundancies; this architecture improves analysis of the data. This is related to the fewer number of joins required to obtain the results of a query. The snow-flake schema may be interpreted as an extension of the star schema [14]. The reason for this is that the snow flake schema attempts to reduce on redundancies in its architecture by introducing a degree of normalization. Star schemas may be extended to snow-flake (star-flake) schemas when there is a significant increase in the number of rows for a dimension that would impede the performance of the data warehouse. It is due to this that during dimensional modeling, the dimension in question could be normalized to reduce the size of the resultant table in the data warehouse. [14]notes a third schema, the 3NF schema, but contends that it is possible to present the schematics of any application in either of the schemas.

The star schema is considered to be the most efficient design and is suited for modeling data marts. The snow-flake schema may suffer from potential performance issues from the relatively higher number of query joins needed as opposed to the star- schema. Star schemas with more than 1 fact table are commonly referred to as constellations.

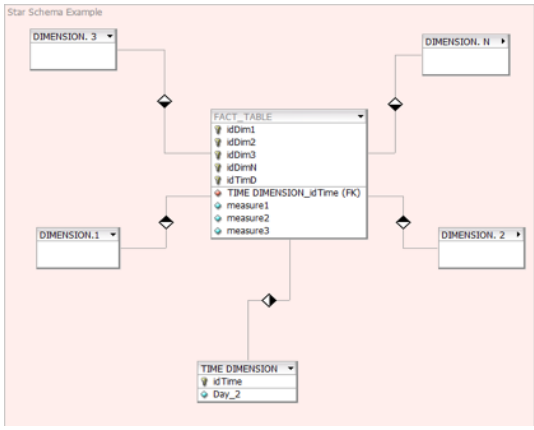


Figure 1: Star Schema

OPEN SOURCE

The cost of commercial software is at times a stumbling block for information systems. This is a lot evident in the moderately young field of data warehousing and data mining. Commercial vendors of different database management systems have developed data warehousing and data mining capabilities for instance Oracle, IBM, and Microsoft. The costs of such software, especially license fees, render the acquisition process prohibitively expensive for the resource constrained settings.

The response to the above has been the development of open source software [3]. Areas that are resource constrained can take advantage of this to acquire information systems that would otherwise be considerably expensive for them. For the case of data warehousing and data mining the use of open source has been scarce and literature on the above limited. However several open source database management systems (DBMS) have come out to compete against the commercial versions. These, to mention a few, include MySQL, PostgreSQL, Firebird, Ingres and Berkeley DB; of this MySQL is by far the most successful.

They (open source DBMS) have however lagged behind in terms of dimensional modeling tools suited to these database management systems. In sections to come, the paper shall draw attention to two open source dimensional modeling tools that were sampled. A key factor for their consideration was the flexibility in terms of the database management system where the dimensional models developed by these tools could be implemented.

HIV/AIDS

The HIV epidemic has affected the countries of sub-Saharan Africa both socially and economically. The HIV virus results in the destruction of the body's immune system rendering it unable to fight off opportunistic infections and therefore resulting in the condition AIDS. Although this has resulted in a large number of deaths it has also offered an opportunity in terms of the data available on HIV. The numerous advances in ICT (data warehousing and data mining) mentioned above can be used to put this data to use. Due to the economic situation in some of these countries like Uganda the use of commercial software, consequential maintenance and sustainability of these data warehousing endeavors may outweigh the benefits of the resultant system. It is important to research and identify alternatives to the high cost of commercial software in the data warehousing process, and why not at the early stages of the data warehousing process, that is, at the dimensional modeling stage.

This paper highlights the development of a dimensional model to support HIV data warehousing using open source. This was done using information from the Ugandan HIV scenario and assistance from different health care partners in Uganda dealing with HIV cases. The government of Uganda in a bid to improve access to antiretroviral therapy (ART) by the infected people has championed the provision of free antiretroviral drugs to patients at Health care centers. Some private non-governmental organizations have also taken the lead in supporting HIV patients. These organizations and

government health centers provide support in the form of ART, voluntary counseling and testing, prevention of mother to child transmission of the virus, medical checkups and adherence monitoring for the patients. It is the information generated by these activities that form the basis of the data to be used for analysis.

MODELING PROCESS

A. Methodology

[12] and [11] recommend that collecting the objectives and requirements should be done by involving the end users. This is the case since organizations have a large spectrum of users with distinct needs to be addressed. Selected government and non-government HIV health centers were visited and the professionals interviewed. The views of some prominent health care givers in HIV were sought.

The dimensional modeling process was articulated in the following phases after collecting the objectives and the requirements.

1. Selection of the appropriate open source dimensional modeling tool(s).
2. Analysis with selected sample of stakeholders to identify the HIV cares process to be modeled.
3. Identification of the dimensions, hierarchies for each fact table.
4. Identification of measures for the fact table.
5. Verification of the technical system.

B. Selection of the Modeling Tool

Two open source modeling tools were identified. The emphasis was placed on modeling tools that allows for connectivity with several database management systems as well as enabling capabilities for database synchronization and reverse engineering. Synchronization allows the tool to generate the corresponding data warehouse dimensions and fact tables from the model directly into the database management system it has connected to. Once the dimensions have been generated in the data warehouse, it is not uncommon for changes to be made directly on the data warehouse. Changes such as definition of new dimensions; attribute additions or removal, new measures in the fact tables can then be directly generated onto the dimensional model; this is known as reverse engineering. The functionality of reverse engineering allows for modifications to the dimensional model from changes to the physical data warehouse dimensions and schemas in the database management system.

The two open source dimensional modeling tools studied were SQL Power Architect and DB designer. Both these tools allows for working with open source database backend as well as commercial database backend including commercial versions search as Oracle, SQL, DB2, and IBM. This would provide a huge flexibility for the dimensional modelers to choose whatever platform was more suitable to the data warehouse design problem in question.

Both tools allow for the definition of the appropriate dimensions with their respective attributes. The relationship and interactions between the different dimensions can

then be defined with the appropriate cardinalities. Figure 2 and Figure 3 indicate the screens for the Power Architect and DB designer tools respectively with sample models.

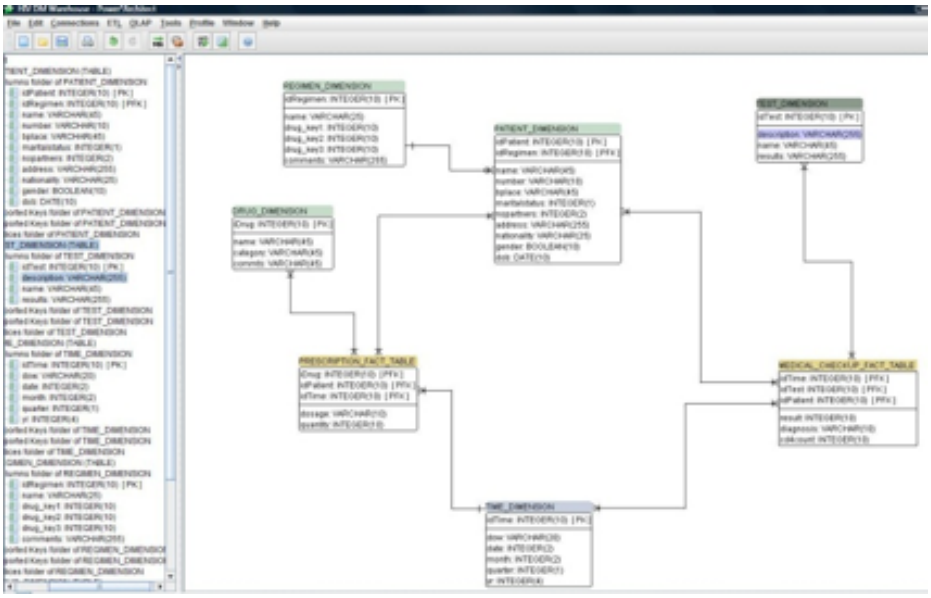


Figure 2: Power Architect Modeling Screen

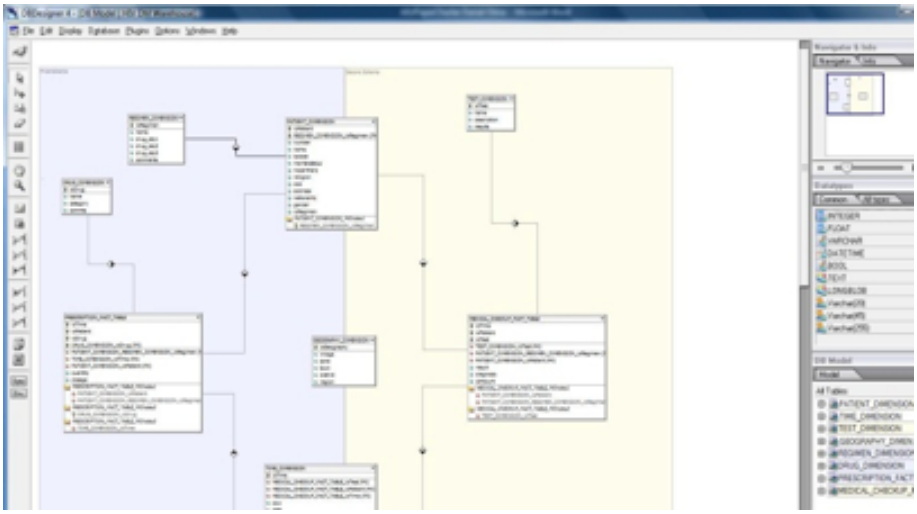


Figure 3: DB Designer Modeling Screen

C. Identification of Process to be Modeled, Dimensions, Hierarchies and Measures

[7] emphasize the importance of selecting the dimensions or features to be used by a data mining algorithm correctly. [11]argues that correctly identifying dimensions and

interrelationships with facts is crucial to coming up with a model that correctly represents users data requirements and the analysis intended on the data. Features that are either irrelevant or unreliable may render the data mining process difficult and make the results complicated to analyze. The objective of dimensional modeling is to represent a set of measurements in a standard frame work. The idea behind this phase is to identify the key process of interest for the HIV health care providers and this was done after repeated consultative sessions with this target group. Analysis of some selected source system[6] from selected HIV health care providers was also done.

Two main processes of interest were noted the i) periodic medical check-ups and ii) Prescriptions to patients. Patients access antiretroviral therapy (ART) from different ART government distribution centers or other NGO assisted ART centers. ART is the treatment of HIV patients with pharmacological agents (antiretroviral drugs) that slow down the progression of the HIV virus in the body. In either of these centers; patients (from here referred to as clients) are given medical checkups at the onset of their ART treatment and periodically when replenishing their ARV drug supplies. Medical checkups may also be conducted in an ad hoc manner whenever the client experiences a relapse of any kind or at the discretion of the care giver. The second process is the prescription; this is given to a client by a physician or medical person in response to a medical checkup or diagnosis that has been done. It may involve the regimen (ART treatment option) that the client is on, or supplementary drugs to assist with opportunistic infections.

The warehouse would then be modeled to monitor the two processes or in this case 'facts' identified above. [8] describes dimensions as entities that are used for analyzing the measurements in the fact table. The dimensions identified include the patient, drug, regimen, test and the time dimensions. In summary the dimensions identified assist in keeping track of what patient underwent what medical checkup and the prescriptions that they were given at what point in time. Items that are being monitored include medical tests that have been done on the patient, the drugs given out as prescriptions and those that make up the patient's ART treatment regimen. The time dimension is necessary to keep track when each of the two processes of interest have been carried out for each patient.

A number of measures were identified for the facts represented by the two processes. The process medical checkup monitors the patients CD4[7] count, weight gain or loss, the tests for opportunistic infections, blood pressure, and pregnancy. The prescription process would monitor measures such as drugs given, the quantity and the dosage dispensed. Dimensional models are extensible because they allow for the addition of new data elements; new facts, dimensions and attributes can be added so long as they are consistent with the present facts. New measures for the facts can be added for increased analytical capabilities.

6 The source systems analyzed include the Adherence monitoring system at reach out Mbuya (an organization that specializes in ART for patients in a Kampala Suburb in Uganda) and Infectious disease institute information system (IDI) in Kampala Uganda.

7 CD4 count is a measure of the strength of the human immune system. HIV continually kills CD4 cells; overtime the body may not be able to replace these lost cells.

Figure 4 indicates the dimensional model generated. This is a constellation with two fact tables and conformed dimensions to enable comparison between the two main processes identified during this stage. This model was generated using the open source modeling tool DB designer. The geography dimension has been normalized from the patient dimension to form a hierarchy along which role up and or aggregation can be done during analysis. Aggregation or role up is done to provide summarized views of the data. It would be important to view the aggregated analysis of each of the two processes, for instance the average CD4 count value for patients in a geographical region on a treatment regimen and an alternate prescription. There are other interesting dimensions that can be analyzed against the time line and tests like the effects of administering ARVs over a period of time, based on the attribute doPTest(date of patient testing positive) in the patient dimension, attributes of time dimension and the different facts in the medical_check_up fact table.

The test dimension enables monitoring of not only the different opportunistic infections but also gives information about pregnancies that could assist with the prevention of mother to child transmission (PMTCT) of the virus. The PMTCT program reduces the risk of mother to child transmission of the virus. It is reported that in the absence of any intervention 15-30% of mothers with HIV will transmit the infection through pregnancy and delivery and others during breast feeding. The response to HIV studies has highlighted ways of reducing this risk one of which is the provision of ART for the HIV positive mothers and the new born babies. The dimensional model offers the opportunity for comparison and optimization on what regimen dimensions for expecting mothers would result in the significant reductions in the HIV virus basing on the test dimension and the largest increments in CD4 count indicated in the medical_check_up fact table. The time dimension would also indicate the most opportune moment to begin the intervention and the progress that is being made during the intervention.

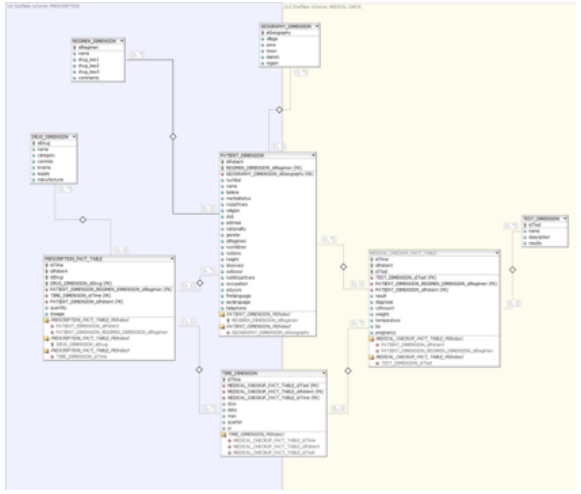


Figure 4: Data warehouse Dimensional Model

D. Verification Technical System

The performance of a model is determined during verification. [9] highlights the different techniques of verification including; general good modeling and programming practices, verification of intermediate simulation outputs, comparison of final simulation outputs with analytical results and animation. The verification process involves checking that the schema is a correct model of the data warehouse. The attributes of the dimensions identified are meticulously cross checked for conformity to requirements, and conformity to the two source systems that were selected for analysis during modeling.

The open source tools selected were flexible in that they offer connectivity with different database management systems both commercial and open source. Part of the verification process was done by connecting to two database management systems MySQL and PostgreSQL as well as trial version of a commercial database Microsoft SQL (MSSQL). The tool allows for the defined models to be generated in the corresponding database management system (synchronization), this was successfully done in all the three database management systems selected.

The reverse engineering aspect was also tested altering the generated data warehouse tables in the different database management system. The changes were successfully reflected in the respective models in the modeling tools.

CONCLUSION

This paper reported on the use of open source tools in building dimensional models for HIV patient information, with the long term result of implementing an HIV patient data warehouse. This is in a bid to reduce on the impact of the high cost of commercial dimensional modeling tools and database management systems in the market and to take advantage of the cheaper open source tools available and the data available on HIV in regions of sub-Saharan Africa such as Uganda.

Star schemas generated using dimensional models are flexible in that they allow for modifications as the data warehouse grows from different data marts organized around the key processes. This is a good property as the dimensional model for HIV patient data would be envisioned to grow as new processes of interest are identified and added to the schema with allowance for new dimensions and hierarchies. This can be done through additions of other new data marts analyzing new processes that are identified with time.

The open source dimensional tools have a weakness in handling complex data types as compared to various new tools that have been researched on and incorporated into some commercial database management systems. These are capabilities to handle complex data types as indicated in [18] and [19]. This would improve on analysis of complex data in open source systems such as patient x-ray screens, brain scans and heart scans. This is still lacking in the open source domain of data warehousing

ACKNOWLEDGEMENT

We would like to acknowledge the assistance of Sida/SAREC, the Swedish research cooperation, Makerere University (Faculty of Technology), Uganda and Blekinge Institute of Technology, Sweden for all the assistance rendered.

REFERENCES

1. Chen, P. (1976). The Entity Relationship model-Towards a unified view of data, ACM Transactions on Database Systems, 1, 1, 9-36.
2. Chilton, M.A. (2006). Data Modeling Education: The changing technology, Journal of Information Systems Educaion, 17,1, 17-20.
3. Coar, K. (2006). The Open source Definition, Retrieved on 18th Nov 2008 from opensource.org: http://www.opensource.org/docs/osd
4. Dash, A.K and Agarwal, R. (2001). Dimensional modeling for Data warehouse, ACM SIGSOFT software engineering notes, 26, 1, 83-84.
5. Golfarelli, M., Maio, D. and Rizzi, S. (1998). Conceptual Design of Data warehouses from E-R schemes, Proceedings of the Hawaii International Conference On System Sciences, January 6-9, Hawaii
6. Gui, Y., Tang, S., Tong, Y. and Yang,D. (2006). Tripple Driven Data Modeling Methodology in Data warehousing: A case study, ACM workshop on Data warehousing and OLAP, 59-66
7. Ilczuk, G. and Wakulicz-Deja, A. (2007). Selection of Important attributes for Medical Diagnosis Systems. *Transactions on Rough Sets*, 7,1, 70-84.
8. Jones, M. E. and Song, I.Y. (2008). Dimensional modeling: Identification, classification and evaluation of patterns. *Decision Support Systems* , 59-76.
9. Kleijen, J. P. (1995). Verification and validation of simulation models. *European Journal of Operations Research*, 82,1, 145-162.
10. Kortinik, M. A. and Moody, D. L. (2003). From ER Models to Dimensional Models: Bridging the Gap between OLTP and OLAP Design. *Business Intelligence Journal*, 8,3, 1-17.
11. Laender H. F., Freitas, G.M., and Campos, M.L. (2002). MD2- Getting Users Involved in the Development of Data Warehouse Applications. *4th International Conference Workshop Design and Management of Data warehouses.* May 27, Toronto, University of British Columbia, 3-12.
12. Lambert, B. (1995). Break Old Habits To Define Data Warehousing Requirements. *Data Management Review* .
13. Malinowski, E. and Zimanyi, E. (2007). A conceptual model for temporal data warehouses and its transformation to the the ER and object-relational model. *Data and Knowledge Engineering*, 64, 101-133.
14. Martyn, T. (2004). Reconsidering Multi-Dimensional Schemas. *ACMs Special Interest Group On Management of Data,* 33,1, 83-88.
15. Nguyen, T. M., Tjoa, A. M., and Trujillo, J. (2005). Data Warehousing and Knowledge Discovery: A Chronological View of Research Challenges. *Springer*, 530-535.
16. Pearson, W. (2008, 1 24). Dimensional Model components: Dimensions part 1. *Retrieved 11 19, 2008, from Database Journal: http://www.databasejournal.com/features/mssql/article.php/3723311/Dimensional-Model-Components--Dimensions-Part-I.htm*
17. Phipps, C. and Davis, K.C. (2003). Automating Data warehouse conceptual Schema Design and Evaluation. Proceedings of the 4th international conference on Design and Management of Data warehouses. May 27, Toronto Canada, 23-32

18. Pokorny, J. (2003). Modeling stars using XML.
19. Riadh, B. M., Omar, B., & Sabine, R. (2004). A new OLAP Aggregation Based on the AHC Technique. DOLAP (pp. 65-71). Washington,DC: ACM.
20. UNAIDS. (2008). 2008 Report on the Global AIDS epidemic. Geneva: *WHO Library Cataloguing-in-Publication Data.*

3.3. Paper III

Implementation of an Open Source HIV/AIDS Data Warehouse in Uganda

Charles D. Otine, Samuel B. Kucel, and Lena Trojer

Abstract:

This paper examines the implementation of a data warehouse for HIV/AIDS patients in Uganda. This implementation is particular focused on a solution applicable to resource constrained settings by using open source software. The system used the open source MySQL database management system as the data store for the data warehouse as well as a staging area for cleaning the data after extraction from different source systems. Scripts were developed basing on the dimensional model for creation of the different dimensions in the data warehouse database and these are presented in this paper. A brief overview of handling population of the time dimension is also presented in addition with two additional scripts showing the loading of the time dimension, as well as the population of the patient dimension using a CSV (Comma Separated Value) file. Finally a flow chat of how adherence is determined using the data warehouse is presented. The paper highlights the focus of MDG 6 on HIV/AIDS management and provision of better access of those affected to treatment. It links this to the use of open source data warehouses to monitor adherence as well as take opportunity of the increasing number of patients accessing antiretroviral therapy whose information can be used in the data warehouse for increase of monitoring effectiveness.

Keywords: Data warehouse, Open Source, MDG 6, HIV/AIDS,

INTRODUCTION

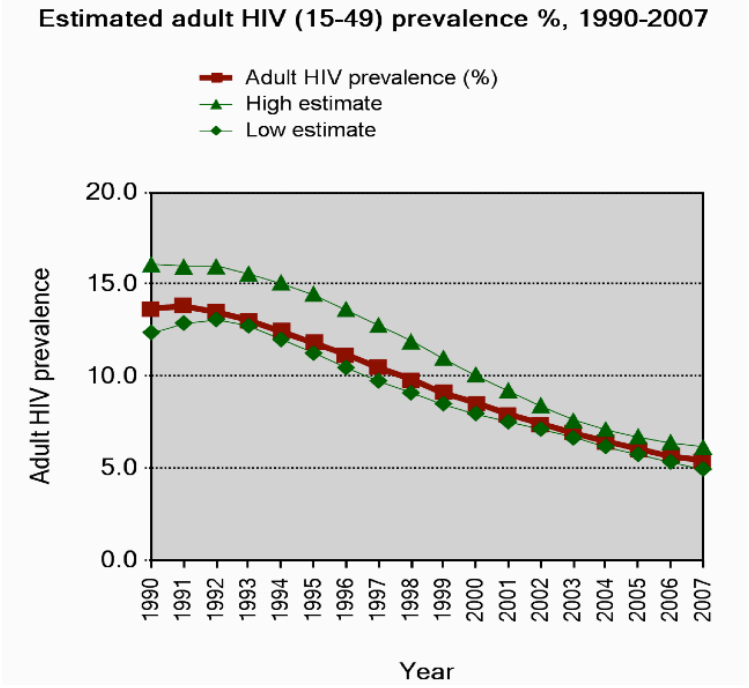
This paper looks at the experience in building an open source data warehouse for HIV data in Uganda. The aim of this research was to develop a platform where HIV/AIDS medical care givers and researchers could use has a base for data mining operations. Uganda is one of the countries whose fight against the HIV pandemic has in the past been praised. This research at a strategic level seeks to leverage ICTs in the fight against HIV aids.

The economic situation in the country and indeed the case for most of the health care providers necessitated a solution that is both affordable and sustainable. This paper therefore showcases the implementation of a data warehouse using open source tools making it affordable for these providers. The results of this research will also contribute to the combined effort of achieving the millennium development goal (MDG) number 6 for Uganda. MDG goal number 6 aims to combat HIV/AIDS, malaria and other diseases (UN 2010).

Background

Uganda in the past has been praised for her success case in the fight against HIV/AIDS. This has been due to multipronged approaches by the government, civil society and the population including a policy of high political support for the fight against the epi-

demic, advocacy for behavioral change communication, interventions to address stigmatization to mention a few. Even with all these praises the HIV infection prevalence rates have remained stagnant with reports placing it between 4% to 8%. This calls for more creative interventions to lower this rate even more or prevent this rate from rising up again has indeed has been in some media [UNAIDS/WHO(Epidemiological fact sheet on AIDS)]



The graph above shows the HIV prevalence rates in Uganda over a 17 year period. Adopted from (UNAIDS/WHO, 2008)

The reduction in the prevalence rate has stabilized between 2000-2005 with reported increases in 2006, which has been a cause for some disparities and scrutiny of this success (Low-Beer, 2002). With a population tipping towards the 31 million mark, Uganda cannot afford a further increase in its prevalence rate. Questions such as, what are the reasons behind this increase in prevalence rates? What is the actual prevalence rate? Why the disparities in prevalence rates? What is the adherence rate? Is a patient adhering? Questions such as these need very clear and accurate answers for AIDS health care provider and researchers. These are some of the questions that a data warehouse with capability for data mining can assist in answering.

The government partnering with development partners has enabled access to antiretroviral therapy (ART) for the population. This has resulted in several ART centers supplying free antiretroviral drugs to the infected people (UNAIDS/WHO, 2008). Ensuring effective therapy has therefore called for careful tracking of patient records

throughout their treatment course at the healthcare centers. Especially in ensuring that the patients adhere to the therapy that is being administered which is crucial to ART treatment (Volberding and Deeks, 2009). This also provides a key opportunity for leveraging ICTs for use in HIV/AIDS by providing data warehouses for patient information.

This paper presents the implementation of an open source data warehouse in MySQL for storage of HIV/AIDS patient's information by HIV/AIDS health care providers in Uganda. The data warehouse was modeled with the future aim of enabling it to support data mining operations by the health care providers to provide additional information to support their therapy activities, an example in point being the adherence to ART therapy.

METHOD

Initial requirements gathering was done through participatory action research with some of the stakeholders in the research. This was carried out through Participatory Action Research (PAR) workshops were selected stakeholders of the research were invited including, policy makers, health care providers and information systems managers for these providers. There were two categories of providers, selected few who had digital patient information systems and a few others who were still using paper based manual systems. Further requirements for the system were identified through one on one meeting with specific HIV/ AIDS healthcare providers. PAR would ensure best practices of project buy-in and ownership, warehouse built incrementally (on concept of data marts), and managing of expectations (Weir et al2003).

Source systems[8] of selected health care Providers[9] were carried out to determine the architecture of the current databases that they were using to hold information on their HIV/AIDS patients. This was carried out for the health care providers that were already using digital databases for their patient records as well as providers that were storing information in paper format. This informed the research on cross cutting important points of analysis for all the providers regardless of their source systems of use.

From the analysis of the source systems from the different providers the dimensional model was generated (Otine *et al* 2010). This dimensional model was the basis for the data warehouse that was implemented. DB professional was the open source tool used to generate the dimensional model. The database management system (DBMS) that was used for the implementation of the data warehouse was the open source DBMS MySQL, using SQL to define the different tables and their interactions. The schematic implementation of the data warehouse was shared with the different providers especially to elicit feedback on the implementation.

RESULTS AND DISCUSSIONS

PAR is a research methodology that is based on a strong desire by the participants and researchers to take a collective effort/collaborate in planning, questioning, reflecting

8 Source systems refer to the current information systems being used by the health care providers in their day to day activities.

9 Providers refer to the organizations providing HIV/AIDS treatment to the infected people

and investigating the key issue that affects them (Cronholm andGoldkuhl, 2004). The implementation of the solution is therefore iterative and leads to more refined solutions that are beneficial to all the stakeholders. The implemented data warehouse was designed with the capability to enable it grow and include options for adding additional models under new business processes. This involved designs based around data marts, each representing an initial business process with the capability to expand this to include additional data marts as needs arise (Breslin, 2004). This ensures a scalable architecture for the system.

This model enabled the interaction between the following tables to be studied more clearly. Queries including the patient table with ARV drug distribution were able to generate important points of information for the researchers. This was then cross checked with the queries against the patient's actual reported drug usage per treatment period.

SQL Scripts for Data Warehouse Dimensions

The following scripts were then run on the MySQL server basing on the dimensional model generated earlier to create the different tables for patient and drug distribution information. The final refined model is given below, followed by the list of scripts used to create the different dimensions in the database server.

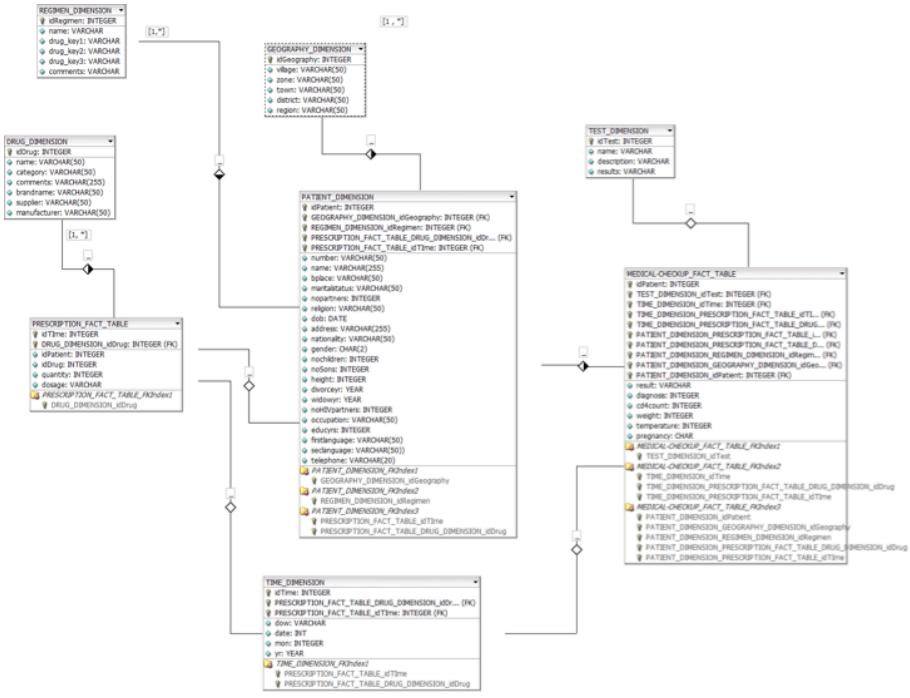


Figure 1: Dimensional Model used to create dimensions in the data warehouse

```
/*Script to Create Data Dimensions in the warehouse*/
CREATE TABLE DRUG_DIMENSION (
idDrug INTEGER UNSIGNED NOT NULL AUTO_INCREMENT,
name VARCHAR(50) NULL,
category VARCHAR(50) NULL,
comments VARCHAR(255) NULL,
brandname VARCHAR(50) NULL,
supplier VARCHAR(50) NULL,
manufacturer VARCHAR(50) NULL,
  PRIMARY KEY(idDrug)
);

CREATE TABLE GEOGRAPHY_DIMENSION (
idGeography INTEGER UNSIGNED NOT NULL AUTO_INCREMENT,
village VARCHAR(50) NULL,
zone VARCHAR(50) NULL,
town VARCHAR(50) NULL,
district VARCHAR(50) NULL,
region VARCHAR(50) NULL,
  PRIMARY KEY(idGeography)
)
TYPE=InnoDB;
CREATE TABLE MEDICAL-CHECKUP_FACT_TABLE (
idPatient INTEGER UNSIGNED NOT NULL,
TEST_DIMENSION_idTest INTEGER UNSIGNED NOT NULL,
TIME_DIMENSION_idTime INTEGER UNSIGNED NOT NULL,
TIME_DIMENSION_PRESCRIPTION_FACT_TABLE_idTIme INTEGER
UNSIGNED NOT NULL,
TIME_DIMENSION_PRESCRIPTION_FACT_TABLE_DRUG_DIMENSION_
idDrug INTEGER UNSIGNED NOT NULL,
PATIENT_DIMENSION_PRESCRIPTION_FACT_TABLE_idTIme INTE-
GER UNSIGNED NOT NULL,
PATIENT_DIMENSION_PRESCRIPTION_FACT_TABLE_DRUG_DIMEN-
SION_idDrug INTEGER UNSIGNED NOT NULL,
PATIENT_DIMENSION_REGIMEN_DIMENSION_idRegimen INTEGER
UNSIGNED NOT NULL,
PATIENT_DIMENSION_GEOGRAPHY_DIMENSION_idGeography IN-
TEGER UNSIGNED NOT NULL,
PATIENT_DIMENSION_idPatient INTEGER UNSIGNED NOT NULL,
result VARCHAR NULL,
diagnosis INTEGER UNSIGNED NULL,
cd4count INTEGER UNSIGNED NULL,
weight INTEGER UNSIGNED NULL,
temperature INTEGER UNSIGNED NULL,
```

```
pregnancy CHAR NULL,
  PRIMARY KEY(idPatient, TEST_DIMENSION_idTest,
TIME_DIMENSION_idTime, TIME_DIMENSION_PRESCRIP-
TION_FACT_TABLE_idTIme, TIME_DIMENSION_PRE-
SCRIPTION_FACT_TABLE_DRUG_DIMENSION_idDrug, PA-
TIENT_DIMENSION_PRESCRIPTION_FACT_TABLE_idTIme,
PATIENT_DIMENSION_PRESCRIPTION_FACT_TABLE_DRUG_DIMEN-
SION_idDrug, PATIENT_DIMENSION_REGIMEN_DIMENSION_id-
Regimen, PATIENT_DIMENSION_GEOGRAPHY_DIMENSION_idGeog-
raphy, PATIENT_DIMENSION_idPatient),
  INDEX MEDICAL-CHECKUP_FACT_TABLE_FKIndex1(TEST_DI-
MENSION_idTest),
  INDEX MEDICAL-CHECKUP_FACT_TABLE_FKIndex2(TIME_DI-
MENSION_idTime, TIME_DIMENSION_PRESCRIPTION_FACT_TA-
BLE_DRUG_DIMENSION_idDrug, TIME_DIMENSION_PRESCRIP-
TION_FACT_TABLE_idTIme),
  INDEX MEDICAL-CHECKUP_FACT_TABLE_
FKIndex3(PATIENT_DIMENSION_idPatient, PA-
TIENT_DIMENSION_GEOGRAPHY_DIMENSION_idGeography,
PATIENT_DIMENSION_REGIMEN_DIMENSION_idRegimen, PA-
TIENT_DIMENSION_PRESCRIPTION_FACT_TABLE_DRUG_DIMEN-
SION_idDrug, PATIENT_DIMENSION_PRESCRIPTION_FACT_TA-
BLE_idTIme)
)
TYPE=InnoDB;
CREATE TABLE PATIENT_DIMENSION (
idPatient INTEGER UNSIGNED NOT NULL AUTO_INCREMENT,
GEOGRAPHY_DIMENSION_idGeography INTEGER UNSIGNED NOT
NULL,
REGIMEN_DIMENSION_idRegimen INTEGER UNSIGNED NOT NULL,
PRESCRIPTION_FACT_TABLE_DRUG_DIMENSION_idDrug INTEGER
UNSIGNED NOT NULL,
PRESCRIPTION_FACT_TABLE_idTIme INTEGER UNSIGNED NOT
NULL,
number VARCHAR(50) NULL,
name VARCHAR(255) NULL,
bplace VARCHAR(50) NULL,
maritalstatus VARCHAR(50) NULL,
nopartners INTEGER UNSIGNED NULL,
religion VARCHAR(50) NULL,
dob DATE NULL,
address VARCHAR(255) NULL,
nationality VARCHAR(50) NULL,
gender CHAR(2) NULL,
```

```
nochildren INTEGER UNSIGNED NULL,
noSons INTEGER UNSIGNED NULL,
height INTEGER UNSIGNED NULL,
divorceyr YEAR NULL,
widowyr YEAR NULL,
noHIVpartners INTEGER UNSIGNED NULL,
occupation VARCHAR(50) NULL,
educyrs INTEGER UNSIGNED NULL,
firstlanguage VARCHAR(50) NULL,
seclanguage VARCHAR(50)) NULL,
telephone VARCHAR(20) NULL,
  PRIMARY KEY(idPatient, GEOGRAPHY_DIMENSION_idGeog-
raphy, REGIMEN_DIMENSION_idRegimen, PRESCRIPTION_FACT_
TABLE_DRUG_DIMENSION_idDrug, PRESCRIPTION_FACT_TABLE_
idTIme),
  INDEX PATIENT_DIMENSION_FKIndex1(GEOGRAPHY_DIMEN-
SION_idGeography),
  INDEX PATIENT_DIMENSION_FKIndex2(REGIMEN_DIMENSION_
idRegimen),
  INDEX PATIENT_DIMENSION_FKIndex3(PRESCRIPTION_FACT_
TABLE_idTIme, PRESCRIPTION_FACT_TABLE_DRUG_DIMENSION_
idDrug)
)
TYPE=InnoDB;
CREATE TABLE PRESCRIPTION_FACT_TABLE (
idTIme INTEGER UNSIGNED NOT NULL AUTO_INCREMENT,
DRUG_DIMENSION_idDrug INTEGER UNSIGNED NOT NULL,
idPatient INTEGER UNSIGNED NULL,
idDrug INTEGER UNSIGNED NULL,
quantity INTEGER UNSIGNED NULL,
dosage VARCHAR NULL,
  PRIMARY KEY(idTIme, DRUG_DIMENSION_idDrug),
  INDEX PRESCRIPTION_FACT_TABLE_FKIndex1(DRUG_DIMEN-
SION_idDrug)
)
TYPE=InnoDB;
CREATE TABLE REGIMEN_DIMENSION (
idRegimen INTEGER UNSIGNED NOT NULL AUTO_INCREMENT,
name VARCHAR NULL,
drug_key1 VARCHAR NULL,
drug_key2 VARCHAR NULL,
drug_key3 VARCHAR NULL,
comments VARCHAR NULL,
  PRIMARY KEY(idRegimen)
```

```
)
TYPE=InnoDB;
CREATE TABLE TEST_DIMENSION (
idTest INTEGER UNSIGNED NOT NULL AUTO_INCREMENT,
name VARCHAR NULL,
description VARCHAR NULL,
results VARCHAR NULL,
  PRIMARY KEY(idTest)
);

CREATE TABLE TIME_DIMENSION (
idTime INTEGER UNSIGNED NOT NULL AUTO_INCREMENT,
PRESCRIPTION_FACT_TABLE_DRUG_DIMENSION_idDrug INTEGER
UNSIGNED NOT NULL,
PRESCRIPTION_FACT_TABLE_idTIme INTEGER UNSIGNED NOT
NULL,
dow VARCHAR NULL,
date INT NULL,
mon INTEGER UNSIGNED NULL,
yr YEAR NULL,
  PRIMARY KEY(idTime, PRESCRIPTION_FACT_TABLE_DRUG_DI-
MENSION_idDrug, PRESCRIPTION_FACT_TABLE_idTIme),
  INDEX TIME_DIMENSION_FKIndex1(PRESCRIPTION_FACT_TA-
BLE_idTIme, PRESCRIPTION_FACT_TABLE_DRUG_DIMENSION_id-
Drug)
)
TYPE=InnoDB;
```

Surrogate Keys and Foreign Keys:

The script above also generates the surrogate keys to define relationships as specified in the dimensional model. These are the IDs (Identification Keys) that are defined in the different dimensions. The surrogate helps to maintain patient history during loading of the data warehouse. This ensures that information on slowly changing dimensions is traced and monitored. For example change in patient geographical residence, or marital status and the linkage to treatment and adherence to medication can be monitored.

Time Dimension and Date Script:

There are three common techniques of populating the Time dimension in a data warehouse; these include i). Pre-population, ii). One-date-everyday and iii). Loading of the dates from the data sources (Darmawikarta, 2007). When using pre-population the time dimension is preloaded with dates for a certain time frame at a go. During loading of data into the warehouse this is automatically linked to the different fact tables that are being compared. The approach of using one-date everyday loads the dates one day at a time for each time; the third option is the loading of the dates from the

data sources which is done when the data warehouse is being loaded with data. The dates are then picked directly from the data sources that are being used to populate the data warehouse. This approach is mostly used when we need to manage the number of entries that are in the data warehouse and in turn the volume of data that we have.

For this implementation the one-date-everyday approach was used. A script was prepared that should be run each day at the moment of loading data into the data warehouse.

```
/*************************************************************************/
/* one_date_everyday.sql      by Otine Charles              */
/*************************************************************************/
USE DATAWAREHOUSE
INSERT INTO TIME_DIMENSION VALUES
(NULL
, CURRENT_DATE
, MONTHNAME(CURRENT_DATE)
, MONTH(CURRENT_DATE)
, QUARTER(CURRENT_DATE)
, YEAR(CURRENT_DATE)
);
```

The script above loads a specific date every day in the data warehouse to be linked into by the different stakeholders. Once this is loaded, entry of data into the warehouse for that date was then automatically linked to this date

Initial Data Warehouse Loading Script:

The following indicates a selection of routines that were developed for operation of the data warehouse. These routines were mainly for the initial loading of the data into the data warehouse from the source systems and the subsequent loading of the patient information. Depending on the source system, the data would first be exported to CSV (Comma Separated Values) format and then loaded using the appropriate script depending on which dimension the data was being input into. In this paper we only present the loading into the patient dimension from a CSV format patient file.

```
/*************************Intial load Script by Otine Charles ********************/
USE DATAWAREHOUSE
LOAD DATA INFILE 'PATIENT_DIMENSION.csv'
INTO TABLE PATIENT_DIMENSION
FIELDS TERMINATED BY ','
OPTIONALLY ENCLOSED BY '"'
LINES TERMINATED BY '\r\n'
IGNORE 1 LINES
(NULL
, number
```

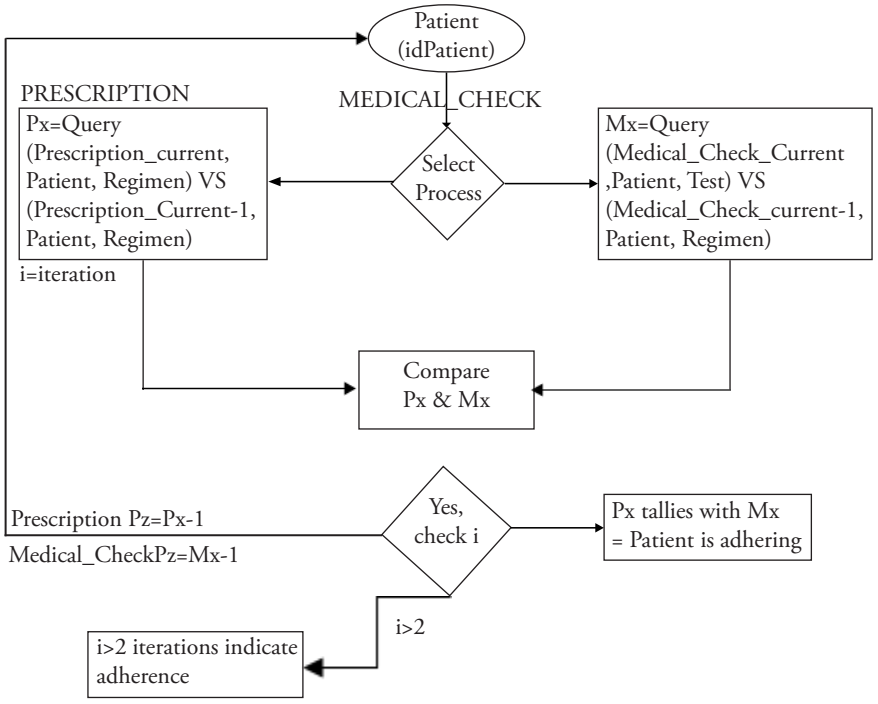
```
,name
,bplace
,maritalstatus
,nopartners
,religion
,dob
,address
,nationality
,gender
,nochildren,noSons,height,divorceyr,widowyr,noHIVpartn
ers,occupation,educyrs
,firstlanguage
,seclanguage
, telephone)
;
INSERT INTO PATIENT_DIMENSION
SELECT
NULL
, number
,name
,bplace
,maritalstatus
,nopartners
,religion
,dob
,address
,nationality
,gender
, nochildren,noSons,height,divorceyr,widowyr,noHIVpart
ners,occupation,educyrs
,firstlanguage
,seclanguage
, telephone
FROM
PATIENT_STG
;
/********************END OF Initial Loading Script*************************/
/*********************************************************************/
```

Each of the different dimensions has a specific script that is run to enable data loading into the warehouse. These were in total 7 scripts, 5 for each of the dimensions and 2 for the fact tables. In each case the script reads the data from CSV format files that have been exported from the source systems. The Time dimension already uses

the one-date-everyday script. From the script we can see that the patient data is captured and loaded from the PATIENT_STG. This is a dimension that is created in a staging area database STG. This holds the data from the source systems on its way to the data warehouse; while in the staging area the data is cleaned and adjusted to the same format from all the data sources. Data in the staging database comes from different data sources and because of this difference each requires an appropriate cleaning method. (Xiang and Min, 2010) Suggest the approach of cleaning the data depending on the category that the data belongs to; either quantitative data or categorical data. Quantitative data in the source systems were for instance patient age, number of children, income, partners, the quantity of prescription, CD4 count, weight, height and other attributes. The Pauta criterion (Xiang and Min, 2010) was adopted to clean the quantitative data in the staging database. Most of the cleaning involved conversion to similar units of measurement. The categorical data required mapping to the correct category names, for example marital status attribute with different source systems referring to it differently.

Prescription and Medical Checkup Flow Chat

When a patient is being monitored for adherence their information is queried against the data warehouse in the following steps.



The patient id is queried against the business processes in this case the two fact tables (Prescription and Medical Checkup fact tables). This involves a comparison from the last two fact information recorded for the patient in question. This then presents the current state Mx of the medical check and the Prescription Px. The prescription and the medical check is then compared. Anomalies in the current and last prescriptions would already flag that dosage of the medication does not match. Additionally the medical checkup also picks on spikes in patient CD4 count, temperature or opportunistic diagnosis and infections such as tuberculosis, herpes simplex, Herpes zoster, Kaposi sarcoma (Avert 2010).

In the event of a non-tally between the prescription and the medical check then a new iteration is started with a prescription value of a previous patient visit. The system maintains the value of the iteration and on the 3rd iteration of a non adherence then the patient is flagged to be a non-adhering patient. System users can then intervene with counseling and corrective actions for the patient.

CONCLUSION

The health care industry has been on the cutting edge of technology and using it to rip benefits commercially (Amit, 1999) in terms of profit as well as socially and morally by helping the sick and giving or prolonging life. From heart, skin and bone transplants to laser treatments the application of technology to health care is ever increasing. The drawback to this however is the cost of investment in some of these technologies and a resultant prohibitive cost to the individual patients in resource constrained settings.

The health industry in such settings can only take advantage of the cutting edge in technology by getting access to cost effective and relevant solutions to their challenges. By presenting the implementation of an open source data warehouse in Uganda this paper highlights the opportunity available to health care providers in monitoring adherence to therapy of HIV patients in an area greatly affected by the disease and limited resources available to invest in high cost technologies.

The Millennium Development Goal (MDG) Report 2010 on goal 6 that focuses on combating HIV/AIDS, malaria and other diseases reports two key issues: The stabilization of the spread of HIV in most regions and an increase in the number of people surviving longer (UN, 2010). This can be attributed to the increase in the number of people accessing antiretroviral therapy by a percentage of 42% between 2003 and 2008 as indicated in the report. To avoid a backward drop in this achievement on access to treatment, different approaches should be used to manage the disease by ensuring adherence to therapy. Sub-Saharan Africa is still a region with many challenges to monitoring therapy let alone providing access to the drugs. These two scenarios therefore provide two opportunities: by employing the use of open source data warehouses, adherence monitoring can be improved in a cost effective manner without the need to spend on license costs for software. Additional the increased number of individuals surviving for longer provides an opportunity to use their therapy information (prescription and medical checks) to inform the data warehouse system in monitoring other patients.

The dimensional model and selected fact tables were focused on mainly two processes, the prescription of during ART and periodic medical checkup carried out during the course of treatment. Further areas of research include additional business processes that may be interesting to the end users. Adjustments to the scripts can also be done to focus on slowly changing dimensions such as patients who change their residence, occupation and salary incomes, and marital status. Further research can then be made on the implications of these changes to their medical conditions and prescriptions. Information systems that are implemented through open source may be affected by the limited support available and in many cases the limitation in documentation (Gacek and Arief, 2004). This was a problem that was experienced during the implementation of this system.

Acknowledgements

We acknowledge the support of SIDA and the School of Graduate studies Makerere University for the funding rendered towards this research. The technical support and linkages made available to the researcher in Sweden by the staff of Blekinge Institute of Technology is also noted and highly appreciated.

References

1. Amit D. (1999). For Pharmaceutical companies a Data Warehouse can just be what the Doctor Ordered-Industry Trend or Event. *Health Management Technology*
2. Avert, Averting HIV and AIDS. (2010). HIV Related opportunistic infections: prevention and treatment. Retrieved on 31st Jan 2010 from http://www.avert.org/hiv-opportunistic-infections.htm
3. Breslin, M. (2004). Battle of the Giants: Comparing the Basics of the Kimball and Inmon Models. *Business Intelligence Journal* Winter 2004
4. Cronholm, S. and Goldkhul, G. (2004). Conceptualising Participatory Action Research – Three Different Practices. *Electronic Journal of Business Research Methods*. 2(2).
5. Darmawikarta, D (2007). Dimensional Data Warehousing with MySQL: A Tutorial. ISBN-13:978-0-9752125-2-0
6. Gacek, C and Arief, B. (2004). The many meanings of open source. IEEE Software Computer Society 34-40.
7. Low-Beer, D. (2002). HIV-1 Incidence and Prevalence Trends in Uganda. *The Lancet* Volume 360:1788-1789
8. Otine C.D. Kucel S.B and Lena T. (2010). Dimensional modeling of HIV Data using open source. World Academy of Engineering and Technology 63. 2010
9. UN (2010). MDG Report 2010. United Nations Department of Department of Economic and Social Affairs (DESA). ISBN: 978-92-I-I02I87
10. UNAIDS/WHO (2008). Epidemiological Fact Sheet on HIV and AIDS. *Core data on epidemiology and response: Uganda*.
11. UNAIDS (2009). AIDS Epidemic Update. Published by WHO. ISBN 978 9291738328
12. Volberding P.A and Deeks, S. (2009). Antiretroviral therapy and management of HIV infection. Lancet 2010; 376:49-62.
13. Xiang, G. and Min, W. (2010). Applying data cleaning in Changqing Oilfields Company's data warehouse. *Second IIATA International Conference on Geoscience and Remote Sensing.IEEE*
14. Weir, R., Peng, T. and Kerridge, J (1999). Best Practises for Implementing a data warehouse: A review for strategic alignment.Proceedings of the *5th International Workshop on Design and Management of Data warehouses 2003 Berlin Germany.*

3.4. Paper IV

Stakeholder Engagement and Participation in Determining Requirements for a Data Warehouse System for HIV/AIDS: Uganda's Experience

Charles D. Otine, Samuel B. Kucel, and Lena Trojer

Abstract

Data warehouse systems face a higher risk of failure than other information system implementation. This is due to many reasons including the fact that data warehouse systems are normally long term implementations. Because of this the need for user acceptance is crucial for the success of such systems. This paper looks how user engagement and participation in developing health data warehouse systems can enhance system acceptance and improve the probability of success of such projects. The paper show cases the development of a health care data warehouse in Uganda to enhance data mining and improve health care provider task of monitoring adherence to medication.

Key Words: User Requirements, Data Warehouse, Database, HIV/AIDs, Health Care

INTRODUCTION

Data warehouse information systems are reliant on the data that is entered into the data warehouse from the different source systems available [11]. In the health care industry the management and ownership of these source systems may vary ranging from government institutions to private institutions. When dealing with a collaborative research project such as a health care data warehouse then it is crucial to involve as many stakeholders as possible in the system lifecycle, giving them the opportunity to learn from each other which is a big determinant of the information system's success probability [3].

Information system projects such as data warehouse systems require adequate planning. This is hampered by the fact that at the onset, the requirements may not be clearly defined which further points to the importance of user involvement [2]. This paper looks at the development of a health care data warehouse on HIV/AIDS in Uganda using open source and the lessons learnt in user engagement and active involvement in refining the requirements for the system. When developing data warehouses based on dimensional modeling [4], we are essentially examining facts about the business processes in question. The user community is therefore very relevant to this because they draw upon the source systems of the organization that they represent [3].They report on the different business processes that they are involved in on a daily basis in their different roles in the organization.

With new systems the cost of change management and training can add to the already high cost of project implementation. When there is end user collaboration in requirements definition and system lifecycle, then organizations may tap into spinoffs such as increased end user knowledge of the system. This is as a result of their involvement

and constant interaction with the system. Additional information system costs such as training may be limited or carried out more quickly due to end user knowledge about the system.

Background

Information system use is ever increasing in almost every aspect of life. This is however hindered by many aspects encountered both after and during the project implementations of these various systems. [1] Report that each year billions of dollars are wasted in failed projects in the developing countries. In some of these cases the wastage is not only financial but time as well due to the investment in user trainings and change management preparations. This further puts in jeopardy the success of future projects due to the loss of confidence by the users who are supposed to embrace the project. Countries like Uganda with limited resources and very many competing priorities for these resources need to avoid such pitfalls by crucially learning from these experiences of failed projects. These situations also call for more caution when defining goal and requirements for an information system project.

Stakeholders and users of the system are very crucial towards mitigating information system failures. For one the end users of the system can fail the system if they just simply reject the final system. Rejection of the system can happen due to a variety of reason including, failure of the system to meet end user requirements, suspicion amongst the end users and users unwilling to let go of competing systems. [5] Categorized these failures into four main categories due to the different reasons behind the failures. The categories include i).Correspondence failure ii). Process failure iii). Interaction failure and iv). Expectation failure. Correspondence failure relate to failures of the system to meet the requirements. Process failure occurs when the system runs over budget, time and or their performance is not satisfactory. A system may be implemented and end up getting hardly used, this type of failure is indicative of an interaction failure. The last type of failure happens when the stakeholder's expectations cannot be met. This paper examines how stakeholder involvement was used during the implementation of the data warehouse system to mitigate against the correspondence and expectation failures.

The data warehouse system has different end users with different expectations of what they need from the system. From government, to medical care providers, researchers to patients, all these should end up with a solution that most closely matches their needs and expectations. Some of these expectations may be conflicting with other stakeholder expectations; these require negotiation and consensus building. The HIV/AIDS data warehouse systems depend on large quantities of data and information that is collected from different sources with the data warehouse being more useful with increasing information. This in essence hinges the success of the data warehouse system on stakeholder acceptance of the system otherwise we could essentially end up with system interaction failure.

The fight against HIV and AIDs has progressed significantly in the last decades with the introduction of antiretroviral therapy (ART) using antiretroviral drugs (ARVs).

These help to slow down the disease progression. In Uganda the government in partnership with non-governmental organizations and development partners has tried to provide antiretroviral treatment to the affected in the community but a large number of the affected still do not have access [10]. This has resulted in a number of ART centers where HIV positive patients can access services such as treatment, counseling, testing and so on. The amount of data collected by these centers of treatment offer an opportunity for data warehousing and data mining. These ART centers were crucial in helping to determine requirements for the data warehouse. In addition to these they helped to determine the business process.

METHODOLOGY

Literature on different techniques and experiences in user engagement in information system implementation was examined from the association for computing machinery (ACM) digital library. This includes journal articles, conference and workshop proceeding and book publications. These and other documents were collected and analyzed. This helped to inform the initial discussion with selected stakeholders to enable them visualize the potential of an HIV/AIDS data warehouse.

A selection of stakeholders of the data warehouse was made including: Antiretroviral therapy (ART) health care provider like nurses, pharmacists, doctors and councilors, Civil society organizations dealing with AIDS, antiretroviral drug manufacturing company, government, researchers and AIDS service organizations. Selected ART providers source systems were also examined, both the manually based systems as well as few with basic computerized systems ranging from simple spreadsheets to database management systems. The research worked with Uganda network of AIDS support organizations (UNASO); this assisted in coordinating and building solidarity with Uganda AIDS organizations.

Participatory Action Research (PAR) methodology was employed primarily because it ensures the participation of all the users [7] and puts into consideration different stakeholders view points. PAR is holistic and allowed the use of additional tools and techniques as the project was conducted. A participatory workshop was used to bring all the different stakeholders together to focus the direction that the project should take while keeping in mind their different needs for the system. During the participatory workshop focused grouping was used to determine the key needs and expectations of the different stakeholders from the system.

Individual selective semi structured interviews and one on one meeting were scheduled with selected users. Screenshots and the dimensional model of the system were shared with selected users to get feedback on user expectations of the system. These were carried out iteratively to refine the needs and manage expectations. The interviews helped to elicit the stakeholders experience and perceptions in system development. This offered the interviewees a chance to offer their specific insights about the project and also discuss their fears and worries.

USER REQUIREMENTS

Common Understanding and Alignment

During the participatory workshop to help refine requirements the users were introduced to the concept of data warehousing and the opportunity for data mining available from the integration of the different process information that they had accumulated from their daily activities. The importance here was to generate a common understanding of what data warehousing involved by all the stakeholders. This enabled the different users to appreciate the fact that on an individual basis they were limited in terms of data warehousing but as a result of collaboration there was a lot of potential. Developing a common understanding of the problem faced by the stakeholder also assists in alignment of the stakeholders to a common goal.

Expectations and Requirements

The stakeholders were then divided into groups. These were requested to come up with their expectations of the system. This was discussed in a plenary session and all results compiled as depicted in the table below.

Stockholder	*Expectations of the system*
Government/Development partners (Policy level):	• AIDS policy and decision making (Planning and Budgeting) • Control over the system/Monitoring • Provision of better service to citizens • Facts and figures to source for funds • Guidelines for policy making • A guide on drug procurement (nationally)/ based on adherence to medication • Establish standards in HIV/CARE • Proper record keeping • Evidence based policy formulation • Mobilization of resources
Medical Personnel: Nurses, Pharmacists, Counselors, Doctors	• Health Improvement of clients/patients • Client tracking and adherence to medication (2) • Improved record keeping • Good decisions on appropriate drug combinations (2) • More informed decisions • Provision of better care, better diagnosis • Timely access to information and further research opportunities

Patients:	• Receipt of better care • Receive standardized and specialized treatment. • Improved diagnosis. • Improved ART Management and service delivery • Improved patient follow up on adherence
NGOs and Donors	• Better Research and improvement in Monitoring and Evaluation. • Better data collection and filling in the gaps and loopholes • Basis for improved funding. • Receive feedback on the effectiveness of their programs and assist in interventions in areas of their strategy. • Ease of project monitoring • Auditability of funding to patients especially on ART provision. • Program Evaluation and Fund allocation

From the above the issue of monitoring, diagnosis, informed decisions and standardization came out. Some of the key expectation were liked, for instance the governments need to plan procurement nationally being independent on records showing prescriptions of ARTs across the board and their usage. This is also linked to adherence, in the sense that if patients are adhering to therapy then this would be an indication of actual usage of the drugs and planning for procurement and budgeting can be based around this. While this was not immediately incorporated into the dimensional model [8] and data warehouse [9], future revisions could incorporate a cost dimension for the drugs. Reports on this would assist in budgeting and planning and policy formulations at the national level.

The expectations on better care, improved diagnosis and decisions on treatment came out as an expectation for both the medical personnel as well as the patients. This highlighted the importance and need for continuous medical checkup, testing and tracking of the patients. With the information in the data warehouse this could be tracked individually as well as in categories and groups to view common patterns.

Concerns and Threats

Some of the stakeholders raised concerns regarding the sharing of data. Some ART offering centers had partnered with foreign based research organizations and institutions. These claimed right of ownership to the patient data and would not share this in a common pool such as a data warehouse. The workshop enabled a forum to initiate discussion on possible negotiation on how sharing of the information could be achieved. Data warehouses properties such as their read only quality after they are loaded with data from the source systems was shared. Those with concerns regarding possible changes to their data were then made aware that essentially with the data warehouse they would be providing a copy of their data. Which meant their original data remains unique to them and in the form that they require. At the same time the data is cleaned and transformed into a form that could be integrated with the other stakeholders. The benefit of this collaboration was highlighted to them in that com-

mon processes such as adherence monitoring could be improved by their collective use of the system.

One on one interviews also raised stakeholder concerns on the issue of becoming irrelevant to their current roles. This is due to fears that the data warehouse would replace their role. This again required open communication and more understanding of what the data warehouse actually involves. The data warehouse essentially depends on their continued role and should not be viewed as a replacement of their role. The data warehouse feeds on data from the source systems [10] therefore the stakeholders' different roles become ever more critical towards the survival of the data warehouse.

Review and Realignment

Determining the initial vision of a data warehouse project is only the first step towards implementing the project. The more process involves the initial alignment of the user requirements to the vision of the project and a frequent revisit of the requirements and expectations to ensure non deviation. Failure to do this means creates an opportunity for correspondence, interaction and expectation failure. With novel ideas such as is the case with data warehouse projects, loss of confidence to the project can be a big blow towards the project's success.

Beginning the discussions on fears and challenges in an open manner and with every stakeholder involved ensures that the potential fears and threats are brought to light and dealt with openly. This improves the chance of the project success and acceptance by the end users.

CONCLUSION

The project highlighted the importance of a HIV/AIDS data warehouse being able to assist in managing adherence. [4] advocates for incrementally building the data warehouse based on the concept of data marts and conformed dimensions. This can be linked to the PAR methodology and through collaborative efforts the stakeholders can continually refine their requirement needs from the system. The system can be developed incrementally as the requirements are refined and new business processes added to the dimensional model.

The engagement of the stakeholders initially ensures that training needs for the end users of the system are more manageable than training users who are hearing about the system for the first time. The stakeholders who are continually involved in the process of developing the system are more aware about the system and provide a good basis for feedback and enhance the system growth.

Acknowledgements:

We thank SIDA for funding the project through Makerere University Directorate of Graduate studies. We also acknowledge the input from BTH and the faculty of technology Makerere University. In a special way we thank the ART centers that participated in the participatory workshop including reach out Mbuya, Infectious Disease

Institute, Mild May, and other partners such as AIDS information center for coordinating the stakeholders.

References:

1. Dalcher, D. and Devin, L.(2003). Learning from Information System Failures by using narrative and ante-narrative methods. *SAICIST '03 Proceedings of the 2003 annual research conference of the South Africa Institute of Compute Science and Information technologies on Enablement through technology.* ISBN:1-58113-774-5
2. Freitas, G.M., Laender, A.H.F. and Campos. M.L. (2002). MD2- getting users involved in the development of data warehouse application. *In Proc. Of the 4th International Workshop on Design and Management of Data Warehouses,* Pages 3-12.
3. Gallagher, K., Mason, R.M. and Vandenbosch, B. (2004). Managing the Tensions in IS Projects: Balancing Alignment, Engagement, Perspectives and Imagination. *HICSS 8:80254a, Proceedings of the 37th Annual Hawaii International Conference on Systems Sciences.*
4. Kimball, R. and Ross, M. (2002). *The Data warehouse Toolkit: The Complete Guide to Dimensional Modelling 2nd.* John Wiley & Sons, Inc. New York, NY, USA ISBN:0471200247
5. Lyytinen, K. and Hireschheim, R. (1987). Information Systems failures: A Survey and Classification of Empirical Literature. *Oxford Surveys in Information Technology*.4:257-09
6. Lugan-Mora, S. and Trujillo, J. (2003). A Comprehensive Method for data warehouse design. In Proc. Of the *5th International Workshop on Design and Management of Data warehouses*.
7. O'Brian, R. (1998). An Overview of the Methodological Approach of Action Research. Retrieved on 10th January 2010 from http://www.web.net/~robrien/papers/arfinal.html
8. Otine, C.D, Kucel, S.B and Trojer, L. (2010). Dimensional Modeling of HIV Data Using Open Source. Published in the *Proceedings of World Academy of Science, Engineering and Technology* Issue 63. March 2010
9. Otine, C.D, Kucel, S.B and Trojer, L. (2011). Implementation of an Open Source HIV/AIDS data warehouse in Uganda. Submitted to *African Journal of Information Systems.*
10. UNAIDS/WHO (2008). Epidemiological Fact Sheet on HIV and AIDS. *Core data on epidemiology and response: Uganda.*
11. Zhang, X., Ding. L. and Rundensteiner, A. (2004). Parallel multisource view maintenance. *The VLDB Journal* (2004) 13:22-34

PART IV – CONCLUSIONS

4.1. Summary of Papers

Paper I introduced the concept of knowledge discovery and data warehousing in health systems. It presented the opportunity available for Uganda as a country with many affected by the AIDS epidemic and with years of information on treatment and success stories in some cases. While developing this paper the status of health information systems in Uganda was survey. This enabled the documentation of their strengths, weaknesses and opportunities. The paper recommended the introduction of electronic medical records, databases and data warehouses for the data available cost effectively using technologies appropriate to the situation. A multi-stakeholder approach to the research was recommended due to the different stakeholders involved in HIV and AIDS service provision in Uganda.

Paper II Developed a dimensional model for data warehouse focusing on HIV/AIDS patients. The tool reviewed the available open source dimensional modeling tools available. These were few and with limited functionality as compared to their off the shelf counterparts. Different dimensions for the dimensional model are identified and proposed and finally modeled using both tools with one tool being recommended. The model generated was one focused on monitoring patient adherence to ART by focusing on two processes: prescription information and medical checkup information. The limitation of the tool in handling complex data types is highlighted. The focus on a star-schema based dimensional model is recommended since it ensures that the system can be incrementally built with the addition of new functionality and requirements based on data marts.

Paper III reports on the development the data warehouse using open source. This builds on the previous paper as it implements the dimensional model recommended. An algorithm that monitors adherence is proposed. This operates on the patient information in the database from prescriptions and medical checks performed during continuous patient monitoring. The paper also highlights the linkage with MDG goal 6 that aims to combat the spread of HIV, malaria and other diseases. This is critical in Uganda where resources are limited and many competing priorities. The implementation using open source limits on the need for off-the-shelf software and prohibitive costs required for license fees.

Paper IV is concerned with the definition of user requirements by ensuring stakeholder engagement and participation. This provides three crucial reasons for this, the first being the need for user acceptance of the system given that the level of the data warehouse success is dependent on the number and involvement of different stakeholders in providing and sharing their patient information. The second involves the need to limit training requirements and costs of using the system as much as possible given the resource constrained aspect of Uganda as the research area. Therefore involvement of the end users and stakeholders from the onset not only assists in clearly defining the goals and requirements of the system but results in stakeholders who are more aware and knowledgeable about the system. Training needs for such users is therefore easier to accomplish. Furthermore engagement and collaboration based on PAR ensures that stakeholders get to learn from each other. The third reason for engagement is to ensure system ownership and prevent rejection by the end users.

4.2. Concluding discussions

Uganda has significant health information but it is fragmented, weakly organized and largely paper based. The Ministry of health (MOH) is responsible for the management of this information however this is lacking, with limited focus on the importance of the information. This was evidenced from the status of information systems in the country. The efficiency of reporting and analysis of critical information is greatly reduced by this fragmentation. This may result in loss of life when simple diagnosis of patient treatment progression and failure during ART is missed. This also affects planning and allocation of resources for government institutions due to limited information on past usage by the medical institutions.

On a more basic level the sustainability and access of the data in paper form is greatly limited. This is crucial as we move to the era of significant information and data sharing and its advantages.

The incidence or infection rate of HIV is quite high for the sub-Saharan African countries including Uganda. This presents an opportunity in terms of the data on patients that can be analyzed for patterns using data mining. This can only be made possible by reorganizing the fragmented data and conversion to a more analysis friendly format such as a data warehouse. Dimensional modeling tools facilitate the development of data warehouses that can be used for data mining. These dimensional modeling tools

are few with a large number being commercially based with stringent license requirements and costs. In resource constrained settings the accessibility of such tools is a major challenge. This research highlights open source tools that can be freely accessed to enable the dimensional modeling and development of appropriate data warehouses. Data mining can then be more easily carried out on this information to generate new knowledge on the disease. Open source dimensional modeling tools such as decision studio professional or data architect can be used at cheaper costs to model data warehouses for study. These tools are versatile in allowing integration with different database management systems both commercial and open source that can be used to house the data warehouse modeled. Open source database management systems such as MySQL and PostgreSQL can be used for the data warehouse. The research therefore helps in addressing the MDG 6 goal of combating HIV, malaria and other diseases.

There were challenges encountered in accessing support for the products and access to documentation was limited for some of the tools used in the research. Since most of these tools are free there is limitation on documentation (Stol and Ali Babar, 2010) and support for end user needs. Therefore a need for adequate training for those intending to use these tools. Additionally the functionality of some of the open source tools is not as advanced as the functionality of their commercial counterparts. Bugs and errors in some of these products are tedious to correct given the budget limitations on these products. The conversion of data from manual paper based system to electronic information in data warehouses was a constraint because of different interpretations and storage of the information with the different health care providers. This took considerable time during the process of extracting the data from the source systems and cleaning it before loading into the data warehouse.

The research methodology, modified PAR was found to be essential for this research given the need for project to be relevant, functional, and sustainable and in developing a feeling of ownership amongst the users. If all the stakeholders develop a feeling of ownership for the system then relevance, sustainability and use of the system will be greatly enhanced. This process is greatly facilitated by using the PAR methodology that ensures inclusion of all stakeholders involved during the entire project. Building trust in groups of stakeholders is not easy with different people having different views and interests on the direction that the project should take. Building consensus amongst all these teams is crucial towards the success of the project and the system. For instance some medical centers linked to research were not comfortable sharing their data with others. This reduces the impact of analysis since data mining is more effective with larger data sets.

4.3. Statement of Scientific Contribution and Originality

The development of an HIV/AIDS data warehouse dimensional model that can be used to monitor patient adherence to medication can be contributed to this research. This model can be technically implemented in different data warehouse capable data-

base management systems with different users adjusting it to their own needs as they identify new dimensions to include to this model.

The research also introduces an algorithm for monitoring adherence from the database through patient prescription information and the continuous medical checkup using the different patient tests. CD4 count values and drug prescription disbursement are compared to classify a patient as either adhering to the medication regimen being allocated or not. Non adherence is flagged and caregivers can then seek new methods of intervention.

Uganda is a country with limited resources and very high demands from the health sectors in terms of human resource, equipment and medication. This research provides a demonstration of the development and implementation of an otherwise costly project in resource constrained settings. This situation is mirrored across many countries in sub-Saharan Africa. This research may also be used to address similar challenges in ART such as adherence in these regions. This is important since the MDG report indicates that most sub-Saharan Africa is still the most heavily affected region. This research provides one of the synergies in the acceleration of goal 6 of the MDGs.

The research also introduces a modified PAR approach for data warehouse development. This includes phases of system design, development and testing. This approach is applicable to the data warehouse systems in health with need for refinements and additions to models that are developed.

4.4. Way Forward

The research methodology PAR supports the identification of additional processes to monitor and their inclusion as additional data marts into the system, giving the warehouse a "growing" property. In effect as the research progresses new problems can be identified and added to the model and the resultant effects from the reports studied. This would mean new data marts focusing on the new processes. Additions to the dimensions can also be explored for example giving the patient dimension spatial properties depending on their location. This will also create opportunity for new data mining algorithms and routines to be developed to address specific stakeholder requirements.

As the system grows there is need to address the training needs of all the stakeholders. This will also ensure continuity as with new additions to stakeholders. Further areas of research will involve the development of training modules on using the system especially in light of the fragmented and different information systems used by the different stakeholders.

The linkage to strongly support policy at the strategic level in terms of government, donor and NGO planning can be explored. This would mean moving from predominately prescription information in the data warehouse to cost implications and forecasting use requirements basing on the number of additional patient requirements

on monthly, quarterly and annual basis. Since there are already dimensions based on geographical location, the addition of spatial data can be used to compare the location of health care centers and treatment facilities with a patient's level of access to these locations.

References

Cabena P., P. H. (1998). *Discovering Data mining : From Concepts to Implementation.* NJ: Prentice Hall.

Connolly, T. a. (2002). *Database Systems: A practical Approach to Design, Implementation and Management.* Addison-Wesley.

Fayyad, U. G. (1996). From Data mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence* , 37-53.

Fayyad, U. G. (1997). Knowledge Discovery and Data Mining: Towards a Unifying Framework. *American Association of Artificial Intelligence* .

Hand D., H. M. (2001). *Principles of Data Mining.* Cambridge: MIT Press.

Hippel, E. V. (2005). *Democratizing Innovation*. Massachusetts: MIT Press Cambridge.

Huang, M. C. (2007). Integrating data mining with case-based reasoning for chronic disease prognosis and diagnosis. *Expert Systems with Applications*, 856-867.

Inmon, W.H (1996). The data warehouse and data mining. *Communications of ACM*. Volume 39(11)

Ivanka, O.-B. J. (2006). Maternal Vacination and preterm birth: using data mining as a screening tool. *Pharm World Sci*., 205-212.

Jean, M. (2002, November 7). Action Research for Professional Development. Retrieved November 7, 2007, from jeanmcniff.com: http://www.jeanmcniff.com/booklet1.html

Jonsdottir, T. H. (2006). The feasibility of constructing a Predictive Outcome Model for breast cancer using the tools of data mining. *Expert Systems with Applications*, 108-118.

Joy, M. W. (2005). *The Microsoft Datawarehouse Toolkit*. Pearson Education.

Kimball, R. and Ross, M. (2002). *The Data warehouse Toolkit: The Complete Guide to Dimensional Modelling* 2nd. John Wiley & Sons, Inc. New York, NY, USA ISBN:0471200247

Kriegel, H. B. (2007). Future trends in data mining. *Data mining and Knowledge Discovery*, 15, 87-97.

Kusiak, A. D. (2005). Predicting survival time kidney dialysis patients: a data mining approach. *Computers in Biology and Medicine*, 431-451.

Larose, D. (2005). *Discovering Knowledge in Data: An introduction to Data Mining*. John Wiley & Sons, Incorporated.

Laudon, K. a. (2004). *Management Information Systems*. Pearson Education.

Lundin, J. (1998). Artificial Neural Networks in outcome prediction. *Anns Chir Gynaecol*, 128-130.

Lyle, H. a. (1995). Using management Information Systems to enhance health care quality assurance. *Journal of Management in Medicine*, 9 (1), 27-36.

MIT (2001). Technology Review:Emerging Technologies That Will Change The World. Published by MIT. Retrieved from http://www.technologyreview.com/Infotech/12265/ in January 2007

Noya, H. M. (2005). Applying Data mining Techniques in the Development of a Diagnostic Questionnaire for GERD. *Dig Dis Sci*., 1871-1878.

O'Brian, R. (1998). An Overview of the Methodological Approach of Action Research. Retrieved on 10th January 2010 from http://www.web.net/~robrien/papers/arfinal.html

Riel, M. (2007, November 8). Understanding Action Research. Retrieved November 7, 2007, from Center for Collaborative Action Research: http://cadres.pepperdine.edu/ccar/define.html

Siri Krishan Wasan, V. B. (2006). The Impact of data mining techniques on Medical Practice. *Data Science Journal*, 5, 119.

Smith, K. a. (2000). Neural networks in business: techniques and applications for operations research. *Computers and Operation Research*, 1023-1044.

Stol, K.J and Ali Babar, M (2010). Challenges in Using Open Source Software in Product Development: *A review of Literature. ACM* 978-160558-978

Tan, K. Y. (2003). Evolutionary computing for Knowledge discovery in medical diagnosis. *Artificial Intelligence in Medicine*, 129-154.

Turban, E. a. (2001). *Decision Support Systems and Intelligent Systems*. India: Prentice Hall of India.

UN (2010). *MDG Report 2010*. United Nations Department of Department of Economic and Social Affairs (DESA). ISBN: 978-92-I-I02I87

Wadsworth, Y. (1998, November 7). What is Participatory Action Research. Retrieved November 7, 2007, from http://www.scu.edu.au/schools/gcm/ar/ari/p-ywadsworth98.html

Wright, P. (n.d.). Knowledge discovery in databases: Tools and Techniques. Retrieved 10 30, 2007, from Association for Computing Machinery: http://www.acm.org/crossroads/xrds5-2/kdd.html