

HIV Patient Monitoring Framework through Knowledge Engineering

Charles Daniel Otine

Blekinge Institute of Technology Doctoral Dissertation Series

No 2012:12

ISSN 1653-2090

ISBN 978-91-7295-241-6

HIV Patient Monitoring Framework through Knowledge Engineering

Charles Daniel Otine



School of Planning and Media Design
Department of Technology and Aesthetics
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
Technoscience Studies
Telecommunication Systems

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

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

© Charles Daniel Otine 2012
School of Planning and Media Design
Department of Technology and Aesthetics
Graphic Design and Typesetting: Mixiprint, Olofstrom
Publisher: Blekinge Institute of Technology
Printed by Printfabriken, Karlskrona, Sweden 2012
ISBN 978-91-7295-241-6
um:nbn:se:bth-00540

Acknowledgements

This thesis would not have been possible without the assistance and guidance of individuals who in one way or another contributed in part to this study. First and foremost I give thanks to God and I am exceptionally grateful to my Swedish supervisor Prof. Lena Trojer and Ugandan supervisor Dr. Samuel Baker Kucel for giving me the guidance and patience through this PhD journey.

I acknowledge the support of the College of Engineering Design Art and Technology (CEDAT) at Makerere University and the Department of Digital Media and Aesthetics at Blekinge Institute of Technology (BTH) in Sweden. Appreciation goes to all the staff in the two departments.

I acknowledge the support extended to me by the following people during the course of my study; Associate Professor Barnabas Nawangwe Principal of CEDAT, Professor SS Tickodri-Togboa, Associate Professor Mackay Okure, Dr. Peter Lating, Paul Kent, Dr. Peter Giger, Silvio Ocasic, Anita Carlsson, Madeleine Persson, Rebecka Molin, Pirjo Elovaara, Fredrik Gullbrandson, Jonas Svegland and Dr. Niklas Lavesson from BTH. I also acknowledge the support and discussion with my fellow PhD students at BTH, Makerere University and University of Dar es Salaam.

The support from Sida, Makerere University and BTH is recognized and greatly appreciated.

Lastly I appreciate the support of my family and the two women role models in my life, my mother Korina and big sister Consolata. My family for praying for me and supporting me through the travels and long periods in Sweden Palma, Joe, Andrew, Jacinta, Peter, Francesca and lastly the first two Doctors in the family Gloria Otine and daddy. You all helped lay the foundation for this study.

This thesis is dedicated to my Grandmother who died in 2011 before the work on this was completed. It is also dedicated to my dear uncle and great role model, Peter Oryema who did a lot for those affected by HIV in Uganda at the onset of the epidemic. Lastly it is dedicated to the more than a million who lost their lives to HIV in Uganda.

Abstract

Uganda has registered more than a million deaths since the HIV virus was first officially reported in the country over 3 decades ago. The governments in partnership with different groups have implemented different programmes to address the epidemic. The support from different donors and reduction in prices of treatment resulted in the focus on antiretroviral therapy access to those affected. Presently only a quarter of the approximately 1 million infected by HIV in Uganda are undergoing antiretroviral therapy. The number of patients pose a challenge in monitoring of therapy given the overall resource needs for health care in the country. Furthermore the numbers on antiretroviral therapy are set to increase in addition to the stringent requirements in tracking and monitoring of each individual patient during therapy.

This research aimed at developing a framework for adopting knowledge engineering in information systems for monitoring HIV/AIDS patients. An open source approach was adopted due to the resource constrained context of the study to ensure a cost effective and sustainable solution. The research was motivated by the inconclusive literature on open source dimensional models for data warehouses and data mining for monitoring antiretroviral therapy.

The first phase of the research involved a situational analysis of HIV in health care and different health care information systems in the country. An analysis of the strengths, weaknesses and opportunities of the health care system to adopt knowledge bases was done. It proposed a dimensional model for implementing a data warehouse focused on monitoring HIV patients. The second phase involved the development of a knowledge base in form of an open source data warehouse, its simulation and testing.

The study involved interdisciplinary collaboration between different stakeholders in the research domain and adopted a participatory action research methodology. This involved identification of the most appropriate technologies to foster this collaboration. Analysis was done of how stakeholders can take ownership of basic HIV health information system architecture as their expertise grows in managing the systems and make changes to reflect even better results out of system functionality.

Data mining simulations were done on the data warehouse out of which two machine learning algorithms (regression and classification) were developed and tested using data from the data warehouse. The algorithms were used to predict patient viral load from CD4 count test figures and to classify cases of treatment failure with 83% accuracy. The research additionally presents an open source dimensional model for monitoring antiretroviral therapy and the status of information systems in health care. An architecture showing the integration of different knowledge engineering components in the study including the data warehouse, the data mining platform and user interaction is presented.

KEY WORDS: Action Research, Antiretroviral therapy (ART), AIDS, Classification, Databases, Data Mining, Data Warehousing, Health, ICT, IT, HIV, Open Source, Knowledge Discovery, Knowledge Engineering, Machine Learning, Regression.

Thesis Structure

This doctoral thesis is structured in 3 parts.

Part I: Presents an introduction to the area of study, specifically dealing with a background to HIV/AIDS in Uganda, the problem being studied and the objectives of the study. It further examines the theoretical framework of the research and the Methodological approach adopted for this research.

Part II: Presents the results of the research. This is presented in the form of papers prepared and submitted for publication in conferences and journals. It further presents additional results that may not have been covered in the submitted papers.

Part III: The last part of thesis presents the discussion of the results, the contributions of the research as well as the areas for further research. The last chapter is the conclusion chapter that answers the research questions, provides recommendations from the research and finally presents areas for improvement and further research.

List of Acronyms

ABC – Abstinence, Being Faithfull and Using Condoms
AIC – AIDS Information Centre
AR – Action Research
ART – Antiretroviral Therapy
ARV – Antiretroviral Drugs
DFID – Department for International Development
DHS – Demographic Health Survey
DHS2 – District Health Information System
EMR- Electronic Medical Records
GIPA – Greater Involvement of people living with AIDS
GOU – Government of Uganda
HIV- Human Immunodeficiency Virus
ICT – Information Communication Technologies
IT – Information Technology
MACA – Multisectoral Approach to Control of AIDS
MOH- Ministry of Health Uganda
MoICT – Ministry of ICT Uganda
NSP – National HIV/AIDS Strategic Plan
PAR – Participatory Action Research
PD – Participatory Design
PMTCT – Prevention of Mother to Child Transmission
PRA – Participatory Rural Appraisal
QCIL – Quality Chemicals Industries Limited
RRA – Rapid Rural Appraisal
TASO – The AIDS Support Organization
UAC – Uganda AIDS Commission
UK – United Kingdom
UNAIDS – Joint United Nations Program on HIV/AIDS
USA – United States of America
USAID – United States Agency for International Development
WHO – World Health Organization

Table of Contents

Acknowledgements	v
Dedication	vii
Abstract	viii
Thesis Structure	ix
List of Acronyms	x
Table of Contents	xi
List of Figures	xiv
List of Tables	xv
List of Equations	xv
Part I	1
1 CHAPTER 1 INTRODUCTION	3
1.1 Background	3
1.1.1 History of the Epidemic	3
1.1.2 Response to the Epidemic	5
1.1.3 Definition of Key Concepts	7
1.1.4 HIV and Uganda's Health Care System	9
1.1.5 Client ART Lifecycle	11
1.1.6 Information Communication Technologies	13
1.1.7 Data mining and Data Warehousing	16
1.2 Statement of the problem	18
1.3 Objectives	18
1.4 Research Questions	19
1.5 Scope	19
1.6 Ethical Considerations	19
1.7 Justification of Study	20
1.8 Concluding Remarks	21
2 CHAPTER 2 THEORETICAL ANALYSIS AND STANDARDS	23
2.2 Information Technology in the Ugandan Context and Health Care	23
2.3 Data mining and Knowledge Discovery Databases	27
2.4 Data Mining techniques and Applications in health care	29
2.4.1 Regression	29
2.4.2 Classification and Clustering	30
2.4.3 Visualization	30
2.4.4 Summarization	31
2.4.5 Artificial Neural Networks (ANN)	32
2.4.6 Health Care Cases involving Data Warehousing and Data Mining	32
2.5 Technology Considerations and Requirements	33
2.6 Ethical Challenges in Health Care Data Mining	34
2.7 Status of Health Information Systems in Uganda	35
2.8 Human Resource Requirements	37

2.9 IT Training Needs	38
2.10 Concluding Remarks	39
3 CHAPTER 3 METHODOLOGICAL CONSIDERATIONS	41
3.1 Action Research	41
3.2 Participatory Action Research	43
3.3 System Process	46
3.2.1 Phase I: System Design	46
3.2.2 Phase II: Implementation and testing	47
3.2.3 Phase III: Stakeholder Training, Maintenance and Growth	47
3.4 Concluding Remarks	48
Part II	49
4 CHAPTER 4 PRESENTATION OF PAPERS	51
4.2 Introduction	51
4.3 Paper I	53
4.4 Paper II	61
4.5 Paper III	72
4.6 Paper IV	87
4.7 Paper V	94
4.8 Paper VI	103
4.9 Summary of Papers	113
5 CHAPTER 5 ADDITIONAL RESULTS	115
5.1 Introduction	115
5.2 Additional Results	115
5.2.1 Overview of System Architecture	115
5.2.2 Sample Screen shots	121
5.2.3 Materialized Views	124
5.3 Matrix implementation of Linear Regression Algorithm	126
5.4 Supervised Learning for HIV Treatment Failure	128
Part III	131
6 CHAPTER 6 DISCUSSIONS	133
6.1 Summary of Main Finding	133
6.2 Conditions for Knowledge Engineering Systems in ART	134
6.3 Dimensional Model for HIV/AIDS	134
6.4 Viral Load from CD4 Count History	135
6.5 Patient Treatment Failure	135
6.6 User interaction and participation in fostering learning, growth and sustainability	135
6.7 Decision Support at Policy and Operational Level	136
6.8 Contributions	136
6.8.1 To Science and Academia	136
6.8.2 To Policies and Strategies	137
6.8.3 To ART stakeholders	137
6.9 Challenges and Weakness	138

7 CHAPTER 7 CONCLUSIONS AND RECOMMENDATIONS	141
7.1 Introduction	141
7.2 Conclusions	142
7.3 Recommendations	144
REFERENCES	147

List of Figures

Figure 1-1: National Backbone Infrastructure (Adapted from MoICT)	15
Figure 2-1: Categories of visualization techniques (Keim 2002)	31
Figure 3-1: Progressive Problem solving with Action Research (Riel 2010)	42
Figure 3-2: PAR as a component of Action Research	44
Figure 3-3: System Lifecycle process adopted from (Kimball & Ross 2002)	47
Figure 4-1: An adjustment of the Knowledge Discovery process (Fayyad et al, 1996)	55
Figure 4-2: Star Schema	64
Figure 4-3: Power Architect Modeling Screen	66
Figure 4-4: DB Designer Modeling Screen	67
Figure 4-5: Data Warehouse Dimensional Model	69
Figure 4-6: Estimated adult HIV (15-49) Prevalence % 1990-2009 (UNAIDS)	73
Figure 4-7: Dimensional Model used to create dimensions in the data warehouse	75
Figure 4-8: Prescription and Medical Checkup Flow Chat	84
Figure 4-9: IPM Development Process	98
Figure 4-10: Summarized view of Dimensional Model	105
Figure 4-11: Process of addition of new dimension	109
Figure 4-12: Changes to the process fact tables	110
Figure 5-1: System Architecture	116
Figure 5-2: Security Options	121
Figure 5-3: Patient Reporting Screen and ARV Statistics	122
Figure 5-4: Patient Medical Report and Regimen Options	122
Figure 5-5: Dimensional Manager for Ensuring Adding Data warehouse Dimensions	123
Figure 5-6: Wiki System Screen shot	123
Figure 5-7: Medical Check Fact Snapshot Table Prior to Feature Scaling	124
Figure 5-8: Medical Check Fact Snapshot post applying Feature Scaling	125
Figure 5-9: Graph of Cost of using Algorithm & Iterations	128
Figure 5-10: Classification for ART treatment failure	129

List of Tables

Table 1-I : Provision of ART by government facilities	10
Table 1-II : CD4 level to commence treatment	12
Table 1-III: WHO clinical staging and immunological criteria for initiating ART	12
Table 4-I: Stakeholder Expectations	90
Table 4-II: User Categories in the Wiki System	97
Table 4-III: Patient Dimension Record	106
Table 4-IV: Patient Record Changing Dimension (Method 2)	106
Table 4-V: Patient Record Changing Dimension (Method 3)	106

List of Equations

Equation 1: Prevalence	9
Equation 2: Projection formula for HIV patient viral load	127
Equation 3: Cost Formula for finding optimization values	127
Equation 4: X and as Matrices	127
Equation 5: Cost Function for Classification	128
Equation 6	128

Part I

CHAPTER 1

INTRODUCTION

1.1 Background

Our capability and ability to scan the environment and other conditions, question it and learn from it is a critical reason behind developments in our world today. The environment around us offers many opportunities for us to learn from as we examine the different problems that we face, case in point the HIV (Human Immunodeficiency Virus) that is at the center of this research. The knack of collecting data available about different situations and to analyze and make sense of it is crucial to the process of dealing with world challenges. This dissertation is about the interdisciplinary research of using information technology (IT) concepts of knowledge engineering in managing HIV patient monitoring one of the challenges of antiretroviral therapy (ART) in Uganda.

The multipronged approach used to deal with the epidemic in the world over has led to the accumulation of lessons learnt and information related to different aspects of the disease. This provides the opportunity to take this accumulated information a step further by using it to build knowledge-based systems that leverage the advances in information communication technologies (ICTs). The complex analysis capabilities available in the field of knowledge engineering can be used on the knowledge bases to facilitate the solution of complex problems during patient therapy requiring high level of human expertise, critical analysis and quick turnaround time for results.

1.1.1 History of the Epidemic

The timeline for the HIV virus before conclusively documented reports of the virus is broad with some reports going back to the 1930s. Pence (2007) reports on the first potential case of a death due to HIV virus strain HIV-1 in Congo, (present day Democratic Republic of Congo) neighbor to Uganda in 1959. Later in 1960 strains of the virus HIV-2 is said to have been transferred to people in the Guinea Bissau. It

is believed that the virus was first transferred to the USA (United States of America) around 1966 by some Haitians who were known to have been working in Congo and could have contracted the virus from this region. It is later in 1975 that symptoms and other reports of wasting began to be reported in residents of Africa. In 1980 the first AIDS (Acquired Immune Deficiency Syndrome) case was noted and reported in the USA, with the patient code named "Patient Zero". The controversies and counter arguments on the theory of patient zero notwithstanding the HIV virus has been a fact and has heavily affected different corners of the world. Starting from Africa where the virus is said to have originated to America where it was first heavily documented and the virus finally got the match stick that set it on fire in the media.

For the case of Africa, the number of deaths has been staggering to say the least with UNAIDS (Joint United Nations Program on HIV/AIDS) and WHO (World Health Organization) putting the figures between 2.2 and 1.8 million people per year from 2005 to 2010 (WHO, 2012). The death toll in Uganda is estimated to be at about 1.77million people at the end of 2010. The social and economic impacts from this have been devastating. The death toll has left significant number of orphans. A study conducted in 2005, showed that of the 2.18million orphans in Uganda, an estimated 1,009,345 (approximately 46.4%) were caused by HIV (Hladik et al. 2008).In Uganda the first diagnosed case of AIDS was in 1982, but prior to that doctors had become aware of a disease which caused severe loss of weight amongst the affected, this was locally referred to as the "slim disease". The disease was accompanied by several opportunistic infections such as tuberculosis, fungal infections such as candidiasis, viral infections such as herpes zoster and HIV associated malignancies such as Kaposi's sarcoma to mention a few. By 1986 the country was in the midst of a major epidemic with reports of the prevalence of the disease being up to 29% in certain urban areas (Bond, 1986; Carswell & Lloyd 1987; Sewankambo et al. 1994; Nalugoda et al. 1997).

Uganda formed the first AIDS country program in 1987 with a three pronged approach to the epidemic: 1) the campaign ABC: Abstinence, Being faithful and Using Condoms; 2) Screening of blood transfusions and 3) HIV surveillance. The Multisectoral response also involved the establishment of grass root organizations such as TASO (The AIDS support Organization) an NGO (Non-Governmental Organization) originally run by a 16 volunteers who had been personally affected by HIV/AIDS. TASO went off to become one of the largest indigenous AIDS service organizations providing both medical and emotional support to many HIV positive people. Many other grass root organizations tackling the epidemic were also established including AIC (AIDS Information Centre), GIPA (Greater Involvement of people living with AIDS), Mild May, Kitovu mobile home Care, THETA to mention a few. Presently the coordination of all these different organizations is being overseen by UNASO (Uganda Network of AIDS service Organizations).

Between 1992 and 2001 Uganda registered remarkable drops in the epidemic prevalence rates from high levels of 19% and even 30% amongst pregnant women to 5% in the population. This has been attributed to a number of reasons including the strong leadership and the multipronged approach to the epidemic. However it is also noted

in some circles that the lack of effective treatment and medication meant that a significant part of the population perished due to the virus. During this period, specifically in 1998 the government of Uganda began to run a test to study the feasibility of providing free ART.

Since 2004 the government of Uganda has had in place a program for the provision of free antiretroviral drugs (ARV). Some experts have argued that this has led to the complacency about HIV and AIDS. There has also been a shift away from the previous ABC campaign towards the abstinence program backed by some donor countries. The focus on comprehensive sex education and condom promotion is no longer mainstream and this may have led to some risky behaviors. As a result there have been reports of increased prevalence of the disease.

Initially few people in sub-Saharan Africa had access to ART, primarily because of the prohibitive costs of the dosage due to the international patents that stopped them from being manufactured at a cheaper cost. After 2001 drug manufacturers in the developing countries began producing generic drugs under special terms in international law. This further greatly improved the access to ART in the world and parts of sub-Saharan Africa like Uganda. Generic drugs are the identical copy (bioequivalent) of a brand name (or proprietary) drug. They are exactly the same as their counterparts in dosage form, safety, and strength, route of administration, quality, performance characteristics and intended use. India is one of the largest suppliers of generic ARVs to low and middle income countries with other major producers being Brazil, Thailand and South Africa. A small number of African nations have also established local HIV drug manufacturing facilities including Zambia, Ghana, Tanzania, Kenya and Uganda. In Uganda the government assisted with the establishment of Uganda's Quality Chemical Industries Limited (QCIL), producing triple-combination of ARV drugs.

The manufacture of generic ARV drugs was a turning point for the accessibility of the drug and helped to address the treatment of those infected in poor resource settings. In spite of the significant reduction in the cost of the dosage for HIV patients, still a large percentage of the infected are not on treatment. In the case of Uganda, the statistics indicate that only 43% (UNAIDS, 2010) of people in need of ART are on the treatment. Though for the case of Uganda this indicates that more needs to be done, this figure is better than the 11% accessing the treatment in sub-Saharan Africa overall.

1.1.2 Response to the Epidemic

The government of Uganda (GOU) responded by openly addressing the epidemic which was affecting the country. At the policy level the GOU formed the Uganda AIDS Commission (UAC) in 1992 and this operated national strategic frameworks to counter the epidemic and above all provided strategic leadership. UAC provided and continues to provide the overall coordination, monitoring and evaluation of HIV and AIDS related activities in the country. UAC developed the Multisectoral Approach to the Control of AIDS (MACA). Based on this a national AIDS policy was drafted. A national HIV/AIDS Strategic Plan (NSP) running from 2007/08-2011-12 was developed, in addition to a health sector strategic plan.

The national health sector strategic plan (2010) spells out Uganda's priority in comprehensive and evidence based prevention interventions. The goal of the NSP is to ensure universal access to HIV/AIDS prevention, care and treatment with a specific target of reducing the HIV incidence rate by 40% by 2012. The NSP additionally aims to improve the life of people living with HIV/AIDS (PLWHA) by doing the following:

1. Reducing the health effects of HIV/AIDS.
2. Mitigating social, cultural and economic effects of HIV/AIDS at the individual, household and community levels.

To achieve the two aims highlighted above requires improvement to the coverage of ART. The donor communities under different funds and projects assisted Uganda financially to combat the epidemic. The sources of funding for HIV/AIDS fight include donations from national governments, multilateral funding and private funding. This includes more recently the USA under the Presidents Emergency Plan for AIDS relief (PEPFAR) with the main goal of saving lives of those suffering from HIV around the world and the UK government's Department for International Development (DFID). Multilateral funding organizations such as the Global Fund have supported the country with funding to combat the epidemic even if this support has been marred with corruption scandals involving the funds. Private sector funding comes from corporate donors, individual philanthropists, religious groups, charities and NGOs.

In Uganda 95% of the national response to ART is covered by donor funds, and the country still continues to rely heavily on this to cover its ART needs for the future (Bamuturaki 2008). In spite of this, challenges continue to be faced in the supply chain management of the distribution leading to shortages in facilities, expiry, and overstocking in some cases. Staffs in the health facilities have reported difficulties in forecasting the ARV needs of the facilities given the lack of data and systems as well as insufficient knowledge of safe drug substitution (Bamuturaki, 2008; Windisch et al. 2011). Planning for the needs of the ART center is therefore a challenge as analysis of usage over a period is not done quickly and in a timely manner.

The government encouraged the policy of HIV testing and counseling. The first HIV testing and counseling center became available in Uganda in 1990, later on the national policy on testing and counseling was developed in 2003. All though this has expanded to many areas in the country there are still some challenges in accessing the service. Furthermore the problem of transmission from mother to child of HIV cannot be overemphasized. This is a serious issue in Uganda at the moment as figures indicate that 76% of new infections are due to sex whereas 22% of the new infections are as a result of mother to child transmission. As such in 2000 GOU developed a prevention of mother to child transmission (PMTC) policy. The PMCT focuses on prevention, family planning, the provision of antiretroviral prophylaxis and care and support to the affected.

1.1.3 Definition of Key Concepts

The following are some key theoretical concepts and terms relevant to this research and used through different sections of the thesis. The terms and concepts are presented in alphabetical order. Some definitions may use terms and or concepts that are themselves defined later. Some terms are defined here and may have already been used before.

Adherence: This is the level to which an HIV patient follows the treatment prescribed to them. This can be calculated as a percentage of the prescribed medication that the patient correctly takes and follows. For instance if a patient is allocated 10 tablets to be taken over a period of 10 days then their adherence is 100% if they take the 10 tablets correctly each day over the 10 days at the times prescribed.

AIDS: 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.

ART: This is the treatment of individuals infected with the HIV virus with pharmacological agents known as ARV drugs to slow the progression of the HIV virus. These drugs inhibit the progression of the disease.

ARV: These are pharmacological agents that are given to HIV/AIDS patients to slow down the progress of the disease. The ARVs do this by inhibiting production of the retroviruses and thereby block the replication of the virus.

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.

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, Piatetsky-shapiro, et al. 1996).

Data Warehouse: 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). This is sometimes known as knowledge discovery databases.

Dimension: This is a specific category of information about the subject under study. Dimensions 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.

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.

HIV: Human immunodeficiency virus. This is the virus that causes the disease AIDS (A.D.A.M 2011).

Generic drugs: Generic drugs are replicas of original brand name drugs. They are exact bioequivalent of the brand name drugs in that their strength is the same. They are used the same way and have the same quality and perform and behave the same way as the generic brand name drugs. The difference is that they cost significantly cheaper compared to their counterpart brand name drugs. Manufacturers of generic drugs do not need to spend the costs involved in researching and developing the initial drug, since they are copies of drugs already developed and tested.

ICT: This refers to the integration of telecommunications, IT, broadcast media, communications media, applications and other contemporary devices to meet different needs or purposes.

IT: The use of computers, networking, hardware and software programming, telecommunications and other equipment to assist with information capture, storage, retrieval, processing, presentation and protection.

Incidence rate: This is an epidemiological term that measures the occurrence of new cases of the disease (Roe and Doll 1995). Incidence rate can be calculated as the number of new cases of the disease in a specified period (usually a year) divided by the size of the population under consideration who were initially disease free.

Knowledge Engineering: The process of structuring, organizing and formalizing data and information available on a particular problem to construct a program or algorithm that can perform a difficult task quickly and adequately.

Monitoring Information System: In the context of this research a monitoring system refers to that system that continually tracks down a dimension. WHO (2006, pp11) in its guidelines on HIV patient monitoring and ART places states that “*patient monitoring is the routine collection, compilation and analysis of data on patients over time and across service delivery points, using information directly from paper forms or entered into a computer*”. During routine visits data such as patient demographics, treatment history and care, contact information and so on is collected by the monitoring system and analyzed over the therapy period. The system may give warnings in face of an anomaly in the dimension, or an indicated value.

Prevalence rate: This is the proportion of people in a population who have a particular disease at a specified point in time, or over a specified period of time. In this regard the prevalence rate in Uganda is the proportion of people in Uganda who have the HIV virus. “*Prevalence measures how much of some disease or condition there is in a popula-*

tion at a particular point in time” (Roe and Doll 1995). In summary prevalence can be calculated by the following formulae, expressed as a percentage:

$$\text{Prevalence} = \frac{\text{all new and pre-existing cases during a period}}{\text{Population during the same time period}} \quad \text{Equation 1: Prevalence}$$

1.1.4 HIV and Uganda’s Health Care System

The last 30 years have seen tremendous changes in the HIV epidemic and different responses to the virus taking place, ranging from an epidemic where little was known and huge deaths to ART and clinic trials of potential HIV vaccines. The landscape has improved and continues to improve with the recent announcements of approval of drugs that reduce the chances of infection (Kasozzi & Agencies 2012). The need for analysis and close monitoring of the different interventions and the patients on these different interventions is now more important than ever.

The 2010 UNAIDS report indicates that the number of new HIV infections have dropped by 19% since the high values of 1999 (UNAIDS 2010). This is due to the range of interventions being used to combat the HIV virus. Even so, a big reason for this is the introduction of ART and the focus the international community has placed on it, given its inclusion as an MDG. Examining this more closely, ART means reduction in the viral load of individuals enabling them to live longer, healthier and more meaningful lives. When the viral load in an HIV patient is low, then their chances of infecting others is also significantly reduced and directly reduces on incidence. Consequently the larger the number of PLWHAs we can get onto ART the better as this directly links to the set that can potentially infect others. This highlights the importance of ART in the fight against HIV.

The GOU has moved from providing ART as a pilot project to providing ART to close to 250,000 (UNAIDS 2010) individuals affected by the HIV virus in comparison to the more than 1 million PLWHA in the country (GOU 2010a, pp 10) who will eventually also need to access treatment sooner or later. From this data more is required to bring about improved access of ART and management of treatment to those affected and undergoing treatment. There is a need to not only address the virus but also work on preventing further infections. The focus has been on the prevention of infection first and the treatment of those affected, these two in turn acting as a cause and effect to each other. By treating those who are infected we indirectly work on the prevention of infections to those not infected. Given the importance of ART to the HIV problem, it is imperative to address its challenges and ensure conditions that enhance results for ART programs.

1.1.4.1 Uganda’s Health Care system

ART is administered through Uganda’s health care system. Uganda’s health care system is based on the referral system, with a bottom up approach and patients being transferred to more specialized treatment in higher level centers in cases when the lower level treatment centers fail to handle the patient’s needs. Ideally the most basic health

centers are referred to as the health center II (HC II). These deal with the needs of a parish in the Ugandan context and are supposed to be equipped with a midwife and an enrolled nurse and should be able to provide basic clinical care like antenatal clinics.

When dealing with ART, antenatal clinics are important as they offer the first point to deal with prevention of mother to child transmission (PMTCT) of the virus. The next level health facilities are the health center III (HC III) and these provide medical facilities to the sub-counties in Uganda. HC III is supposed to be equipped with basic laboratory facilities and a clinical officer amongst other staff. At the county level or parliamentary constituency level we have the health center IV (HC IV). The most striking capability of a HC IV is the fact that they are supposed to have a functional surgical theater and a doctor with wards for patient admission. HC IV is essentially a small hospital in the Ugandan context.

Beyond this we have the district hospitals that receive referrals from the different HC IV in the district. The district hospitals further refer to the regional referral hospitals of which there are 11 in the country. Finally there are 2 national referral hospitals namely Mulago Hospital and Butabika Hospital. Even with these planned infrastructures many of these health facilities are understaffed, underequipped and poorly constructed in some cases. The services provided in these health facilities is not comprehensive and far in service delivery from the planned range of services that they were established to provide.

The table below shows the coverage of ART by government health centers and hospitals.

Health Centre/Facility (Total Number in Country)	Number Providing ART	Percentage
National Referral Hospitals (2)	2	100%
Regional Referral Hospital (11)	11	100%
District General Hospitals (98)	98	100%
Health Centre IV(166)	130	78%
Health Centre III (905)	40	4%
Health Centre II (1887)	2	0.1%

Table 1-I : Provision of ART by government facilities

(MOH Report as quoted by GOU 2010a ,pp 33)

Aside from the government facilities Uganda also has a range of private health facilities supported by different groups of individuals and organization. The organizations include NGOs, religious groups, and private individuals or health specialists. The religious organizations have a number of health facilities around the country. Some of these health facilities form the main source of health care for the people in remote regions of the country. Examples include [Bwindi community hospital](#) in western Uganda, [Matany Hospital](#) in North East of the country and [Lacor Hospital](#) in the North of the country. There are also other private hospitals and institutions like [Reach Our](#)

[Mbuya \(where the research is based\)](#), Rubaga hospital, Nsambya hospital, Mengo and others. These health facilities have been providing ART to clients over the years. Each of these having started providing the therapy at different points since ARVs became available in the country.

The health centers begin by offering voluntary counseling and testing to the clients/patients¹ who visit the ART center. Once the counseling is given and the client approves and gives consent, testing can then be carried out and depending on the outcome the client is counseled more on how to avoid risky behavior, or more information is provided on how the client can live positively.

When clients are placed on ART adherence to the therapy is very important. A research carried out by US through USAID (Livesley et al. 2008) found that in a sample of ART providing centers approximately half of the ART centers monitored adherence using only one method of monitoring. The manual technique of monitoring adherence using patient monitoring cards that use the results of the adherence by checking times and dates when drugs are taken was being used by only 13% of the health centers. Meaning that majority of these ART centers were either not monitoring therapy or not carrying out adequate levels of monitoring. Furthermore, the effort required in monitoring adherence across different methods and the number of patients involved does not provide incentives for effective adherence monitoring. Given the importance of adherence, there is dire need for a carefully thought solution for monitoring adherence and treatment progression.

1.1.5 Client ART Lifecycle

Voluntary counseling and testing (VCT) was adopted initially during care of HIV clients. VCT is an approach based on the concept that testing must be initiated by the client who wish to know their HIV status. However because of challenges of late diagnosis, testing initiated by providers have also been adopted for HIV testing in health unites. This has been adopted in many clinics that offer routine services like antenatal clinics and in clinics treating sexually transmitted diseases. The patient is offered the option of testing for HIV and the clients have the choice of not taking the test.

At antenatal clinics, testing is important because it can help with the issue of PMTCT. Based on MOH (2009) and WHO (2006) guidelines when a new client visits an ART center like reach out, the following typically happens:

1. A center counselor available is tasked with speaking to the client about their potential condition. The client is counseled and told about the HIV testing options available to them. Counseling enables the client to cope with the potential stress of going through the testing experience.
2. When the client voluntarily agrees to undergo testing to determine their condition, they are then taken to the laboratory for a blood sample to be drawn. The test checks for antibodies HIV-1 and HIV2 using either ELISA (Enzyme-linked immunosorbent assay test) or simple rapid test. A comparison check test is then done using either ELISA or Simple rapid test having used a different antigen preparation.

¹ Clients/Patients: Patients who visit ART centers or Hospitals providing ART. From here on clients will be used to mean these patients.

3. The client is then counseled again about the meaning of the result. If negative the client is encouraged to live positive and how to avoid infection.
4. If positive then the client is counseled about how to live a healthy life. A family based approach is adopted during HCT, whereby the client is advised to bring the immediate family i.e. spouse and kids for treatment. This improves adherence because there is disclosure and the client does not need to hide the ARVs being taken in.
5. The center follows strict MOH guidelines before introducing a client to ART, based on the results of the test administered in (2). These tests are explained below.

Some HIV Related Treatment Tests

There are broadly 3 categories of tests conducted for the HIV client:

The first test identified in (2) above checks for the HIV virus and confirms whether an individual is infected (+ve) or healthy (-ve). The second test determines what is known as the viral load in the body, this by Ugandan standards is an expensive test and not available in all the medical and ART centers country wide. The last type of test is a measure of the immune deficiency. This last test measures the CD4² (number of CD4 cells or T-helper cells in the blood) count on the client and is an important focus in this study. It is the test that determines at which point in the client's treatment life, since the first test when the client should be enrolled onto the ART.

When a client's test returns a CD4 count of less than 250 then they are placed on ART. The table below shows the considerations placed before deciding to place a client under ART, as adopted from the MOH guidelines.

Clients CD4 Cell Count	ART Options Available
>350	Client not yet eligible for ART treatment
CD4 between 350 and 250	Consider treatment for clients suffering from TB, showing symptoms, or pregnant women
CD4 less than 250	Begin ART

Table 1-II : CD4 level to commence treatment

Aside from the considerations of the CD4 count as an outcome on the client, WHO has also developed additional guidelines for beginning ART on clients based on clinical staging and immunological criteria. A summary of this is depicted in the table below.

Clinical Stage (WHO 2006 guidelines)	CD4 Cell Count	Action
I	CD4 Guided	Treat if ≤ 250
II	CD4 Guided	Treat if ≤ 250
III	Consider CD4	Treat if Pregnant or symptomatic & ≤ 350
IV	Treat	Treat

Table 1-III: WHO clinical staging and immunological criteria for initiating ART

If the client meets the conditions stated above for enrolment into the ART then another session of counseling is organized dealing with specific issues such as ensuring:

2 Cluster of Differentiation 4: This is a type glycoprotein found on the surface of T-helper cells.

1. The client understands that the treatment using the ARVs is not a cure for the condition and though viral loads may be low they still have the potential to infect others and responsible behavior is required.
2. That adherence is crucial to the success of the treatment and should be taken seriously.
3. The importance of proper feeding and responsible living for the client and knowledge lack of these will place the client or others at unnecessary risk.
4. That sometimes the ARVs drugs have side effects and people may react differently to the drugs depending on each individual clients experience and circumstance.

Additionally the client once again goes through a thorough medical and physical examination in addition to the laboratory investigations. This is important in order to determine a range of variables and indicators such as potential pregnancy, past illness, Tuberculosis, weight, height, CD4 tests, history of opportunistic infections amongst others as well as client vital statistics.

A lot of emphasis and care is required from the moment clients visit an ART center to the point when they are enrolled for ART and for the duration of the therapy. Furthermore the level of information captured, stored for each person is considerable over a period of treatment. Given the number of clients that the ART centers have to deal with followed by continuous analysis and follow up with adherence monitoring while factoring all the important indicators and treatment options; monitoring becomes a challenge. Similarities and commonalities in the therapy progression need to be tapped to improve the levels of care offered to clients during therapy. More complex analysis of these different segments of data available and collected continuously from the clients during therapy is required to provide better insight on the client condition.

1.1.6 Information Communication Technologies

Advances in the field of knowledge engineering offer opportunities for addressing healthcare deficiencies. Knowledge engineering offers the opportunity of using recent IT advances in integrating knowledge into computer systems to assist in solving complex problems. The GOU (2010) draft IT policy states its vision as *“a knowledge-based economy where national development and governance are effectively enhanced by harnessing and adoption of IT-led economic transformation”*. The GOU’s IT mission is to transform the economy of Uganda through the utilization of Information Technology (IT). This research is in line with this.

1.1.6.1 Policy

The IT policy specifies priority areas such as the setup of the legal framework, IT infrastructure, human resource development, research and development, promotion of IT and the development of national IT standards. Additional focus areas include E-Waste management, hardware and software industry, IT security and resource mobilization. The spin off from this policy has been development of electronic transactions and electronic signatures Act of parliament, the E-government framework, Computer misuse Act and the Business Process Outsourcing (BPO) Strategy. More recently the Ministry

of ICT (MoICT) has been working with technical groups in the Ministry of Health (MOH) to develop an e-health policy to manage Information systems in health care.

Uganda has a dynamic ICT industry with increasing investments in ICT and high growth rates as reported by the Uganda Investment Authority (UIA 2009). The growth in the sector has been remarkable especially with the mobile and telecommunications industry, with increased number of players and actors and stiff competition.

The competition has resulted in numerous innovative products in the industry including, mobile applications, mobile banking and mobile cash transfers. This growth in the mobile and cellular phone segment has not been replicated in the internet access and connectivity despite the fact that they were both introduced to the Ugandan market at around the same time.

1.1.6.2 Infrastructure

The government has been working on nationwide connectivity through the draft IT policy and the formation of the national information technology authority (NITA) is pushing for improvements in this sector. This is evident from government strategies such as the Uganda Broadband Infrastructure strategy and the implementation of projects such as the National Data Transmission Backbone infrastructure NBI/EG to improve communication speeds. A number of town and districts have been connected through fibre optic cables. The arrival of the submarine cables connecting the east African coast and inland has meant faster and cheaper access to the world. The map below (Figure 1-1) shows the plan for the National Backbone Infrastructure (NBI).

The NBI has improved and continues to improve the connectivity between the different local government headquarters in the country, indirectly improving the access to centralized information from central government. Local and regional health facilities in the connected locations can benefit from this infrastructure to access and share national information with and from central government ministries including MOH. This infrastructure has been used to support the district health information system (DHS2) (HISP 2012) project at MOH, providing central access to health information from the different areas in the country.

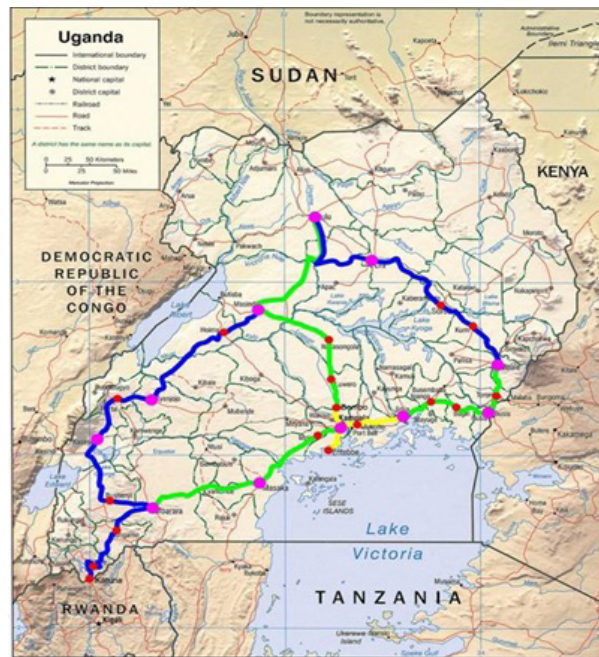


Figure 1-1: National Backbone Infrastructure (Adapted from MoICT)

1.1.6.3 Open Source Software Options

When establishing ICT systems, the costs due to ICT can include; infrastructure setup costs, hardware and software procurement, training, maintenance and support costs over the system lifetime. Open source license software can significantly reduce some of these costs such as the software and license costs. The Open Source Initiative (2009) (OSI) defines open source software as one that meets a set of criteria including; free distribution, source codes that can be accessed and does not discriminate against persons or groups. The license does not restrict on the use of other software.

The OSI is the organization that promotes the use of open source software and under its license review process has approved more than 50 licenses. These includes widely used categories like GNU General Public License (GPL), FreeBSD License, special purpose licenses like Educational Community License and adaptive public license making up the miscellaneous category. There are further OSI license categories like Academic Free License and Microsoft public license. Through these open source licenses, many innovative groups and communities have developed useful software products in different areas. Developers collaborate on software projects and incrementally produce software, updates and new versions. This collaboration and sharing is fostered by the open software licenses.

The free distribution criteria of open source 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, some of which would not otherwise be able to afford the prohibitively expensive license fees involved.

Taking an example of the GPL, when a software product is released under this license the recipients are granted the following rights:

- i. The right to use the software anywhere and in any situation
- ii. The right to redistribute the software to others as long as the source code is included and the distribution license remains GPL.
- iii. The right to create derivative work of the code and redistribute it as long as the resulting software code is made available at redistribution time and as long as the resulting code is licensed under the terms of the GPL.

By making use of the open source licenses available developing countries like Uganda can tap into some of the plethora of innovative software products available to address to some context problems with limited cost investments in expensive software. This is possible in the field of health care and most especially HIV/AIDS client monitoring and therapy.

1.1.6.4 Concluding Remarks

Beyond the infrastructure levels, ICTs can assist with the generation of new knowledge through analysis of data which is collected during day to day operations by health care facilities. Large amounts of data are collected on a daily basis by health care facilities (Tan et al. 2003), but the form of this information and the way it is organized does not facilitate the generation of new knowledge (Wasan et al. 2006). The use of ICTs in health care in the developing world has been minimal. This can be attributed to many reasons including the high costs of the ICTs. Software license options such as open source can help improve accessibility of some systems through the reduction of the costs involved.

In the developing country 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 health 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.

1.1.7 Data mining and Data Warehousing

The MIT technology review (MIT 2001) named data mining as one of the ten emerging technologies to change the world. Data mining is a technique of identify unique patterns in large datasets (Cabena et al. 1998; Hand et al. 2001; Larose 2005) by methodically sifting through this data. This can be done on data which is generated by health care providers during their day-to-day activities. Data mining is supported by categorization. Categorization involves the organization of the data into a format

suitable for data mining by formation of dimensions. Through aggregation and continuously storing data in a data warehouse or a central repository over time (Kimball and Ross 2002; Connolly 2002), we create a knowledge base where data mining can be carried out. The amount of data collected by different organizations has increased considerably over the recent years (Palpanas 2000). This can be explained by the increased volume of capture and storage of information and data during organizational processes. The advancement in the IT industry has facilitated more increased capacities of storage possible for organizations. This landscape has resulted in increased implementation of data warehouses systems. These offer data collected specifically in relation to a subject area e.g health, sales, treatment and HIV collected over a period of time. This data is constant and doesn't change but new data is added to it over time as more data is collected offering a good basis for data analysis.

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 clients that are facing chronic illness such as AIDS. The AIDS disease has adversely affected the nations of Africa. UNAIDS (2010) reports that sub-Saharan Africa has close to two-thirds (2/3) of all the people infected with HIV in the world. In the case of Uganda the country has had over a million deaths due to the virus with current estimates of close to 1 million infected and 250,000 undergoing ART. With this scenario in mind, sub-Africa has had the largest potential percentage of AIDS patient's information in the world and this culminates into crucial information that can be used for study. This data if carefully stored could be used to generate a variety of information including the disease progression and number of people infected as well as give 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 even with this potential for large patient data sets for data knowledge engineering in Africa and consequently Uganda, the framework for data mining in affordable 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. The format of data storage is not conducive for quick retrieval and large scale analysis. This is because of varying standards and differences in source systems and operating standards across different health facilities. Ranging from paper based systems to loose electronic medical records the architecture of these different systems does not facilitate analysis levels required for knowledge generation. Given the above one can argue that 'resource' in terms of data is not being made adequate use of. *This research therefore aims to study, propose and develop a way forward in integrating knowledge engineering in health care to assist in monitoring AIDS patient treatment.* An open source health care data warehouse was developed for data analysis using data mining which forms a standard framework on which to develop future open source knowledge management systems health care. Two data mining algorithms of regression and classification were used on the data warehouse to facilitate ART monitoring.

1.2 Statement of the problem

During ART, patient monitoring and evaluation is paramount in assessing intervention and to ensure safety of the patient (MOH 2009, pp 39) especially in light of related toxicities of ARV drugs. The importance of strict adherence to treatment cannot be overemphasized given the strict adherence requirements of ART (WHO 2006, pp 10). A patient's progress has to be consistently monitored from the moment of diagnosis to the end of treatment. This can be monitored using patient viral load values. 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 like HIV, 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 limited incorporation of the broad advances in information systems and analysis techniques.

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 HIV where medical cures are still being sought and all new knowledge and techniques about managing the disease during treatment is welcome.

1.3 Objectives

The main objective of the research was to develop a framework for adopting knowledge engineering in information systems for monitoring of HIV/AIDS patients.

1.3.1 Specific Objectives

The specific objectives of the research were identified as follows:

- i. To survey the status of information systems in health care in Uganda.
- ii. To identify the constraints to incorporating knowledge engineering in monitoring of HIV patients in Uganda.
- iii. To develop a dimensional model for HIV patient monitoring in Uganda
- iv. To explore the deployment of data warehouses for use in knowledge discovery in HIV.
- v. To develop, simulate and run an open source health care data warehouse targeting HIV patients in Uganda
- vi. To develop and propose a cost effective data mining model for HIV patient adherence monitoring in Uganda.

1.4 Research Questions

To achieve the objectives of this research, the answers to the following questions were sought.

- What is the status of information systems in health care in Uganda?
- What are the constraints in developing an HIV patient data warehouse using open source software?
- What are the tools for modeling HIV data in resource constrained settings?
- How can open source data warehouses be deployed for knowledge discovery in HIV?
- What dimensional model and data mining technique in an HIV/AIDS data warehouse can be used for in monitoring patients during ART?
- What is the importance of user engagement and participation in developing cost effective knowledge engineering systems for HIV/AIDS patient care?

1.5 Scope

The research involved the development of a data warehouse for HIV patient monitoring in Uganda. Simulations were then carried out to establish the most effective data mining algorithm to use for monitoring adherence to ART therapy. The data focused on analysis of HIV/AIDS patients (adult women and men) records from ART centers and hospitals in Uganda. To enhance co-evolution of the system cooperating partners were identified from different areas in the HIV/AIDS domain. This included researchers from Makerere University, selected health care professionals, patient organizations and NGOs dealing with HIV/AIDS and the Ministry of Health. These cooperating partners were involved in the development process enhancing cooperation, coevolution and ownership of the outcome functional system.

The development took into consideration the information system used in healthcare in Sweden specifically in Blekinge Hospital and researchers were also contacted at Karolinski Institute in Sweden. Data mining through linear regression and classification through logistic regression were tested and run on the data in the data warehouse, forming a basis for proposal on a suitable algorithm to adopt for monitoring adherence through projection of viral load and treatment failure during therapy.

1.6 Ethical Considerations

Given the sensitive nature of the research domain and the privacy concerns regarding patient information, steps were taken to ensure very high levels of confidentiality and privacy in this research context.

For security issues, specific patient data such as names were obfuscated using the Advanced Encryption Algorithm (AES³) standard using a 6 character symmetric encryp-

³ This algorithm has been classified to protect USA classified information

tion key. This data is therefore always gibberish when accessed and the patient is only referred to by a number and the link to the number requires a reverse decryption with the symmetric encryption key. This ensured that knowledge of patient confidential details like names became irrelevant and not required. When mapping the patient treatment progression and monitoring adherence, what was taken into consideration were the characteristics that the patient displayed and results from tests administered, as opposed to the specifics like name and private data.

The research was based on the MOH and WOH guidelines for counseling and testing, which requires patient's consents for them to access testing and counseling from health providers.

1.7 Justification of Study

The last decade has seen IT playing an increasingly important role in our daily lives including health care. Medical treatment faces challenges of learning and adapting quickly from the ever growing volume of data and sharing the generated knowledge amongst the people concerned. By making use of the daily information generated from patient care, the research develops a framework to provide insight during ART. This places care givers at informed points of view and helps in minimizing incidences of human error in treatment. The framework developed provides a platform for monitoring and evaluation of interventions in ART therapy and offers a basis for future planning.

An important function of knowledge discovery databases is their use in prediction and forecasts (Fayyad, Piatetsky-Shapiro, et al. 1996; Wasan et al. 2006). This research provides insight into ART patient monitoring and empowers treatment providers by offering them potential treatment options and directions. A key issue of ART is adherence and the problems associated with it which the system helps to address by getting the health care providers to "learn using past data". The ability to learn based on historical data is important in dealing with HIV and AIDS, where no known cures exist and more information about the disease is always welcome. The result enables different stakeholders to take advantage and learn from different commonalties in the disease amongst different patients thereby assisting in treatment monitoring.

Even as we approach the end of the millennium development goals (MDG) in 2015, the results of this research contributes towards the MDGs given the fact that Uganda and many other sub-Saharan African countries lag behind on some of these goals. This research is directly linked to MDG goal number 6 focusing on combating HIV/AIDS, malaria and other diseases (UN 2010). MDG goal 6 target A and B both deals with HIV/AIDS, with target 6A working towards halting and reversing the spread of HIV. Target 6B works with ART and seeks to provide universal access to the treatment for those infected. This research is directly in line with this by providing decision support for managing ART. In addition to assisting with monitoring therapy, this will also form a basis for planning and establishing exact numbers on the therapy which is a shortcoming of Uganda's country ART health care system as indicated by Bamuturaki (2008) and Windisch et al. (2011).

The resources lost due to human capital when an adult individual dies due to HIV and or treatment failure is on two fronts. The investment by the government in providing the treatment to the individual goes down the drain, when the individual succumbs to the disease either because of treatment failure, non-adherence or inconclusive or unavailable tests and testing facilities. As a consequent of death to HIV, affected families lose means of income especially if the death is of a bread winner in the family. By adopting a mechanism of addressing potential problems due to adherence, treatment failure and poisoning, we ensure investments in treatment are worthwhile.

1.8 Concluding Remarks

We are at the cusp of the intersection of the HIV epidemic in Uganda and the advent of high speed communication and related benefits including access and distribution of information. Coupling this with the penetration of the mobile phones into the country and this sets a stage for more information sharing and access across the country. After 3 decades of dealing with the epidemic there is a need to move the interventions in new and innovative dimensions while factoring in the new advancements in IT.

While still in need of a finalized ICT proposal, the government has moved forward in ensuring that some minimal legislation are in place to provide the framework for the use of IT in the country. The move to establish the e-health technical working group by MOH to guide and manage the development in the health information sector is important. The liberalization of the telecommunications sector in 1996 has also created the conditions for innovative products in the market with the intense competition, as many players join the industry.

As more sectors of the economy continue to appreciate the potential crucial role that IT can play in their different activities interdisciplinary studies that shed direction on applicable technologies for specific problems and knowledge engineering solutions and studies will hasten this process.

A review of the applications and software systems so far implemented in the country especially in relation to health systems shows them to be disparate and disaggregated with varying architectures and no single control framework. The main philosophy behind these health information systems especially the ones that focus on HIV is advocacy and sharing of information with the target stakeholders. The advocacy calls for changes in behavior or provision of information on where different services may be accessed. The level of analysis that goes into the data collected by the information systems for knowledge expert users to be able to make informed decisions depending on their roles is limited. However some of these implementations on can be scaled up and implemented across the country forming source systems for use as inputs into knowledge bases.

For the case of HIV and the adherence issues that affect the treatment of clients a knowledge base model that directly deals with this problem is still not present.

CHAPTER 2

THEORETICAL ANALYSIS AND STANDARDS

This section looks at the theoretical framework of the research. It begins by examining the IT and health care in the Ugandan context. Furthermore the concepts of data mining, data warehouses, knowledge discovery databases and selected data mining techniques are covered. The section ends by looking at some specific applications of data mining techniques to different challenges faced by health care providers.

2.2 Information Technology in the Ugandan Context and Health Care

Uganda has seen notable advancements at policy level, infrastructure access and access to IT. At the policy level several Acts of Parliament on different IT aspects have been adopted and 3 bodies have been established to deal with the regulation of the sector. This includes the Uganda communications commission (UCC), the broadcasting council and the National information technology authority (NITA). NITA is in charge of the information technology regulation. The development in infrastructure has led to greater access of information and an increase on the number of internet users in the country. UCC places this at almost 13%, at the end of June 2011.

Notwithstanding this the access of computer hardware equipment, software and other equipment in the country has been growing steadily since the government lifted the taxes on ICT goods in the country in the last decade. This has inevitably improved access to simple ICT equipment in the country like computers, laptops, and other specialized peripherals. This further resulted in products from brand name manufacturers as well as the so called “cloned” computers. The cloned computers are assembled from

components made by a variety of manufacturers and are normally lower in quality and cheaply affordable. There has also been importation of second hand computers which has resulted in the presence of very low quality and very old computers in the market. The public, legislature and other civil society organizations have decried this; in any case the government established through the MoICT a policy on e-Waste Management to curb this practice.

This increased level of access to basic IT equipment like computers coupled with the improvement in network access infrastructure has meant that more people have access to information on the web and from other sources. This on the flip side has meant that there is a substantial increase in piracy from illegally downloaded software and media. In the area of software the lack of sensitization on the different license options available in terms of open source still remains very limited. The result is continued piracy on software that majority of the ICT users in the country are accustomed to and piracy continues unabated due to ignorance of available alternatives. Improved internet access to information opens up the possibility of making use of the plethora of open source solutions. However their adoption still depends on increased sensitization and research in open source alternatives.

Concerning Internet access we have seen the development of linkages with software developers worldwide leading to some limited development of applications in the country. The improved internet access in the last 3 years has also played its part and resulted in a bust in the world wide web, with many web portals coming up and a number of successful web applications being developed in the country. Many of the organizations, individuals and teams highlighted previously have moved into the web sector. The government ministry have followed suit under the guidance of the ICT ministry and now UCC has been championing the development of web presences for 78 districts in Uganda.

A selection of companies in the country have web and short message system (sms) applications offering different functionality to the more than 16.5⁴ million telephone subscribers in Uganda. These sms applications have mainly been in the area of providing information such as forex, daily news, sports alerts and bulk sms transmission. Examples of earlier integration of sms applications include a system of using the sms applications to check for vehicle ticketing balance championed by trueafrican.com a local software company. Other examples include the attempt at linking sms to banks in the concept of mobile banking by trueafrican.com. However this attempt did not enjoy as much success as the mobile money concept introduced by the mobile telephone providers and championed, by the provider MTN. This was because the concept of mobile banking required the individual to have two things; the bank account and the telephone. Consequently with Uganda's banking network being accessible to only a small segment of the population this idea was never truly picked up. The mobile money concept however was dependent on only one key concept and that is the mo-

⁴ Taken from 2011 UCC report on the Performance of the telecommunications sector in Uganda.

mobile phone, or even better yet the subscriber's mobile number, essentially the mobile SIM card. Consequently given the 16.5 million subscribers in the country this went on and continues to be a very successful undertaking in the mobile and IT industry in Uganda and has changed the landscape of banking in Uganda.

In relation to health care and sms software development an example of a software implementation involves a system developed to check the credibility of health care providers (CapacityPlus 2012). Developed by a local NGO in collaboration with MOH, this system boasts of over 3000 records of health care practitioners. Using sms, mobile subscribers can query the system using a keyword of the medical provider to check if the user is registered in the system or not. A response sms system is then sent back to the mobile user indicating the medical practitioner or health provider's details. This system is worthwhile noting because of two key issues. The first is the fact that this is a good example of an open source implementation and secondly is that it provides a direct linkage to registered and authorized health centers. This system focuses on the human resource provided, ensuring that they are quality, have the possibility to be tracked and are authorized to provide care to clients.

The text to change project (TTC 2012) is another example of using open source and sms technology to provide affordable solution to advocacy in health care. The project has supported the implementation of a number of sms systems dealing with distribution of information to the population using sms on the subject of HIV, Malaria and Child health in different regions and areas of Uganda (TTC 2012). In the south west of the country in Mbarara a total of 15,000 mobile subscribers were targeted with simple sms broadcasting simple questions to test their knowledge on HIV and simple answers (Henriquez 2009). At the end participants were encouraged to seek voluntary testing and counseling at selected health care facilities. These are systems that focus on advocacy to the population, raising awareness on HIV, treatment options available and how to remain safe. In Kakira town near Jinja in Uganda sms applications have been used to disseminate information on maternal health, antenatal care and services available through the village health teams in the area (HealthChild 2012). These techniques focus more on the advocacy aspects and distribution of information to the users of the system as opposed to analysis of fact to facilitate decision making during client therapy. The use of mobile phone technology is innovative and worthwhile as it takes advantage of a communications medium that has quickly and heavily penetrated the Ugandan Market.

In western Uganda there has been an attempt at the implementation of electronic medical records system (Heatwole 2011; Otushabire 2011) under the millennium villages project in Isingiro district in Uganda. The project has facilitated quick retrieval of client information and avoiding the need for clients moving back and forth with their information.

The MOH DHS2 (HISP 2012) government project also provides an implementation of information systems in health. DHS2 stands for the District Health Information System, implemented managed by MOH, is open source software released under the

BSD open source license. It has been implemented in several African countries including Uganda. The current version runs a web based JAVA powered frontend with a PostgreSQL database backend. DHS2 enables remote locations in the country to log in through the web and carry out different functions including ordering for drugs and supplies. It is a transactional based system.

In the private sector a number of individuals have also developed different ranges of software for different functionality in the Ugandan market including specialized software for microfinance, finance management, loans, and management of schools and so on. Strong examples are information systems to manage schools, finance and micro-finance applications, loan applications and a few limited clinical management systems. These are mostly characterized by being standalone; non-integrated and customized for the specific organization they are developed for. These applications have been developed by small teams of 1 to 5 persons, this impacts on the product lifetime in the event there is only 1 developer. Majority of these software applications are marketed through word of mouth, with few registering massive sales and are under reported. However this limited success can also be blamed to the uptake of ICT in managing processes in organizations which is still in its infancy in the country.

Consequently accesses to these specialized softwares are limited and many of these variant systems exist in different departments, in governments, in companies and in selected health care facilities, especially the privately owned health care facilities. These softwares have varying underlying architecture and implementation and therefore miles apart and addressing different and sometimes similar end user needs and requests. Given that most of these source systems are closed, common development and growth is lost. Moreover these developments are based on single companies or individuals and only last as long as the company or individual is viable. As is common in the harsh industry many of these companies go “bust” with time and there ends the life of an otherwise good product.

Nonetheless some of these systems are still in operation and developers have continued to bring upgrades to these products over time. One such system is the clinic master software under close source. Clinic Master helps to manage clinical operations and financials at a health care facility. Clinic Master is built on windows .NET platform and Microsoft's SQL database platform. The system has a component that deals with distributing ARVs to clients. The gap in this system is that it stops short at monitoring and tracking the client, and instead focuses on the financial aspects of the health care center. The concept of deeper analysis and knowledge engineering on the client is also lost in this case. Furthermore the system is more of a transactional based system ensuring management of single client data, as opposed to an analysis based system carrying out large set retrievals and analysis. This and any such systems form a good source system for data into a data warehouse with an architecture that is more geared towards supporting analysis of stored information.

2.3 Data mining and Knowledge Discovery Databases

Data mining involves providing automatic pattern recognition and attempts to uncover patterns in data that are difficult to detect with traditional statistical methods (Fayyad, Piatetsky-Shapiro, et al. 1996; Fayyad, Piatetsky-shapiro, et al. 1996; Wasan et al. 2006). Wasan et al. (2006) argues that without the help of data mining the full potential of data that is collected by organizations would not be realized. They further state that for the case of HIV/AIDS data mining can be used to check the spread of the virus. Wasan et al. (2006) however does not elaborate clearly on how this spread of HIV can be addressed through data mining. The concern of various source systems and modes of storage of data by the different health care providers is noted. The full potential of the data collected by the organizations cannot be realized because different organizations maintain their massively collected data in a highly dimensional, distributed and uncertain way. Furthermore the analysis of the data would involve more than the available more familiar statistical available methods. To harness the full potential of data mining the data collected from business operations have to be reorganized in a way that supports analysis to establish these hidden patterns. This can be done through the creation of data warehouses that organize the data in a framework that highly supports drill down and analysis. Data mining is not a single technique but rather a group of heterogeneous tools and techniques used for different purposes. The techniques are based on statistical techniques, visualization and machine learning. Different techniques suite can suite different purposes with a need to analyze and compare algorithms to establish which provides the best answers depending on the problem.

Data mining is sometimes synonymously referred to as knowledge discovery. Fayyad et al. (1996) one of the first to coin the term defined knowledge discovery as “the non-trivial extraction of implicit, unknown and potentially useful information from data.” Automated analysis of large quantities of data in a database has possibility of discovering knowledge, but the distinction is that the discovered knowledge must be of interest; it must have potential value to the user (Wright 1998). Thus, the knowledge discovery process is often iterative as noted by Fayyad et al. (1996); with subsequent iterations refining on discovered knowledge. The knowledge discovery process goes through selection of target data, data cleaning, data transformation and finally data mining. Discovered patterns in data are then evaluated to enable the interpretation of this information to enable and inform actions of stakeholders of the research.

Since the data warehouse provides a good location for data mining 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 case of health care, the data can be enhanced by adding new attributes as well as by judicious aggregation of existing attributes, this has been shown by Rajagopalan and Isken (2001) to result in higher quality of discovery. 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. In this research context coming up with a methodology that ensures this ownership, incremental growth, and co-learning is important.

Data warehouse growth is an iterative gradual process involving sequential collection and aggregation of data from different relevant source systems (Fel! Hittar inte referensskälla.). 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.

Whereas data mining has been incorporated largely in different organizations in the developed world its use in Uganda, a developing nation is still limited. This can be attributed to a number of reasons; from lack of knowledge about the potential for generation of knowledge from the huge sources of data available to lack of expertise in developing data warehouses where data mining techniques can be applied. The lack thereof or limited use of electronic medical records in health care is another reason. The use of traditional techniques of documentation of patient care greatly hampers the effectiveness of analysis of patient information. Sensitization and education of the stakeholders in health care about the benefits of electronic means of documenting patient care may be required for the successful incorporation of data mining in health care organizations.

Cost in terms of specialized software and hardware for undertaking data mining project has also contributed to its adoption globally as well as in the country. Specialized off the shelf software for use in data warehousing projects can be quite daunting for resource constrained settings like the Uganda. In a data warehousing project implemented in health care, Ewen et al. (1999) estimated the cost of establishment at \$777,121 with the costs due to software and data modeling accounting for more than 80% of the project cost. In this particular project referenced by Ewen et al. (1999) the researchers recommend the purchase of commercial software on a server process based approach. The end users of the system would then access the system using a web –based interface reducing the costs due to multiple client licenses.

Development of a data mining culture in an organization or a sector requires long-term commitment. Developing successful data warehouses involves collecting data for long periods and the data warehouse gets more accurate with even larger amounts of data. This opens up a possibility to use data warehousing and data mining in areas like the health sector in Uganda, which has had a track record dealing with the epidemic with large numbers infected with the HIV virus; this is greatly compounded by the fact that even with treatment these numbers continue to increase.

2.4 Data Mining techniques and Applications in health care

Data mining facilitates the discovery of interesting facts from large data sets by making use of algorithms. Algorithms that enable learning are integral parts of knowledge discovery (Wright 1998). Machine learning is a branch of artificial intelligence that is concerned with developing algorithms that enable machines (computers) to 'learn'. Mitchell(1998) defined machine learning as follows: "A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E". Using a machine learning algorithm specialized tasks carried out by health care providers can be accomplished with higher performance and accuracy by developing knowledge engineering frameworks for a specific problem domain.

Learning can be either inductive or deductive, with inductive learning the computer learns by extracting rules and similar patterns from massive data. Deductive learning on the other hand bases learning on set norms, general principles or premises. Using these algorithms data mining is able to accomplish different tasks by trying to fit a model that is closest to the characteristic of the data being studied.

Models can be either descriptive or prescriptive. Prescriptive models are used to make predictions for instance making a diagnosis for a particular disease or to predict relapse in a particular patient's condition which patient is undergoing treatment monitoring. This means that patients can be subjected to a particular type of treatment basing not on their conditions but on conditions that were depicted by other patients but only discovered using data mining. Descriptive models on the other hand are models that are used to identify patterns in data using tasks such as clustering, association rules and visualization.

A number of machine learning techniques exist even then; there are continuous introduction and refined algorithms for the different methods. The two most common types of machine learning algorithms are supervised and unsupervised learning algorithms. In supervised learning the algorithm the data given to the learning algorithm is labeled or the correct output is given to the algorithm so it can differentiate between the 'right' and the 'wrong' answer. Once the algorithm has learnt from the data, then new unclassified data provided to it can be mapped to the correct answers based on the learning data. In unsupervised learning algorithms the data is not labeled and the algorithm has to independently discover the different labels. Each of these categories (supervised and unsupervised) of algorithms are applied differently depending on the context and the problem.

Some of the more common machine learning methods include but are not limited to the following:

2.4.1 Regression

One of the specific goals of data mining is prediction; the regression technique is paramount to achieving this goal of predicting attribute values for new objects(Cios

and Moore 2001). Using regression a data value can be mapped to a future prediction value (Fayyad, Piatetsky-Shapiro, et al. 1996; Wasan et al. 2006). This property becomes essential in the establishment of patterns between different variables under study. The relationship between two different symptoms can be established using regression. Patient test results can be used to predict other conditions.

In regression the data that we mine is considered a pair $P(X, y)$ where $X \in R^d$. The features of X indicate the answer y . In this case the algorithm is given these correct features and uses it to learn the characteristics of the data. The regression technique is adopted when we need to estimate the value of a datum using another variable or set of variables. However regression technique has the risk of generating wrong predictions if there were errors in the data used to train the model.

2.4.2 Classification and Clustering

Classification 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 health care by mapping the conditions being depicted by a patient to one of the known health care conditions. The algorithm uses the training data provided to group and define classes. In this case the prediction is where a training example belongs to a particular class or not. This is normally adopted when we have a known group of classes and we just need to fit the training examples to the correct classes. An example could be the classification of tumors into malignant or benign basing on a feature such as size of the tumor. That means we also have a set of training classes $\in (X, y)$ and the feature y is either true or false corresponding to the classes malignant or benign tumor.

Classification is similar to the clustering technique. The key aim when using the clustering technique is to find natural groupings (clusters) in large dimensional data. Clustering involves the identification of clusters for previously unclassified data basing on their set up of similar attributes (Wasan 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. There are many clustering algorithms and methods in use; some of these include hierarchical methods, partitioning methods, and model based clustering methods including decision trees and neural networks. The main concern here is how to incorporate the knowledge of the health care providers that is the domain knowledge into the mechanisms for clustering

2.4.3 Visualization

Using the technique of visualization data miner is able to gain insight using visual images (Maimon and Rokach 2005). This technique attempts to give the data some visual representation. The visual representations provide an opportunity for analysts to get insight on the data and make some preliminary hypothesis about the data relationships. The hypothesis can then be tested later with other techniques and algorithms. Visualization techniques are helpful when we deal with large datasets such as data warehouses.

Keim (2002) analyzed visualization by providing 3 categories based on the data to be visualized, the visualization technique applied and the level of interaction allowed from the end-users. Data that is multidimensional in nature present multivariate problems in machine learning. In a health care research such as this the data under analysis is multidimensional in nature calling for parallel coordinate visualization technique. (Keim 2002) defines parallel coordinate visualization technique as one that “displays each multidimensional data item as a polygonal line which intersects the horizontal dimension axes at the position corresponding to the data value of the corresponding dimension” (Figure 1-1).

When visualizing multivariate problems then a multidimensional system of parallel coordinates is used. This can be used to discover patterns in medical data sets using scatter diagrams and Cartesian planes different attributes can be compared and analyzed.

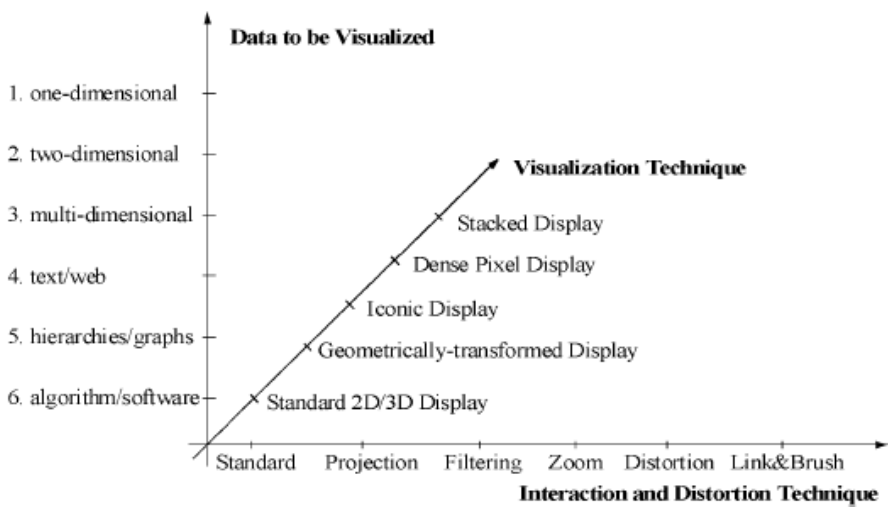


Figure 2-1: Categories of visualization techniques (Keim 2002)

2.4.4 Summarization

Data mining involves interaction with large data sets. Dealing with large data sets calls for summarization. The goal here is to characterize the data in terms of a small number of attributes, which have been aggregated (Cios and Moore 2001). 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.

Summarization can also be looked at as an attempt to generalize the data set through common features. For example in the clustering or classification techniques discussed previously we are trying to summarize the data sets into two or more classes based on the features in the data. In ART as we analyze the HIV data warehouse we attempt to

summarize the symptoms of the treatment to for example categorize the clients into groups corresponding to different treatment progression categories.

Summarization can be done on multidimensional data sets by providing summaries across the different hierarchically defined dimensions. Patterns can be depicted more clearly through summarization. For example in a healthcare system dealing with drug projection requirements for a district, average summaries of needs in a data warehouse over similar districts can be generated using summarization.

2.4.5 Artificial Neural Networks (ANN)

Artificial neural networks (ANN) have been applied to health care. ANNs are based on mathematical models that take an input and generate a corresponding output, they attempt to model cognitive system and neurological functions of the brain. ANN is based on layers with each layer taking inputs and producing outputs, ANN with many layers are sometimes referred to as multilayer neural networks. Each layer produces outputs that are then passed onto the next layer and becoming inputs into the new layer. Depending on the number of layers in the ANN, then different outputs are generated depending on what the previous layers are providing as output.

Er et al. (2008) reported that the use of multilayer neural networks with many hidden layers produced better results. ANN learns how to produce the correct output after being 'trained', using some learning data set. Once given new data sets the ANN mathematical algorithm can then produce the output based on the training data set it has used. The layers use a combination of backward and forward propagation techniques to determine outputs and inputs from corresponding layers of the neural network.

2.4.6 Health Care Cases involving Data Warehousing and Data Mining

In health care data mining has been used as a diagnostic tool(Wasan et al. 2006), in breast cancer(Jonsdottir et al. 2006) and in chronic illness prognosis(Huang et al. 2007).

Jonsdottir et al. (2006) reported on the development of a predictive model for breast cancer using machine learning to predict patient treatment up to 5 years after diagnosis. Patient information is analyzed in a database to determine which a class category for the patient. This enables the classification of the patient into the different survival groups that exist. Advanced predictions has been done in kidney dialysis (Kusiak et al. 2005). In this case, data mining is 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. The model proposed by Kusiak et al. (2005) had one main downside in that the data considered was rather small as noted by the researcher. The data did not consider demographics of the subjects under study. A similar issue was raised by Jonsdottir et al. (2006) regarding the breast cancer model, especially the lack of sufficient data from the 5 year period prior to the study. Maternal vaccination and preterm birth has

been studied using data mining as reported by Orozova-Bekkevold et al. (2007). The researchers studied the link between medicines used in pregnancy and preterm deliveries. This is an important area of research since preterm deliveries have a huge impact on infant mortality. However this study still highlights the need for further studies in data mining showing the linkage between vaccination and preterm deliveries. The study also requires refinements to enable it screen large databases such as data warehouses for relevant information.

ANNs have thus been used in health care to analyze patient blood and urine samples, to study diabetes and to detect conditions like tuberculosis (Lundin 1998;Er et al. 2008), in gastroesophageal reflux disease (Noya 2005), diagnosis of chronic hepatitis B conditions (Raoufy et al. 2011) and in the prediction of high risk preterm births (Catley et al. 2006). Lee and Park (2001) as quoted by (Vararuk et al. 2008) worked on ANN that predicted the symptomatic states of HIV/AIDS patients. A framework for the use of open source data warehouse platforms and data mining in monitoring ART for HIV clients still lacks.

2.5 Technology Considerations and Requirements

Database Management System

The resultant HIV data warehouse will ideally store a significant amount of data related to patient therapy. Performance will be affected by the size in addition to the volume of operations that will need to be conducted on the information retrieved. One technique of dealing with this is the implementation of column-oriented database management systems (DBMS). These store data as sections of columns as opposed to rows of data.

When dealing with data mart fact tables, the ability to select and retrieve a large number of rows over a small number of columns is crucial. There are also times during the loading of data that we require to supply similar data for large number of rows. Such operations are greatly facilitate by the architecture of the DBMS chosen be it column oriented or row based data access.

The research adopted an open source approach to ensure that costs were manageable in terms of software access. It was necessary to adopt a database management system that was open source and with ease of access, as well one that had the capabilities to support column oriented data management. MySQL was adopted to form the DBMS for the data warehouse system.

OLAP systems

These refer to online analytical processing systems and form the technological ideology of data warehouse architectures. OLAP tools allows for the interactive analysis of multidimensional data from multiple perspectives in an efficient and effective way. OLAP systems are geared towards analytical processing of data in a database as opposed to online transactional processing systems (OLTP) that focus on capturing transactions. OLAP attempts to deduce interesting information from sets of transactions stored in a

database. OLAP systems are therefore geared towards the retrieval and analysis of large sets of data from a data warehouse while, OLTP systems are geared towards the recording and capturing of large sets of transactions into the database.

OLAP systems are comprised of the data source (typically the data warehouse or large database), the OLAP server (where the analysis is done) and the user who requests for the analysis.

HTTPS: Secure HTTPS

The Secure Hyper Text Transfer Protocol is a widely used communication protocol that supports secure communication over a network. It is simply a result of layering between the HTTP protocol and the security capabilities of SSL and TLS (Secure Sockets Layer and Transport Layer Security respectively). HTTPS helps in ensuring secure authentication over a network and offers bidirectional encryption between the two communicating parties. This technology is important when dealing with security issues of managing data in a data warehouse that is accessed over a network or internet.

Furthermore given that HTTPS is based on native HTTP, encrypted access is possible with simple HTTP enabled devices. The entire HTTP protocol including resources being requested, query parameters and cookies can be encrypted and decrypted. This simplified access also ensures that training can be quickly and easily done with stakeholders.

2.6 Ethical Challenges in Health Care Data Mining

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.

The column orientation feature of some DBMS is a feature that can be used to facilitate the security. This is possible because the retrievals are quicker and for large number

of rows over a limited set of columns, this enables encryption and decryption to be carried out on the specified columns only because we are not selecting the entire column list for the dimension.

2.7 Status of Health Information Systems in Uganda

The paperless information systems industry in Uganda and indeed Africa is still in its infancy. With a large number of organizations, leave alone health care facilities still preferring to retain paper records storing patient information. This is a big hindrance to analysis, access and sharing of the information, not to mention the fact that this may result in inaccuracies in the data due to redundancy. This hampers quick retrieval, volume of retrieval and the complexity of analysis of data that can be undertaken during studies. Nevertheless even with the limited incorporation of ICTs in the health care sector, there are still some areas, which have adopted the use of information systems.

To mention a few, the infectious disease institute in Mulago Hospital, which deals with a range of diseases has a simple information system, which attempts to digitize the patient information after patient visits with physicians. The nurses in the institute first record the patient visit information in paper files and forms whereupon the information is then entered into computers later. This source system is more suited for transactional processing and not data mining operations. Using this they are therefore able to carry out statistical analysis on the data that they receive.

Some of the first efforts at computer based information systems to assist health care providers in Uganda were through the provision of information to health care providers in remote offices from different head offices. This was facilitated through the use of Personal Digital Assistant (PDAs) in Uganda. These were expensive, unsustainable for many reasons. First of all the PDAs the cost and only selected health care practitioners and projects could afford them. Even if cost was not the issue then the information was only accessible to the doctors and excluded the other stakeholders in health care.

The recent boost in the mobile telecommunications industry in Uganda has seen a number of mobile based applications springing up in the health sector (Henriquez 2009; CapacityPlus 2012; Heatwole 2011; HealthChild 2012) in the last 5 years alone. Just as the original attempts with PDAs these implementations have only marginally moved away from advocacy and information sharing there is still limited focus on analysis of data collected to enhance patient care through facilitating decision making and treatment options for care providers. The implementation of the DHS2 (HISP 2012) by MOH highlights the implementation of a centralized information system that collects and shares information from the different district health facilities. It is an OLTP based system that can be used as source for an OLAP based system with room for addition of a knowledge engineering component.

In summary the status of the information systems in Uganda's health care is as follows:

Strengths

1. There exists basic ICT infrastructure in many districts (78). This has been further improved by the fiber connectivity between certain districts in Uganda both by the government and other private communications companies. Refer to Fell Hittar into referenskölla. showing the National Backbone Infrastructure
2. 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. This has been done using the DHS2 tool.
3. Information from population surveys conducted by Uganda Bureau of Standards (UBOS) available. Including the production of demography health surveys (DHS), with the most recent being DHS 2011.
4. There is regular dissemination of information through meetings /workshops and other forums such as reports.
5. There is evidence and existence of basic functional HIS coordination mechanism present under the resource center for Ministry of Health (MOH) Uganda.
6. Appreciable demand for information from senior managers, policy makers, development partners for decision making.
7. 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.

Weakness

1. 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.
2. Lack of a training policy for health officers at all levels.
3. Lack of a comprehensive strategic plan.
4. No relational data warehouse for all HIS sources.
5. Most of the collected health information is paper based hindering complex analysis and requiring digitization before analysis.
6. 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.
7. Poor integration of regional and national health facilities.
8. Inadequate disaggregation of data by gender, socio economic status and geographical profiles (drill downs, dimensions data warehouse)

Opportunities

1. Improvements to ICT infrastructure country wide by the government.
2. Competitive and active mobile telecommunications sector.
3. Internet Fiber Backbone into the country, facilitating even greater access to open source software resources.

4. Government policy on non-taxation of selected ICT commodities to boost the sector.
5. Opportunities for research in HIV/AIDS given the progression of the disease, Uganda's experience.
6. MOH developing guidelines on e-Health to guide integration of information systems in health.
7. Capacity building for health information officers.
8. Development of national and sub-national web based health data warehouses and repositories.
9. 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.

2.8 Human Resource Requirements

One of the challenges faced by Uganda's industry is the challenge of the human resource. These range from problems in numbers of health care providers available [the doctor patient ratio in Uganda is 1:15,000], their distribution nationally, and the technical expertise of these health care providers. The government must be commended for ensuring that 72% of the population live within a 5km radius of a health care facility however, the next challenge is ensuring that the health care facilities have facilities and the human resource to provide services to the population.

The access to medical service providers can be blamed on many factors but primarily on the poor remuneration. This has resulted in many of the health care service providers leaving the country and relocating to the neighboring countries where the pay is better. However even those who choose to remain in the country have to double between working in the government health care facilities where they are employed and private health care facilities that pay slightly better. The private health care facilities are preferred because providers can levy a cost on treatment. Those who remain in the public health care facilities are demoralized, with few opportunities for training and therefore limited in upgrading their level of service.

There has therefore been a need for an effective method of managing the human resource available. This involving managing the allocation of health care professionals available and their training needs for better service delivery. The government has in 2012 with the support of the United States Agency for International Development (USAID) rolled out a health human resource management information system based on open source (CapacityPlus 2012). Using several partners the MOH has managed to have the system installed 15 hospitals and 68 district health care official offices. The system has also been installed in the MOH officers and select professional organizations. In addition to it being used to verify credentials of health care staff, it will also facilitate strategic planning for the health sector. This enables planning budgeting allocation, hiring of additional staff and analyzing training requirements. Specific objectives include analysis of optimal allocation of limited human resources to the different

health care centers in the country. The implementation of this Human resource system is based on open source, meaning it can be customized specifically to deal with the Ugandan context.

ART monitoring information systems require training on two fronts. The first is the training involving the administering and managing therapy to ART clients. The government has tried to address the first training need through ART guidelines and clinical procedures by the MOH and by WHO (MOH 2009). In a study done by Lutalo et al. (2009) in Uganda sampling ART providers; 49% had not received training on monitoring of patients during ART with a significant 35% not having had training on the initiation of ART. Through WHO guidelines and given the limited number of doctors a policy of task shifting has been adopted to increase access to ART. Lutalo et al. (2009) quotes WHO as having defined task shifting as the process whereby “Specific tasks are moved, where appropriate, from highly qualified health workers to health workers with shorter training and fewer qualifications in order to make more efficient use of available human resource for health. ” This has enabled the movement of ART initiation and monitoring from strictly the domain of the doctors to other health care providers. However the guidelines on how to initiate and monitor clients who are on the treatments needs to be frequently updated and rolled out to health care stakeholders to ensure everyone is working with the most updated approaches to managing the disease.

The second level training involves the use of IT information systems to enhance the level of their operations at the different levels. The health care stakeholders need to come up to speed quickly on how to use the system provided. There are two ways of addressing this, what can be termed the top down approach of having guidelines and ‘lectures’ or the cross cutting approach involving a synergy of self-help and co-learning between and amongst stakeholders. Adopting a top down approach of stakeholder training is recommended at the initialization of the stakeholder training. However for long term and to ensure learning and evolution, a more cross cutting approach is recommended. In the context of Uganda organizing the training of all the health care stakeholders offers serious logistical challenges; however when there is horizontal training and learning between the providers some of these logistical challenges are mitigated. Expertise is also developed through co-learning between stakeholders.

2.9 IT Training Needs

The specific training needs for stakeholders can be grouped under the following:

1. Training on understand the guidelines on initiating ART as specified by MOH(MOH 2009) and WHO(WHO 2006).
2. Training on monitoring clients on ART as specified by MOH and WHO guidelines. Understanding regimen of medications and different therapy combinations available.
3. Training in basic computer skills to manage, input of computer data input peripherals: mouse, keyboard and scanner if required.

4. Training in using web based applications, navigation using links and uniform resource locators.
5. Training on using different system functionality grouped by stakeholder.

Training needs highlighted above are all crucial to the uptake of the system for monitoring clients on ART and are interlinked. When care givers understand the procedures of initiating clients on therapy the clients can be initiated on therapy at the correct time. Once clients have been initiated on therapy, a common understanding of the procedures involved during therapy is crucial to ensuring that the same level and standards of therapy are availed to clients. Caregivers have to at the minimum be able to operate computer systems to input data and output information as these are now minimum standards of operating computer based information systems.

Comprehensive health care systems should be integrated and linked to different treatment centers. When dealing with data warehouse systems that depend on the comprehensiveness of inputs from different sources the ability of secure access from different locations is crucial. Basing on Uganda's context the usage of secure web-based information systems is necessary for health care providers.

2.10 Concluding Remarks

The liberalization of the telecommunications sector in Uganda in 1996 has led to some transformations in the ICT sector in the country. The next few years are going to be crucial as the country continues to improve its connectivity infrastructure as even more reliable connectivity comes to the region through fiber optics sea cables. Furthermore access to ART by those in need increase as government places more focus on this coupled with MOH adopting guidelines for e-health in the country sets the stage for developing frameworks for integrating knowledge engineering in ART patient therapy.

It is now not a matter of if but when we have e-government fully operational leading to full district headquarters and health care centers being linked to single systems. The development of the human resource management system and DHS2 systems already begins to pave the way for the inception of such systems. It is important to begin capturing the data available at the present time. The training needs of end users should be carefully taken into account to avoid ending up with systems that are white elephants and not utilized. Adopting a PAR methodology and ensuring user collaboration, communication and co-learning helps to ensure system ownership and learning between the stakeholders.

HIV surveillance is one of the key focus areas in the governments fight against HIV. A crucial part of HIV surveillance involves analysis of ART therapy. This comes from expert analysis of the different clients on ART in the different regions and areas in the country. This has to be improved hand in hand as access to therapy is also improved. Closely related is the issue of planning as (Bamuturaki 2008) noted that many facilities are still not able to indicate their needs from the center when requesting for ARV needs from the center. Thus needs to improve the supply chain of to the hierarchically organ-

ized health centers. Systems that provide expert answers such as knowledge driven data warehouses can address this shortcoming.

.

CHAPTER 3

METHODOLOGICAL CONSIDERATIONS

This research was approached from an interdisciplinary perspective, with stakeholders from varying sectors including academia, private health care providers, government, civil society organizations and information technology professionals. The problem of patient monitoring and adherence is one that is a major concern to many of the identified groups of stakeholders. Due to this reason a participatory based mode of approach was adopted involving the different interdisciplinary mix of stakeholders.

3.1 Action Research

Action research is one that is stakeholder focused and requires stakeholder involvement in the process to ensure acceptance and satisfactory system functionality. It is research that is conducted over various iterations involving reflective steps that involves a process of taking action and studying the action undertaken (O'Brien 1998; Riel 2010).

Rapoport in 1979 as cited by Avison et al. (2001) defined action research as that research that:

"...aims to contribute both to the practical concerns of the people in an immediate problematic situation and to the goals of social science by joint collaboration within a mutually acceptable ethical framework..."

This emphasized the idea of collaboration and the need to address problems that are affecting stakeholders in a particular area, in this case HIV/AIDS patient monitoring.

Avison et al (2001) proposes 3 areas to control action research, these control areas were used during the research. At the commencement of the research the problem to

be addressed was identified, secondly the warrant of control of the research was clearly defined and finally the degree of formalization of changes to the research. Problem identification can be researcher led though pinpointing an opportunity of leveraging an upcoming area of study to better enhance the degree of care, planning, and monitoring available in the domain of study. It could also be as a result of the stakeholders calling upon the assistance of researchers to help solve a particular problem faced. This research adopted a combination of researcher led and stakeholder assisted problem identification.

The research involved a continuous reflective process of taking study, tacking action and analyzing collected data and evidence.

The knowledge base of the subject understudy was continuously expanded in addition to solving the problem as a direct consequence of the research being done. This is best depicted by the iterative Figure below.

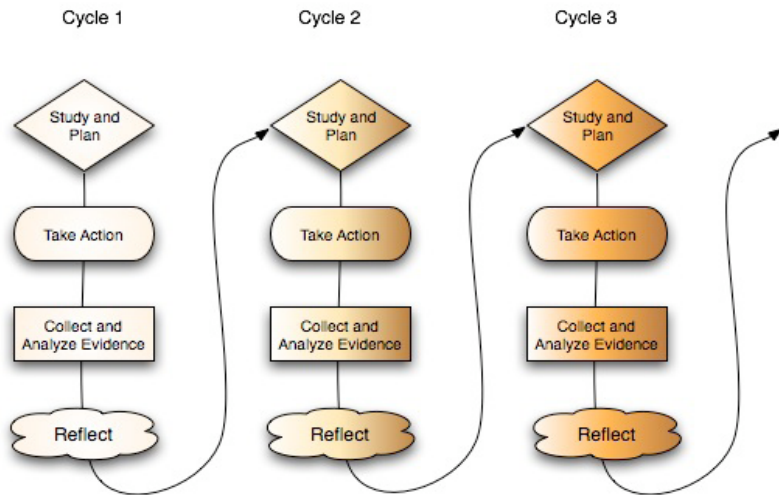


Figure 3-1: Progressive Problem solving with Action Research (Riel 2010)

As action research is focused around problems facing stakeholders in an area, there are many roads to the destination depending on the problem being research. It is therefore agreeable that there may not be a perfect silver bullet or manual of “howtos” when conducting action research. The aim therefore was to ensure that minimal good practices and guidance notes were observed during the research. The research adopted novel ways of approaching a solution based on the problem that was being addressed. In action research each study is novel in its own way with the innovative techniques applied to reach the desired destination.

As this research involved extensive system development, system development methodologies were adopted in the action research process. A more iterative development

model was adopted to complement the action research process. A component of the study indeed involved the development of a knowledge base and therefore this necessitated integrating an appropriate data warehousing methodology. The next sections present this hybrid methodology of study.

3.2 Participatory Action Research

Action research has evolved over the years into Participatory Action Research (PAR) 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 focuses the research methodology on involvement of the entire community. PAR was used to define the requirements of the different components of the system and to involve the users in defining the goal of the project and what the initial steps or direction of the research 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 (Mcniff 2002) with the collaborators/stakeholders being regarded more like co-researchers.

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.

Good principles of PAR were applied including offsetting biases, optimal ignorance, learning from and with people and applying optimal ignorance. When undertaking PAR stakeholders involved in the research have to progressively learn as they act on different stages of the project. Facts were cross-checked using triangulation involving cross checking the correctness with other sources either in person or through analysis of documentation.

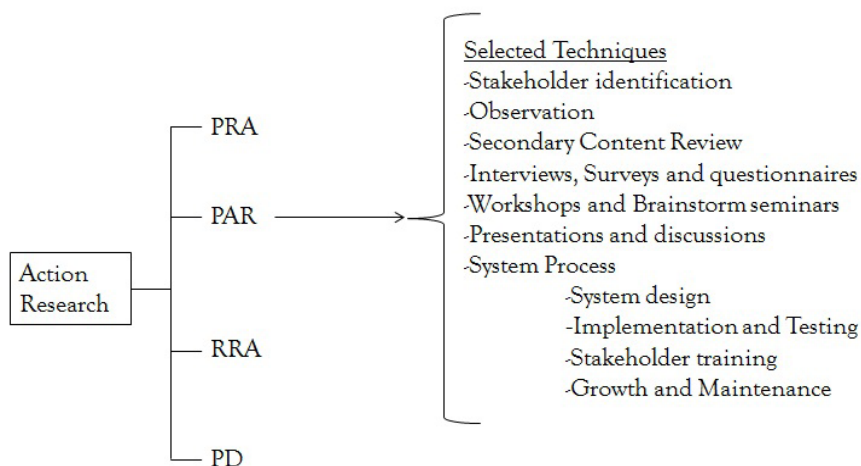


Figure 3-2: PAR as a component of Action Research

As alluded to earlier given the novel nature of AR projects, a hybrid group of techniques were adopted to deliver the different components of the research with special focus on the building of the system. The techniques adopted during the study are presented next in the sections below:

Identification of stakeholders

“By bringing diverse parties to the PR process early on, chances for a more comprehensive identification of relevant issues increase.” (Krishnaswamy pp3 2004). Potential stakeholders were identified through Uganda network of AIDS support organizations ([UNASO](#)) and MOH. [Snowball sampling](#) technique was adopted with referrals to enable identification of stakeholders. This involved the totality of ART provision centers, CSOs involved in HIV/AIDS, development partners, research organizations, and health care IT professionals.

Secondary Data Review

Content analysis was done on data gathered from selected relevant books, articles, journals, and reports. Further information was collected on systems done in the area of information technology, health care and HIV/AIDS therapy. This involved conceptual and theoretical literature review on, information systems, data warehousing, data mining and its impact on patient monitoring in Uganda. 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. Standard paper based information systems on ART were analyzed from MOH and reach out Mbuya a selected ART center. Developers of SMS mobile applications and other software developers were selected and contacted, and information on their system architectures studied.

Observation

Direct observation was adopted in certain situations to view the operations of health facilities providing health care services to HIV patients; this helped to gain a clear understanding of their operation. 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 Uganda as well as studying the conditions for integrating knowledge engineering to HIV patient monitoring. Direct observation and study of modeling systems was carried out at two hospitals in Sweden, Karolinska Institute and Blekinge county hospital.

Semi-structured interviews, Surveys and Questionnaires

Informal sessions with stakeholders were conducted 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 were used where the source of the information was extremely large and structured in form. These were effective in collecting qualitative information from the stakeholders involved, such as important dimensions of observation for HIV clients, therapy procedure and reporting mechanisms. The stakeholders in this case include selected HIV medical researchers, health care professionals, NGOs dealing with HIV patients, college of engineering, Design and Art Makerere and selected HIV patients. Government institutions consulted included MOH, MoICT in addition to regulatory bodies such as UCC, NITA and National Drug Authority.

This technique assisted with the generation of reports on the status of information systems, information technology and studies in health care and HIV. Constraints to adopting knowledge engineering in patient monitoring and the training needs of health care personnel were identified. This technique was important in establishing the process of ART provision to clients in different centers, the planning needs of district/ regional health teams. It also helped to identify dimensions to monitor during therapy, thus informing the dimensional modeling phase.

Workshops and Brain storm seminars

A workshop involving 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 assists 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 for knowledge engineering. Encouraging participation of all the people involved helped in avoiding situation where the researcher could 'impose' 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. An e- approach to encouraging participation was tested through the use of wikis. This enhances training stakeholders and the ability of users to learn from one another.

Presentation

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.

3.3 System Process

3.3.1 Phase I: System Design

The previous techniques in addition to meeting various research objectives helped to define the project scope and to perform analysis. This made clear the requirements, priorities, project plans and resource plan. This informed the system design phase and was used 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 was done in iteration with the involvement of the stakeholders using techniques such as brainstorm sessions, one on one interview, presentations, and observation. This phase assists in developing the system dimensional model that forms the initial foundation for the data warehouse.

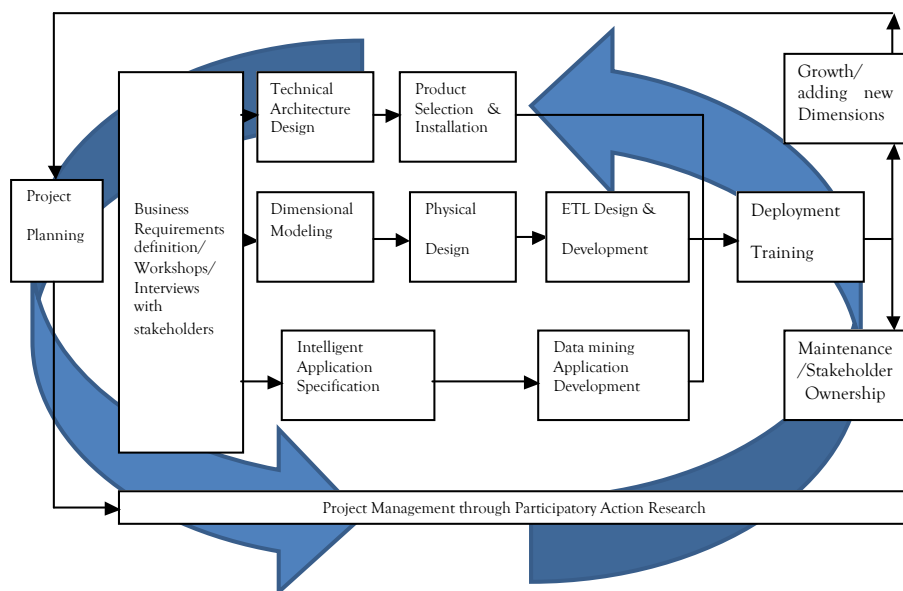


Figure 3-3: System Lifecycle process adopted from (Kimball & Ross 2002)

3.3.2 Phase II: Implementation and testing

The infrastructure setup was done in this phase. This involved analysis of different database management tools that could most efficiently manage the HIV data for the data warehouse. The implementation phase also involves analysis of different sources of data into the data warehouse and development of extraction techniques from these source systems into the physical design of the data warehouse. Given the varied source systems the extensive cleaning was done to gain enable a common standard of data for loading into the data warehouse.

Quantitative adjustments is done to ensure having 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. Appropriate technology to use to develop the user front end and different data mining algorithms to monitor patient adherence to therapy and patient viral load were analyzed. Testing of selected components and system wide testing was also done during this phase.

3.3.3 Phase III: Stakeholder Training, Maintenance and Growth

In PAR the ownership of the solutions to the problem being faced should be accepted by the end users. Solutions to ensure growth of the system and capability of the stakeholders to make changes to the underlying model of the data warehouse and data mining system were identified and tested. Techniques of using collaboration on online systems were examined and tests to enable the stakeholders learn and teach each other as they use the system. Dick (2002) likens PAR research to learning by doing and an

approach to doing this in the context of Uganda. This phase involved having selected stakeholder champions learn to use the system by doing.

The growth of the system meant examining and testing techniques of automatically adding new dimensions to the initial data warehouse dimensional model proposed in the initial phase of the research.

3.4 Concluding Remarks

The research adopted a hybrid of techniques and approaches that suited the problem being faced in the context of the research. The conceptualization and identification of the problem domain was driven by a mixture of researcher and the mix of interdisciplinary stakeholders involved. The research therefore has some arms in the Mode-2 (Gibbons et al. 1994; Nowotny et al. 2003) approach of research. This is research that is problem oriented; context centered and aims to contribute to the body of knowledge. The research is a hybrid of Mode-2 in that the problem identification was still driven by the researcher with the strong university linkage which is one of the hallmarks of Mode-1 (Aken 2001). However the linkage to solving the problem of patient monitoring in the context of Uganda pushes it into the domain of Mode-2.

Part II

CHAPTER 4

PRESENTATION OF PAPERS

4.2 Introduction

This doctoral thesis, focused on the use of knowledge engineering techniques such as data warehousing and data mining techniques of regression and classification to improve patient monitoring during ART. Some key outputs of the research are now presented as papers to journals and conferences. The papers are presented in chronological dates in order of writing. The papers have been revised to conform to the format of the thesis with the content unchanged.

- I.** Otine, C.D, Kucel, S.B & Trojer, L. (2007). Knowledge Discovery in Health Care Using Data Mining. Published in the Proceedings of the 1st International Conference on Collaborative Research for Technological Development
- II.** Otine, C.D, Kucel, S.B & Trojer, L. (2010). Dimensional Modeling of HIV Data Using Open Source. Published in the World Academy of Science and Technology, Issue 63. 2010
- III.** Otine, C.D, Kucel, S.B & 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
- IV.** Otine, C.D, Kucel, S.B & Trojer, L. (2011). Stakeholder Engagement and Participation in Determining Requirements for a Data Warehouse System for HIV/AIDS: Uganda's Experience. In Licentiate Manuscript.
- V.** Otine, C.D, Kucel, S.B, Giger, P., Rydhagen, B. & Trojer, L. (2011). WIKIs in Support of Intelligent Product Manuals for HIV/AIDS Data Warehouse and Data Mining Systems. International Journal of Modern Engineering Research (IJMER).
- VI.** Otine, C.D, Kucel, S.B & Trojer, L. (2012). Enhancing Data Warehouse Capabilities by Automatic Addition of Dimensions to Established Models. International Conference on Data Warehousing and Knowledge Discovery (2012).

The following additional Papers are also related to the research; these are in Manuscript and have not been submitted for publications.

VII. Otine, C. D (2012). Optimized linear regression algorithm to determine HIV patient CD4 count. In Manuscript.

VIII. Otine, C.D (2012). Unsupervised learning algorithm for monitoring HIV treatment failure. In Manuscript.

Paper I

KNOWLEDGE DISCOVERY IN HEALTH CARE USING DATA MINING

C.D Otine¹, S.B. Kucel² and L. Trojer³,

ABSTRACT

The exponential growth of data banks in health care creates opportunities for knowledge generation using data mining. Advancements in information and communication technologies (ICTs) 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 possibility of provision of extra functionality such as backup, encryption, security and complex quick analysis.

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: lana.trojer@bth.se

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 acquisition of considerable volumes of data from patients both past and present information, including patient bio information, treatment 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 possibility 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 (Kriegel et 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 4-1, adapted from the same source. In this adaptation the transformed data is stored in a data warehouse.

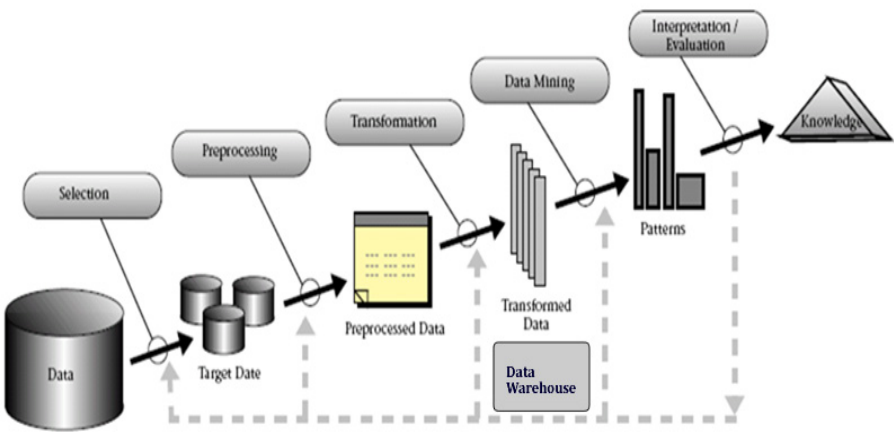


Figure 4-1: An adjustment of the Knowledge Discovery process (Fayyad et al, 1996)

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 Figure 4-1). 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 (Fayyed et 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 gastroepha-geal 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 prediction 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.

Acknowledgements

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.

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

REFERENCES

- Balajo, R & Mark, W.I (2001).Exploiting *Data preparation to enhance Mining and Knowledge Discovery*. IEEE Transactions on Systems, MAN and Cybernetics.
- Brodley, C.E., Lane, T., &Stough, T.M. (1999).*Knowledge discovery and data mining*.American Scientist 87(1):54-61
- Fayyad, U., Gregory, P.S.&Padhraic Smith (1996). *From Data mining to Knowledge Discovery in Databases*.American Association for Artificial Intelligence.Pg 37-53.
- Fayyad, U., Gregory, P.S. &Padhraic Smith (1997). *Knowledge Discovery and Data Mining: Towards a Unifying Framework*.American Association for Artificial Intelligence.
- Ivanka, O-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
- 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
- Keld, B. Finn, K &Jesper, S. (2004). *Participatory IT Design: Designing for Business and Workplace Realities*. MIT Press
- 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
- 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.
- Lundin, J. (1998). *Artificial Neural Networks in outcome prediction*.AnnsChirGynaecol 87: 128-130
- 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
- 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
- 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
- Siri, K.W., Vasultha, B. &Harleen, K. (2006).*The impact of data mining techniques on Medical Diagnostics*.Data Science Journal 5:190-126.
- Wright, P. (2007). *Knowledge discovery in Databases: Tools and Techniques*. Retrieved from<http://www.acm.org/crossroads/xrds5-2/kdd.html> on 3/11/2007

4.4 Paper II

DIMENSIONAL MODELING OF HIV DATA IN 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 potential 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 - Database, Data Mining, Data warehouse, Dimensional Modeling, Open Source.

I. NTRODUCTION

Data models form the foundation of data warehousing and data mining systems since they help to describe how data is to be represented and accessed. 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. Multi-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 updates 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 establishments who may have abundant and continuously growing data from taking advantage of data warehousing due to the high costs involved, 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 available 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.

II. DIMENSIONAL MODELING

Dimensional modeling is used to conceptualize data warehouses which are then implemented using star schemas or snow-flake 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

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

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 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 4-2). 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.

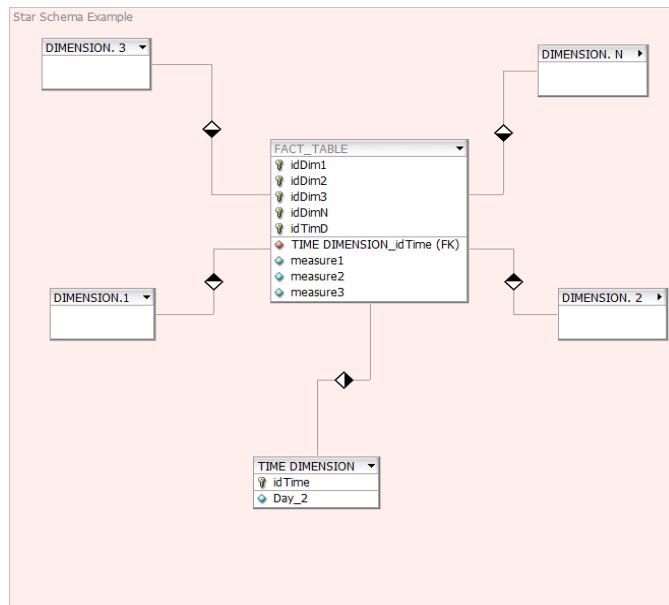


Figure 4-2: Star Schema

III. 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.

IV. 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.

V. MODELING PROCESSS

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 4-3 and Figure 4-4 indicate the screens for the Power Architect and DB designer tools respectively with sample models.

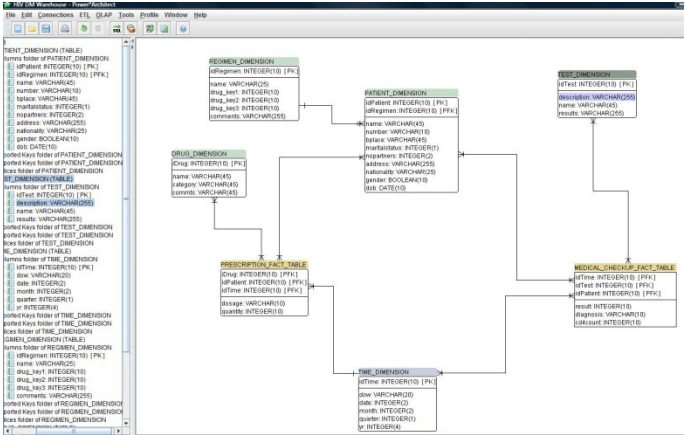


Figure 4-3: Power Architect Modeling Screen

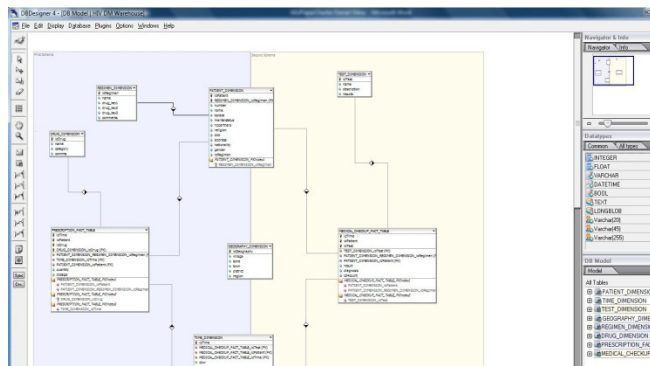


Figure 4-4: DB Designer Modeling Screen

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 system⁶ 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 infectThe 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

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.

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 CD4⁷ 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.

Figure 4-5 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.

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.

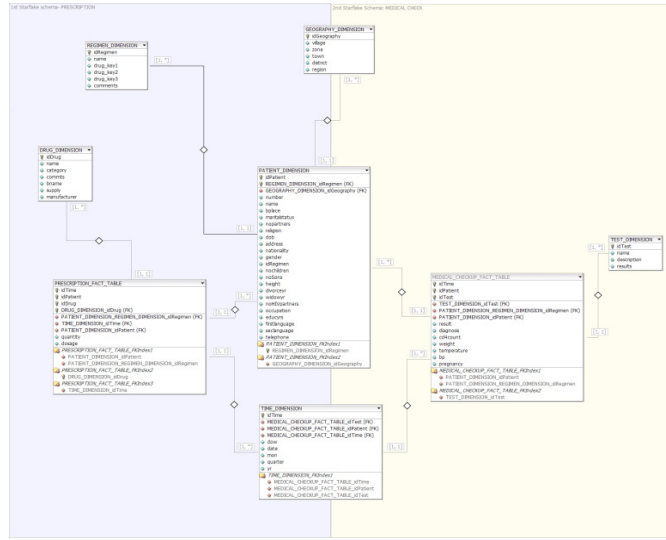


Figure 4-5: Data Warehouse Dimensional Model

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.

V. 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 Education, 17,1, 17-20.
3. Coar, K. (2006). The Open source Definition , Retrieved on 18th Nov 2008 from open-source.org: <http://www.opensource.org/docs/osd>
4. Dash, A.K and Agarwal, R. (2001). Dimensional modeling for Data warehouse, ACM SIG-SOFT 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. Kleijnen, 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.

4.5 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' treatment. 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, Data Marts

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 epidemic, 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)]

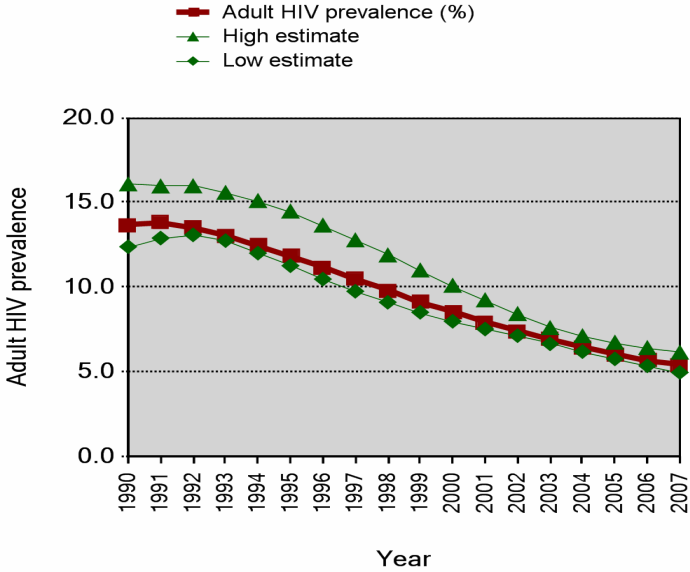


Figure 4-6: Estimated adult HIV (15-49) Prevalence % 1990-2009 (UNAIDS)

The graph above (Figure 4-6) 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 the success against the epidemic fight (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 where 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 al. 2003).

Source systems⁸ of selected health care Providers⁹ 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 (Otin 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.

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 ART to people infected with the HIV virus

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 and investigating the key issue that affects them (Wadsworth, 1998;McNiff, 2002;Cronholm and Goldkuhl, 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 different tables representing the dimensions in the domain of study 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 (Figure 4-7), followed by the list of scripts used to create the different dimensions in the database server.

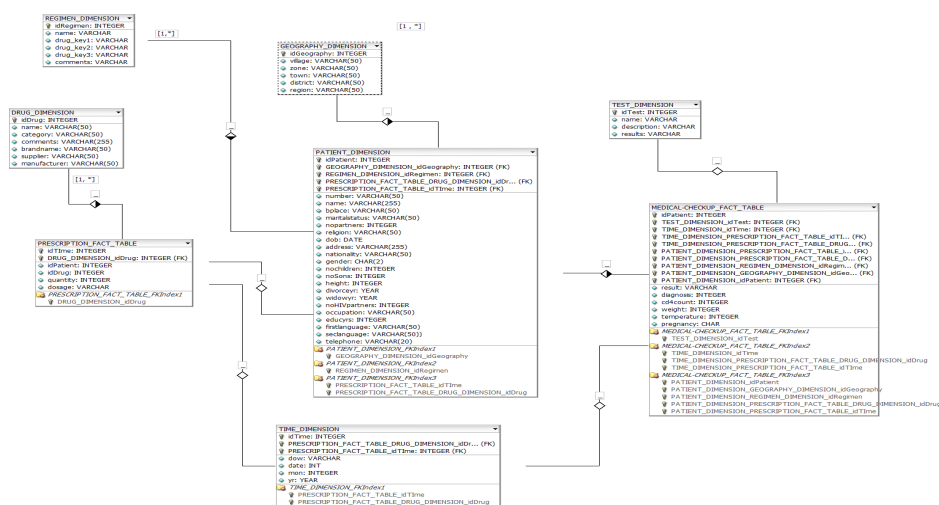


Figure 4-7: Dimensional Model used to create dimensions in the data warehouse

/*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 UN-  
SIGNED NOT NULL,  
TIME_DIMENSION_PRESCRIPTION_FACT_TABLE_DRUG_DIMEN-  
SION_idDrug INTEGER UNSIGNED NOT NULL,
```


PATIENT_DIMENSION_PRESCRIPTION_FACT_TABLE_idTime INTEGER UNSIGNED NOT NULL,

PATIENT_DIMENSION_PRESCRIPTION_FACT_TABLE_DRUG_DIMENSION_idDrug INTEGER UNSIGNED NOT NULL,

PATIENT_DIMENSION_REGIMEN_DIMENSION_idRegimen INTEGER UNSIGNED NOT NULL,

PATIENT_DIMENSION_GEOGRAPHY_DIMENSION_idGeography INTEGER 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_PRESCRIPTION_FACT_TABLE_idTime, TIME_DIMENSION_PRESCRIPTION_FACT_TABLE_DRUG_DIMENSION_idDrug, PATIENT_DIMENSION_PRESCRIPTION_FACT_TABLE_idTime, PATIENT_DIMENSION_PRESCRIPTION_FACT_TABLE_DRUG_DIMENSION_idDrug, PATIENT_DIMENSION_REGIMEN_DIMENSION_idRegimen, PATIENT_DIMENSION_GEOGRAPHY_DIMENSION_idGeography, PATIENT_DIMENSION_idPatient),

INDEX MEDICAL-CHECKUP_FACT_TABLE_FKIndex1(TEST_DIMENSION_idTest),

INDEX MEDICAL-CHECKUP_FACT_TABLE_FKIndex2(TIME_DIMENSION_idTime, TIME_DIMENSION_PRESCRIPTION_FACT_TABLE_DRUG_DIMENSION_idDrug, TIME_DIMENSION_PRESCRIPTION_FACT_TABLE_idTime),

INDEX MEDICAL-CHECKUP_FACT_TABLE_FKIndex3(PATIENT_DIMENSION_idPatient, PATIENT_DIMENSION_GEOGRAPHY_DIMENSION_idGeography, PATIENT_DIMENSION_REGIMEN_DIMENSION_idRegimen, PATIENT_DIMENSION_PRESCRIPTION_FACT_TABLE_DRUG_DIMENSION_idDrug, PATIENT_DIMENSION_PRESCRIPTION_FACT_TABLE_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_idGeography, REGI-
MEN_DIMENSION_idRegimen, PRESCRIPTION_FACT_TABLE_DRUG_DI-

```



```

MENSION_idDrug, PRESCRIPTION_FACT_TABLE_idTIme),
    INDEX PATIENT_DIMENSION_FKIndex1(GEOGRAPHY_DIMENSION_
idGeography),
    INDEX PATIENT_DIMENSION_FKIndex2(REGIMEN_DIMENSION_idReg-
imen),
    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_DIMENSION_id-
Drug)
)

```

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_DIMEN-
SION_idDrug, PRESCRIPTION_FACT_TABLE_idTIme),
    INDEX TIME_DIMENSION_FKIndex1(PRESCRIPTION_FACT_TABLE_id-
TIme, PRESCRIPTION_FACT_TABLE_DRUG_DIMENSION_idDrug)
)
TYPE=InnoDB;

```

Source 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-every day 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 (note that this has been truncated to only indicate a few fields of the dimension).


```

/*****Initial 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,noHIVpartners,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,noHIVpartners,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 an approach of cleaning the data that depends 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 Chart:

When a patient is being monitored for adherence their information is queried against the data warehouse in the following steps depicted in Figure 4-8.

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.

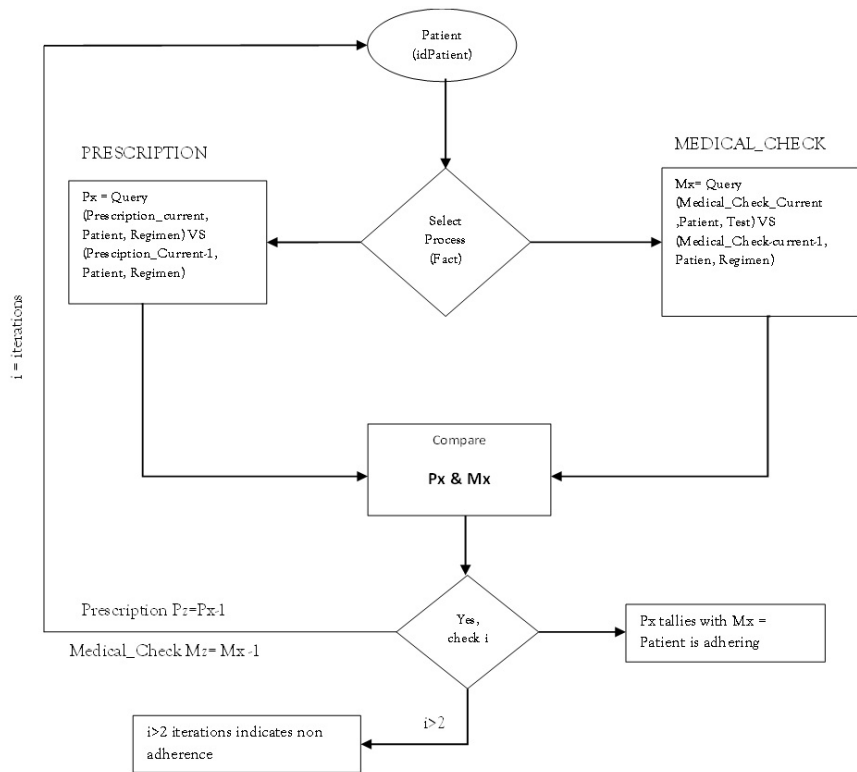


Figure 4-8: Prescription and Medical Checkup Flow Chat

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 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. Additionally 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 drugs 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-I02187
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 ILATA 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.

4.6 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 and 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.

Keywords: 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; this is 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 effective in results of analysis with increasing information. This in essence hinges the success of the data warehouse system on stakeholder acceptance of the system otherwise resulting in 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 part-

nership 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. Successfully implemented and accepted data warehouses would improve the effectiveness of ART programs in the country, providing patient monitoring and planning for ARV resources

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. The goal was to create a framework for dialogue to encourage the participant to engage and learn from each other.

Participatory Action Research (PAR) methodology was adopted because it ensures the participation of all the users [7] and puts into consideration different stakeholders viewpoints. 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. This is best achieved in group settings through discussions, for stakeholders to give their different viewpoints and learn from each other.

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 (Table 4-I).

Table 4-I: Stakeholder Expectation

Stockholder	Expectations of the system
Government/ Development partners (Policy level):	<ul style="list-style-type: none"> – 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	<ul style="list-style-type: none"> – 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:	<ul style="list-style-type: none"> – Receipt of better care – Receive standardized and specialized treatment. – Improved diagnosis. – Improved ART Management and service delivery – Improved patient follow up on adherence

Stockholder	Expectations of the system
NGOs and Donors	<ul style="list-style-type: none"> – 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 expectations were linked, for instance the need by governments to plan procurement nationally being dependent on records showing prescriptions of ARTs across the ART centers 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 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 common 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, but rather a tool or mechanism to enhance their role(s). 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. Furthermore their

Review and Realignment

Determining the initial vision of a data warehouse project is only the first step towards implementing the project. A critical 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 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.

This process is also crucial in coming to consensus about the implementation phases. Due to the interdisciplinary nature and different stakeholders with competing needs from the system. A phased approach of implementation and roll out of different system functionalities has to be embraced. Some high level functionality of expert decision support and forecasting may not be available in initial versions of the system. This would become available in future versions of the product however this process needs to be managed to ensure that some users do not become alienated, as they feel their requirements were not met.

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. Once users buy in to the idea of a phased approach with clear evidence why the phased approach is necessary then 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.

Techniques of militating against correspondence and expectation failures of the system are discussed. This ensures that the final system is used widely and above all meets the stakeholder expectations.

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 Hirschheim, 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*

4.7 Paper V

WIKIS IN SUPPORT OF INTELLIGENT PRODUCT MANUALS FOR HIV/AIDS DATA WAREHOUSE AND DATA MINING SYSTEMS

Charles Daniel Otine*, Samuel Baker Kucel*, Peter Giger**, Birgitta Rydhagen**, Lena Trojer**

*(College of Engineering Design and Art, Makerere University, Uganda)

Email: {hautine, sbkucel}@tech.mak.ac.ug)

** (Research Division of Technoscience Studies, BTH, Sweden)

Email: {lena.trojer, peter.giger}@bth.se)

ABSTRACT

The paper reports on the development of interactive product manuals for an HIV/AIDS data warehousing and data mining system developed in Uganda. It uses the online tool wiki to enhance online participation and collaboration between the different stakeholders. A modus of moving participatory action research to dialogue online by leveraging Web 2.0 is provided. The paper proposes ways of enhancing system usability. Techniques of enhancing learning, idea generation and user satisfaction are examined to improve a data warehouse and data mining system usability.

Keywords: Data Warehouse, Data Mining, Product Manuals, Collaboration, Usability, Stakeholder participation, PAR

INTRODUCTION

A data warehouse and data mining system were developed to help in addressing the problem of drug adherence amongst HIV/AIDS patients [3]. Due to the varied types of stakeholders involved, the development of a comprehensive product manual is vital to the systems success. The need to have a progressively growing system is crucial to a health care data mining system. The collaboration of stakeholders in the development of product manuals would enhance easier additions to the product manual of new functionality and discussions on changes required to the system.

The data warehouse system was built on open source technology to ensure that the costs to the user are minimized as much as possible. An intelligent product manual¹⁰ (IPM) for the system ensures that stakeholders utilize the system easily and effectively with as little costs as possible while minimizing the support and training costs. Providing a platform where different stakeholders can collaborate on a tool to enhance the usage of the system would create an environment to facilitate innovation through new ideas.

To develop the IPM, the research requires a collaborative tool that met the following requirements:

¹⁰ Only the product manual was used on the wiki and not the entire HIV/AIDS data warehouse and data mining system

1. Easy to learn, use and light weight (easy to install).
2. Openness and ability to facilitate brainstorming.
3. Facilitating collective effort by getting ideas and feedback quickly and reducing need for meetings.
4. Ability to enhance the process of building a product together.
5. Freely easily accessible by all stakeholders.

An online wiki was identified as a possible effective tool to provide such a platform that would emphasize user requirements.

In a previous study [1], [2], and [3] to address the challenges of adherence in antiretroviral therapy (ART), an HIV/AIDS data warehousing was proposed and implemented. This was based on open source software to address the issues of implementation in a resource constrained setting. [2] Developed a dimensional model in open source for the HIV/AIDS data warehouse. Different stakeholders were used to define requirements using participatory methods. This paper addresses issue of stakeholder involvement in training and support through the provision of an intelligent product manual.

The paper provides an overview on the state of the art in intelligent product manuals and collaborative systems for the development of user manuals. A section follows on the methodology for the development of data warehouse and data mining manuals and the challenges involved. Recommendations and different techniques are provided on how to enhance usability of data warehousing and data mining systems. A conclusion is then provided on areas for future research.

INTELLIGENT PRODUCT MANUALS

For the case of the HIV/AIDS data warehouse and data mining system the operation and maintenance of the system are critical. The operation and maintenance of the system depends on the training and support available to the stakeholders of the system. Comprehensive training and user support greatly reduces the time taken to bring a system to full operation by its stakeholders. To accomplish this there is a strong need for comprehensive stakeholder tutorials or manuals to cover a range of different system tasks that stakeholders are faced with every day.

User manuals can either be paper based or electronic. The paper based manuals are the conventional manuals that are passive with very static presentation. They are the basic paper based documents with contents, indexes and cross references. Some paper based manuals may also include graphs and texts to enhance communication to the users. The advent of Web 2.0¹¹ and more recently HTML5.0 has seen a move towards the electronic manuals which are more active and responsive based on user requests. The presentations are more dynamic, with easier navigation and integration of multimedia.

An intelligent product manual is an electronic, multimedia, knowledge based system that provides active assistance to the user of a product during any product related ac-

¹¹ Often called social media

tivity [4]. The related activities include installation, operation and maintenance. These are vital towards the success of any system.

There have been attempts at intelligent product manuals such as [5], [6], [7] and [8]. [5] Proposes a web-based approach that uses AJAX technology with event handling methods to capture real time user interaction with the system. This information is then used to construct meaningful sentences that are then put into a database and used to generate user manuals. The approach is limited in that though it enables real-time generation of a user manual, the participation of the user is somewhat limited to their usage of the system which is then captured. The ability of the end stakeholders to edit the manual is also limited and this might affect some innovative idea generation.

[6] Presents a methodology of developing intelligent engineering product manuals with a strong emphasis on user requirements. The use of multimedia technologies is heavily supported in the development of the manuals, even as the emphasis is on development of user manuals for heavy duty mechanical equipment. The approach is also more one directional and not a dialogue with the stakeholders of the end product.

The data warehouse and data mining system proposed in this research were based on open source software. Due to the nature of open source, and the lack thereof an organization directly responsible means the availability of reliable product manuals may be a challenge. Consequently [7] and [8] propose a mechanism for automatic generation of configuration manuals and installation manuals respectively. This approach though only looks at one section of the product manual, which is the configuration while leaving the other important area of stakeholder training in form of a user training manual. The approach proposed by [8] also depends heavily on the initial use by an expert user of the system, while downplaying the input of the yet to be trained user. What happens, if the expert user suggested is not available? In an interdisciplinary research this might hinder user collaboration and contribution.

We address this and encourage learning by making use of Web 2.0 technology, through the use of wikis.

USING WIKIS FOR COLLABORATION IN DEVELOPING IPMS FOR DATA WAREHOUSE AND DATA MINING SYSTEMS

Wikis are browser based tools that enable users to actively participate in writing; editing and linking HTML based documents [9]. Wikis are based on the concept of openness allowing its users to edit and manage documents on the wiki system. Wiki users are able to modify, critique and add to contributions made by other users. This provides the opportunity for dialogue with the stakeholders.

As with any system development the use of wikis needs to have a clear business case. For the sake of the data warehousing system the goal was to ensure that usability of the system was enhanced for current and any new stakeholders as quickly as possible and with limited costs. Interdisciplinary research may have a plethora of stakeholders from different backgrounds, building a consensus amongst these different groups is vital to the research objective. This is evident in the area of participatory action research where

stakeholders are learning, collaborating and coming up with solutions to problems that affect them in teams [12]. The stakeholders are therefore involved in the action and ensuring a common agreement is reached on critical system issues cannot be overemphasized.

In our case the stakeholders of the data warehouse and data mining system were different with a strong need to adopt a participatory approach. These users include doctors, nurses, pharmacists, laboratory technicians, computer information system officers, researchers, development partners, civil society organizations and local government. All these categories of users may have convergent and or divergent needs from their requirements of the system. Product manuals therefore need to be able to address these different needs of the users effectively.

A wiki tool can be implemented on a web domain. The initial usage instructions are then loaded onto the wiki media, with comment sections. These can be categorized by the different operations that can be performed by varied class of stakeholders involved. The landing page is providing the objective of the wiki and the detailed instructions on how to edit and make changes to the information. User accounts for different categories can then be created as indicated in Table 4-II. These can then be invited to use the system and make changes and suggestions on the wiki depending on their usage experience of the system.

Table 4-II: User Categories in the Wiki Syste

Category	Number of accounts	Active
Researchers	10	4
Doctors	5	1
Nurses	10	4
Lab Technicians	4	2
HIV Counselors	8	2
Pharmacists	7	1
Development Partners	6	2
Civil Society	5	3
IT systems Officers	4	2

The main content area of the wiki was used to provide the instructions on the system functions. These different stakeholders had provided their expectations of the system in a previous research [3]. These expectations were used to organize the different categories of functions available to the stakeholders of the system.

The IPM development process can be depicted by the diagram below. The $user^X, Category^X$ indicates the different stakeholders involved.

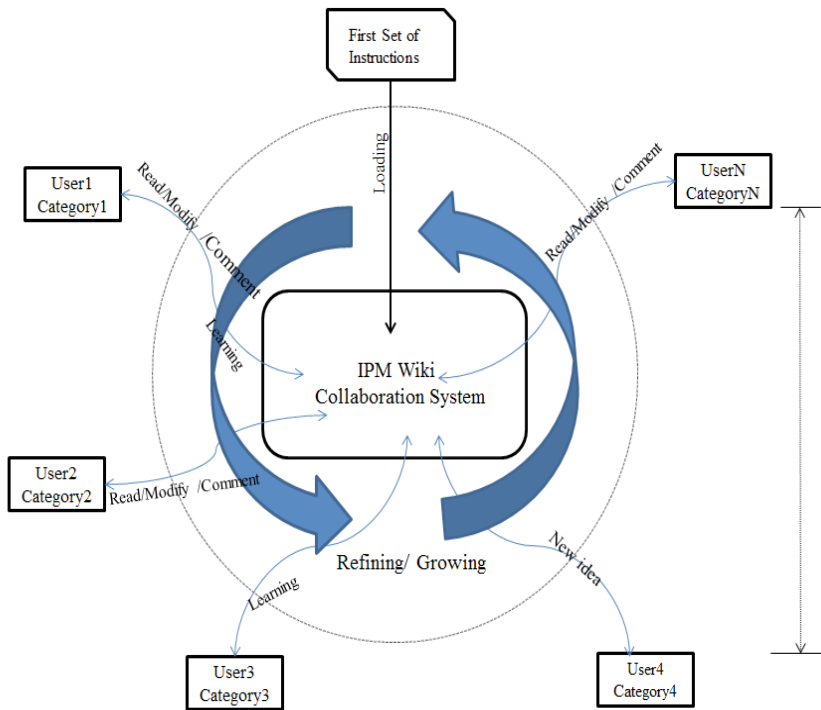


Figure 4-9: IPM Development Process

Steps in the process:

Stage 1: The first set of instructions are developed by the research group and loaded into the wiki collaboration tool.

Stage 2: Involves loading a set of N stakeholders into the system categorized by the different categories in the multi-disciplinary research. An initial list is provided in Table 4-II.

Step 3: This is the iterative process where the individual collaborators, read/edit and make comments on the first instructions submitted using the platform provided. The usage instructions therefore evolve as the different stakeholders edit and add information.

Step 4: Learning and idea generation. The stakeholders committed to using the platform are also made to learn as a result of reading the resources on the wiki. New ideas from certain stakeholders are captured and incorporated using the comment section. As the tool facilitates the growth of the manual the new ideas suggested provide opportunity for innovation.

CHALLENGES

A challenge of Web 2.0 technologies like WIKI is how to get participation, ensure interesting discussions and writings. Due to the mixed set of persons and interdisciplinary

nary nature of the research this compounds the challenge of web 2.0 technologies. The varied set of stakeholders may mean some people are afraid to put their comments out there. How do we get these different people to participate? Solutions to this are still ongoing for WIKIs, and not yet clearly documented and research on. However one approach could be the use of moderations. The system should have clear leaders spreading goodwill in the online community.

The lack of participation, the open nature of this approach and using participatory approach makes it open to monopolization by some participants or stakeholders. The monopolization is as a result of lack of participation from other stakeholders due to low self-esteem and problems of expression. Using Web 2.0 technologies like WIKIS may increase the difficulty in managing this. This is because of the open nature of the wiki tool and the freedom of access to the users. A user monopolizing the input may skew the IPMs generated to his view. Moderators can manage and track this using the history tool that tracks the changes that are being made in the system by the participants in the research.

In a third world country like Uganda the knowledge of ICT and access to internet resources places a strain on the use of web 2.0 collaborative technologies. The speed and progress of editing and adding contributions to the system end up being adversely affected. There is therefore a need for more training for the different disciplines of stakeholders and more patience and time to get on with the research. The use of multimedia technologies such as video tutorials and explanations on how to use the system is more effective than the conventional written instructions. This can easily be done with screen capture tools that log the actions on the screen.

The issue of ownership of the material generated at any point on the platform is difficult to determine. The wiki platform is open and a contribution of all stakeholders involved. The question then is what are the copyrights if any of the IPM generated through the WIKI? What happens if a user decides to use the IPM resources for their individual and for profit use and benefit? Since essentially freely distributed and with open access the issues of copyright [10] on wiki systems is still a challenge. This can be addressed through about GNU free documentation license [11]. This would essentially assure everyone of freedom of access. If the system is already open and freely accessible, then no one can take individual advantage to benefit from it, because it is already freely accessible.

TECHNIQUES FOR ENHANCING USABILITY OF DATA MINING AND DATA WAREHOUSING SYSTEMS

Effective product support and training for use of data warehousing and data mining systems is essential if the full potential of these systems are to be achieved. Highly knowledgeable stakeholders are more capable of innovative ideas of how to enhance the system capabilities through their requests for functionality to the system designers.

The need to enhance usability cannot be overemphasized. For this research some key issues that come to light are discussed below:

Research needs to make use of collaboration enhancing technologies such as web 2.0. These synergize the training and learning process of using information systems by enhancing contribution, building partnerships and consensus not to mention innovation and idea generation that foster system growth. PAR methodology is underpinned by collaboration and users learning with and through each other. This concept is enhanced through web 2.0.

Collaborative learning and idea generation that ensures instant publications can be used to enhance knowledge and improve the training progress and access to information if and when required by the end stakeholders. Simplifying the authorship of product support documentation such as user training manuals is important. This improves the feedback time required by stakeholders who might face challenges using the system. When the process of making changes to an intelligent manual is quick, then correcting an error is also quickly done and stakeholders updated accordingly.

Usability of data warehousing and data mining systems can also be enhanced granting all the stakeholders an opportunity to provide feedback on the information received and having their feedback taken into account. Wikis provide a platform for this using the comment section provided. Conflicting schedules make it easier to use wikis to track and generate comments in a participatory research as opposed to organizing a physical meeting with stakeholders.

Taking advantage of multimedia technologies is also useful. Providing stakeholders with simple videos and visual representations of how they perform simple functions is crucial. In this implementation we provide a link with a simple video on how to download free software to capture user actions on the screen. These simple how to videos can be shared with other stakeholders.

During learning some stakeholders may prefer formal trainers. These stakeholders feel at ease because of the idea of neutrality represented by formal trainers. In other cases people may be more comfortable with colleagues and partners than new formal trainers and especially if these trainers are using static user manuals. The use of a collaborative tool that makes use of participants in an interdisciplinary research offers a good opportunity for enhancing learning between the participants. The feeling of “If my colleague is doing it so can I”, helps to push stakeholders to use the system, and teach others what new “how to” that they have learned. Overzealous stakeholders in these different categories (Table 4-II) may monopolize the process and limit equal contribution from the rest.

CONCLUSION

This paper looks at enabling stakeholders in an interdisciplinary research to develop intelligent product manuals for a data warehouse and data mining system using web 2.0 technology. The need for these systems to grow and the stakeholders grow with it is crucial, both in the development of new requirements but also in the growth of the stakeholders knowledge on how to use the system. It is only from stakeholder’s fully comprehending the system that capabilities can be understood and intelligent requests

lodged for new capabilities and revisions to the initial system made. The requests for revisions are then made on informed decisions and from the perspectives of their different roles and different disciplines. This begins a process of using participatory action research online.

The need to use collaborative technologies that incorporate a range of multimedia can be used to jump start online participation by the end stakeholders. Intelligent product manuals should embrace the visual approach of communication more, and move a step further by involving the stakeholders. End users should be looked at as normal people and people tend to trust the familiar, the colleagues, the associate, and the research partners more than the formal structures. Web 2.0 technologies can go a long way in leveraging and using this trust for learning, idea generation and innovation.

[4] Reported on using AJAX technology to log stakeholder events while using the system. A further research area would be to enhance the automatic capture of user screens as the system is used. These can then be the subject of a user's post onto the wiki as their intelligent contribution to the manual. This would address the issue of speed for users who type and need to format the instructions clearly on the wiki. The integration of this platform with other more widespread social media technologies such as facebook would be of interest.

ACKNOWLEDGEMENTS

We would like to thank and acknowledge the support of Sida, BTH Sweden and Makerere University Uganda. We additionally thank the Research Division of Technoscience studies at BTH and the Departments of Electrical and Mechanical engineering in the College of Engineering Design Art and Technology (CEDAT) at Makerere University.

REFERENCES

1. C.D Otime, S.B Kucel, and L Trojer, Knowledge Discovery in Health Care using Data mining, *Proc. International Conference on Research in Engineering and Technology*, 2007
2. C.D Otime, S.B Kucel, and L Trojer, Dimensional Modeling of HIV Data, *World Academy of Science and Technology*, 2010
3. C.D Otime, *Participatory Approach to Data Warehousing in Health Care: Uganda's Perspective*, Lic. Diss. BTH, Sweden (Printfabriken, Karlskrona Sweden, 2011)
4. D.T Pham, S.S. Dimov, and B.J. Peat, Intelligent Product Manuals, *Proc. Instn Mech Engineers, IMechE 2000*, Vol 214 Part B, Page 411-419
5. R. Zhang, Design and Implementation of an Intelligent User-Manual Maker System, *International Conference on Computer Science and Software Engineering*, 2008 IEEE Computer Society 460-463
6. R.M. Setchi, D.T. Pham, and S.S Dimov, A Methodology for developing intelligent product manuals, *Engineering Applications of Artificial Intelligence*, 19 (2006) 657-669
7. Y. Murakami, E. Kagawa, and N Funabiki, Automatic Generation of Configuration Manuals for Open-Source Software, *International Conference on Complex, Intelligent and Software Intensive Systems*, 2011 IEEE Computer Society 653 – 658
8. Y. Murakami, N. Funabiki, H. Tokunaga, K. Shigeta and T. Nakanishi, A Web-based Installation Manual Management System for Open Source Software, *Fifth International Conference on INX, IMS and IDC*, 2009, IEEE Computer Society 1261-1266

9. K. Kear, J. Woodthorpe, S. Robertson and M. Hutchison, From forums to wikis: Perspectives on tools for collaboration, *Internet and Higher Education* 13 (2010) 218-225
10. N.A Polukarova, The Concept of Open Editing from the Copyright Viewpoint, *Automatic Documentation and Mathematical Linguistics* Vol 41 No 3 (2007)104-107.
11. GNU Free Documentation License. *Why free software needs free documentation*. (2008). <http://www.gnu.org/copyleft/fdl.html>
12. P. Lating, *Hybrid E-learning for Rural Secondary Schools in Uganda Coevolution in the Tripple Helix Process*, PhD Diss, BTH, Sweden (Printfabriken, Karlskrona Sweden 2009)

4.8 Paper VI

ENHANCING DATA WAREHOUSE CAPABILITIES BY AUTOMATIC ADDITION OF DIMENSIONS TO ESTABLISHED MODELS

Charles Daniel Otine^{#1}, Samuel Baker Kucel^{#2}, Lena Trojer^{*3}

[#]College of Engineering Design Art and Technology, Makerere University
P.O BOX 7062 Kampala

¹hautine@tech.mak.ac.ug, ²sbkucel@tech.mak.ac.ug

^{*}Research Division-Technoscience Studies-BTH

SE-371 79 Karlskrona, Sweden, ³lena.trojer@bth.se

ABSTRACT

In this paper we propose a technique of adding dimensions to an established data warehouse dimensional model. It is common to find systems designed with a set of fixed user requirements although these are seldom static. Over a systems lifetime, new needs are identified leading to software revisions, new versions and updates. These have their inherent costs and expenses. For data warehousing systems, at the onset requirements are clear and defined, forming the basis for analysis and the development of the data warehouse dimensional model. However as systems are used over time, users may require even more answers from what the data in the warehouse, and business processes may overtime begin capturing additional data that can be of interest to analysts. We are looking at a system of automatically capturing user requirements by adding new dimensions to an HIV/AIDS dimensional model. The data warehouse system was developed to monitor HIV/AIDS patients' adherence to therapy using decision trees, regression and clustering techniques of data mining. A module was developed that enables the addition of new dimensions to the dimensional model of the data warehouse, defined by stakeholders of the data warehouse. This enables direct management of the architecture of the data warehouse by enabling changes to be carried out on the dimensional model.

Keywords: Dimensional Models, Data Warehousing, HIV/AIDS, Data Marts

I. INTRODUCTION

In a previous study [1] a dimensional model (DM) for monitoring patient adherence to antiretroviral therapy (ART) was developed. ART refers to the treatment of HIV infected individuals using antiretroviral drugs (ARV) to suppress the HIV virus and stop progression of the HIV disease [2]. As in the treatment of any disease adherence to the medication is crucial. In the case of ART failure to adhere to therapy even just occasional non adherence has dire consequences to the patient's treatment progression [3], [4] and [5].

The DM is based on dimensions centred on different concepts in the treatment environment of the HIV patient. These facilitates the analysis and reporting on the two basic processes monitored during therapy namely: i). Periodic prescription of media-

tion and ii). Medical check-up and testing done to detect, diagnose and monitor the disease. The DM is typically locked or in other words in read only mode after the implementation of the data warehouse, this means that at the end of the implementation the data warehouse model is supposed to remain the same. To ensure longevity of the data warehouse systems and for the need to address new requirements, a technique of adding new dimensions to the original model is required.

Maintenance of data warehouse systems is required to ensure that the needs of the stakeholders are being addressed. As the system is continually used, the traditional approach of system maintenance involving technical teams reviewing the requirements and adding new dimensions afresh as and when required may end up being a hindrance to system growth. This may come at great cost of money and time not to mention the prerequisite of having technically competent persons who understand the architecture of the system. In an adherence monitoring system the stakeholders using the system need to be able to add new dimensions to the underlying data warehouse model more cost effectively. The resource constrained environment of the research calls for an approach that is cheap and with more stakeholder ownership. This would give the capability to continuously modify the dimensional model of a warehouse by its end users depending on their needs.

The paper provides an overview of the previous research that culminated in the dimensional model and presents this model. The next section deals with changing dimensions linking it to practical approaches in the HIV/AIDS data warehouse. An automatic technique of adding dimensions to established models is then presented with challenges to dimensional modelling. A conclusion is then provided with identified areas for further research.

II. HIV/AIDS DATA WAREHOUSE DM

This being an interdisciplinary participatory action research a number of stakeholders were used to identify requirements for an HIV/AIDS data warehouse to help monitor patient adherence to ART [1]. The research stakeholders included doctors, nurses, pharmacists, laboratory technicians, computer information system officers, researchers, development partners (DPs), civil society organisations (CSO) and local government (LG).

The process identified 2 key facts centred on 6 key dimensions to be tracked for each patient undergoing therapy. The model generated was based on the star schema with facts centred on dimensions. The time dimension was vital in tracking the periods the patient underwent a medical test versus the prescription given to the patient. The star schema was chosen for its flexibility in allowing for modifications as the data warehouse grows. This research however seeks to move this flexibility from just the technical data warehouse personnel to system stakeholders. Given that stakeholders are directly or indirectly involved with the system, their input on requirements can be harnessed by giving them flexibility in automatically adding new dimensions to the dimensional model.

Given the sometimes significant costs of data warehousing systems [6] there is a need for affordable alternatives. This is needed most in resource constrained settings like Uganda where this research is based. The first dimensional model was therefore developed in open source software to minimize software costs. This research moves a step further in managing the costs that would come as a result of software requirement changes in the future. Automating the process of changing the dimensional model with minimal intervention of information technology professionals and technical review task force would go a step in this process.

Figure 4-10 presents a summarized view of the dimensional model.

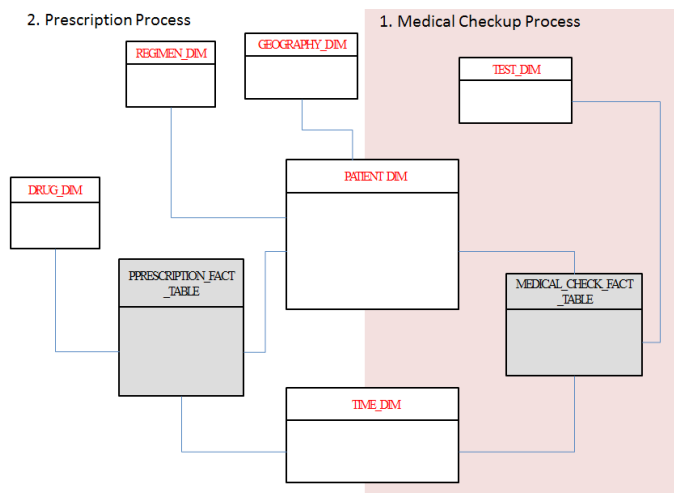


Figure 4-10: Summarized view of Dimensional Model

The dimensional model was verified on different open source database management systems. This involved checking that the schema generated was a correct representation of the model. This involved analysis of user requirements against the model generated, to establish whether answers required from the data warehouse could be generated from the dimensional model in question.

The medical_check_fact process monitors patient information such as CD4 count, weight, and pregnancy for women, haemoglobin values across the time dimension. Different test results are also monitored depending on a medical personnel recommendation. The prescription_fact process in the model monitors the distribution of ARVs to patients, with each patient belonging to a certain drug regimen for a period of time. This is also monitored over the time dimension.

III. CHANGING DIMENSIONS

Dimensions in a data warehouse refer to a category of data in the stated data set. In a data warehouse, dimensions in the DM may change over the data warehouses' lifetime. Dimensions can change in one of two ways, either by having new dimensions added

to the dimensional model, or having what is referred to as slowly changing dimensions [7] and [8]. Slowly changing dimensions are basically values in the dimension that change over time.

Slowly changing dimensions may be dealt with in 3 ways in a data warehouse. The first approach involves overwriting the value in the dimension, in which case the previous information is lost. The second approach would be to insert a new record in the dimension table indicating the change in the dimension. With this approach the new entry becomes the effective value for that dimension. This second approach includes an effective date and obsolete date to mark the history of the dimension value. The third approach is the addition of a new column in the dimension to maintain the history of the dimension. A practical implementation of this with the model can be observed when we analyse the PATIENT, REGIMEN dimensions and marital status (mstatus) in the patient dimension. This record can be depicted as the following select record¹² in Table 4-III

Table 4-III: Patient Dimension Record

Number	Name	Regimen	mstatus
10001	George	1	single

A patient is assigned to a treatment regimen and over a period of ART the patient's ART treatment regimen can change. It is also feasible that the patients' marital status could change over a period. The first approach of dealing with this dimension change would involve over writing the information. In the above patient's case if his change of therapy was made from regimen 1 to say regimen option 2, then the new information would reflect 2. However this approach loses the history of the information, in that we would from then on not be able to know what regimen the patient therapy began with initially.

The second approach of dealing with this would involve the addition of a new record and the dimension also including effective (effDate) and obsolete (obsDate) dates. In this approach we would have a record similar to the one in Table 4-IV

Table 4-IV: Patient Record Changing Dimension (Method 2)

Number	Name	Regimen	mstatus	effDate	obsDate
10001	George	1	single	2/2/2008	2/2/2011
10001	George	2	single	2/2/2011	NULL

The two entries would then have similar patient numbers but with differences in effective dates with the older of the records already having an obsolete date indicated in the record. The history between the two values in the dimension is therefore maintained. In the above example we would also maintain a similar effective and obsolete dates for the variable marital status should we wish to track its change over time.

¹² Patient name (otherwise the field is obfuscated) is indicated to help easier understanding of the example; many columns in the dimension have been truncated to only show those relevant to the example.

The third approach to managing slowly changing dimensions can be depicted by the Table 4-V. It involves adding new columns to the dimension. This then indicate the different successive changes in that dimension value. At the system design the amount of history that we need to maintain and follow can be decided. This is then set, for instance for the patient dimension we can specify that we would only maintain up to 3 regimen changes during therapy. Any further regimen changes are then captured by over writing the first record.

Table 4-V: Patient Record Changing Dimension (Method 3)

Number	Name	Regimen1	Regime2	Regimen3	mstatus
10001	George	1	2	NULL	single

In the above example the most effective regimen would be Regimen2 while the NULL indicates that a third regimen change has not yet been effected. This indicates history for the regimen in the patient dimension. Using the same approach a change in the marital status can also be tracked on the patient dimension, with new changes being monitored as columns.

The three approaches mentioned can be combined into a hybrid technique of monitoring dimension changes. This would include adding columns and monitoring across effective dates and obsolete dates for the changing dimensions. In ART therapy patient history is key in monitoring adherence, changes in the patient's life and treatment options need to be tracked over the therapy. It is therefore imported to use the hybrid technique and maintain both the change in the dimension being monitored versus the time the change was effected.

IV. AUTOMATIC ADDITION OF DIMENSIONS

In the above section we examined the scenario of a dimension changing; however there are situations where a totally new dimension needs to be included into the patient data warehouse. This would entail analysis, and changes to the dimensional model.

The crucial point to observe whenever dimensions are being added to the dimensional model automatically is to ensure that the granularity of the different fact tables being monitored should not be altered. The granularity or grain is defined as the lowest level of information that will be monitored for each of the processes being monitored [9]. In this context the lowest level of prescribing medication and medical checkup on the HIV/AIDS patient should not be altered. Performance of the final model in face of analysis operations should also be considered whenever dimensions are being added to the dimensional model.

A GUI (Graphical User Interface) module was developed for the system named the dimension manager (DMG). This has capability for the user to select and define a new dimension for inclusion into the dimensional model in addition to any attributes for the defined dimension. The process of addition of the dimension goes through the iteration as indicated in Figure 4-11.

The user has to the option to specify from the GUI system what data mart the dimension interacts with and or whether the dimension being added to the star model is a

conformed dimension. Conformed dimension is a dimension that is the same across more than one data mart of the business process and as Kimball is quoted in [10] the conformed dimensions help with analyzing multidimensional data. Then a set of iterative steps are performed whereby the user indicates the different attributes specific to the dimension in question. The data type for each attribute is also clearly indicated. The system then automatically connects to the DBMS system and scripts are created for the generation of dimensions and the automatic linking to data marts. The first phase involves the creation of the dimension with all the relevant attributes and data types as specified earlier. This operation is carried out through structural query language (SQL), which is generated from the selections made by the user using the web-based GUI defined earlier.

The second phase of the operation involves linking of the generated model to the data marts. This is a one step process in the event that the dimension created is not a conformed dimension. In the event that the dimension created is a conformed dimension, then it has to be sequentially linked to each of the facts or processes in the model while observing the grain of each process. For the case of the dimensional model under study here, it would need to be linked to the two processes: i). Prescription and the ii). Medical checkup processes. This involves the addition of the new dimensions primary key as a new column in the particular fact table being changed.

This is then cascaded across all the fields in the fact tables representing the processes being modeled for each data mart. At the point of selecting the dimension to be added automatically to the dimensional model the system request the user to setup an initial default value for all the entries that are already in the data mart. The value for all the entries is then setup to this initial value to indicate the starting value for analysis across the new dimension.

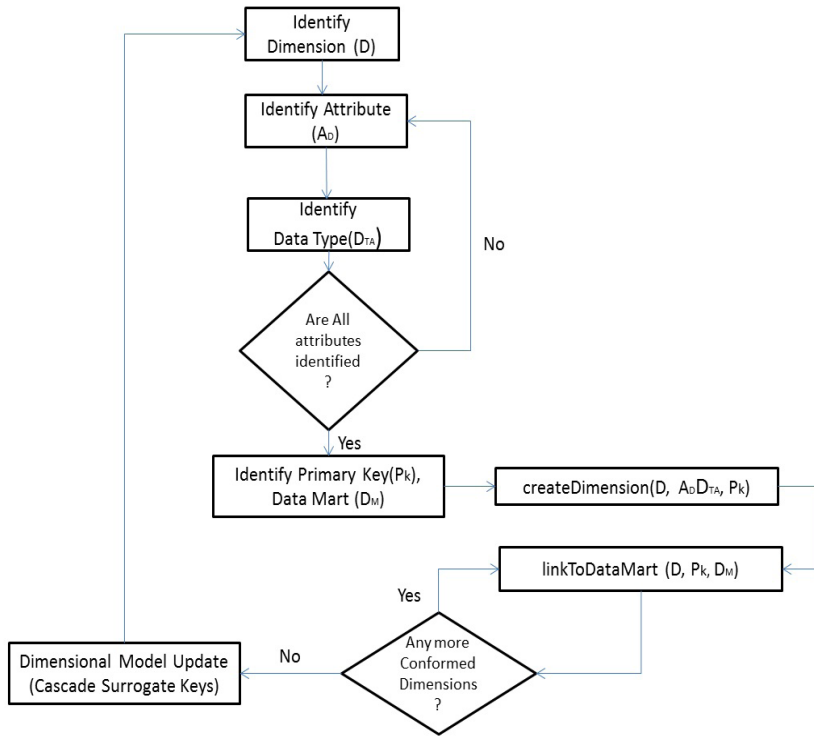
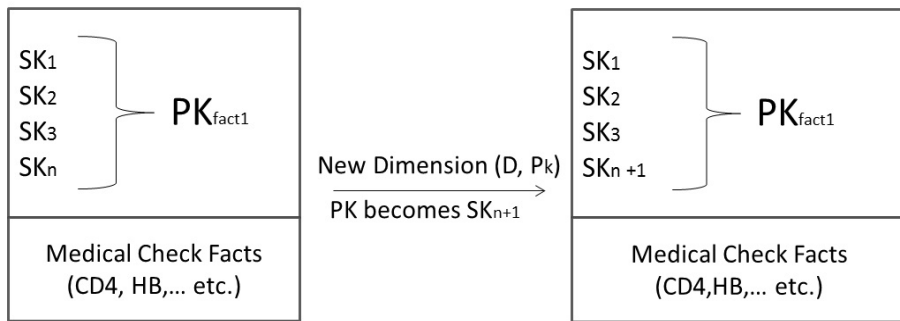


Figure 4-11: Process of addition of new dimension

The DMG module then updates the resultant model accordingly and writes the changes to a temporary database. The newly defined dimension and the attributes are also noted and an update to the data dictionary for the HIV/AIDS data warehouse is done. A final approval process by a data manager then commits the changes directly to the data warehouse and the new dimensional model takes effect and new analysis of the data can be done across the added dimension(s).

Figure 4-12 below depicts the adjustments that are automatically done on the two processes under study. Depending on the conformity of the dimension being added to the module, the adjustment to the group of surrogate key (SK_n) making up the process primary key is then done. In this case the defined primary Key (P_k) from the DMG is then added as the next composite primary key to the fact table that forms the new identifier for the updated fact table.



PK_{fact1} :Primary Key for Data Mart represented by fact table fact1

Figure 4-12: Changes to the process fact tables

When the above automation is being carried out there are some underlying considerations that need to be put into perspective for example the performance aspects as alluded to earlier. If not managed correctly the functionality of automatic addition of dimensions to the dimension model can result in a significant increase in the number of dimensions in the dimensional model.

Since each of the new added dimensions will then have its primary key included as a surrogate key in the composite primary key for the dimensional fact table then resultant fact table becomes extremely large. This therefore has a direct effect on the performance of analysis queries that would be run on the data warehouse. Kimball [11] suggests that when the number of dimensions for a single fact goes above 25 then there is a need to reexamine the dimensional model design. This check is done by the DMG and also the reason for the review phase before committing the changes made to the model.

V. HALLENGES TO AUTOMATIC ADDITIONS

The challenge of a very large fact table with very many dimensions has been identified as a potential bottleneck that may be faced. The review process before committing to changing the dimensional model should help to manage the process of dimension addition. New dimensions should only be automatically added to the dimensional if they will address information gaps that cannot already be addressed and handled by the current dimensions. Even then when absolutely necessary dimensions can still be added and measures of handling overly large fact tables can be employed. This including techniques such as using multidimensional database systems [12], indexing (bitmapped indexes and join indexes) as reported by Poolet quoted by [13] and also be employing the use of materialized views for analysis. Large fact tables can also be partitioned for easier management as reported by [14].

The complexity of the analysis and queries that can be performed on the data warehouse is increased with addition of each dimension. This is compounded by the fact that this being an interdisciplinary research with different stakeholders with many

needs and expectations of the system. There is therefore the ever existing challenge of managing these many different expectations and ensuring that the resultant system and dimensional model is not overly complex to hinder analysis. The need to judiciously know limits and know that the possibility of including all possible dimensions in the spectrum may not be possible.

Maintaining consistency and granularity is also important, especially when dealing with non-technical users. They need to be made to understand the concept of granularity and the level of detail that each of the processes in the different data marts must maintain. This calls for continuous training of the users on DMG.

VI. FURTHER RESEARCH AND CONCLUSIONS

The DMG module moves the optimization of dimensional models into the hands of the end users. Giving the end users more ownership of the system and ability to configure and include the dimensions that they find interesting. This does two things, firstly demystifying the data warehouse dimensional concept as being a focus area for the information technology experts like data modelers and database administrators. By providing capability for changes to the model by the end users, the research also addresses the resource constrained setting of the research. Given the limitation of “IT expertise” to constantly release and bring updates to the data warehousing system a need to build capacity of people who interface everyday with the system is important. This also addresses the gap of turnover of stakeholders who understand the system. If broad spectrums of people are already involved in making adjustments to the dimensional model, this turnaround time can be significantly shortened and will not also adversely affect the operation of facilities making use of this system.

The DMG brings the stakeholders closer to the architecture of the data warehouse system and makes for easier changes to requirements. This also improves learning among the users and brings about some questions that may make interesting areas for further research in terms of the dimensions created by stakeholders. The question of whether handing over system ownership at the architecture level actually does improve the quality of design products generated as the users innovatively try to add new dimensions.

The management of the process by the users is also important especially with the challenge of excessive dimensions being created. Data warehouses have typically been into systems with data being loaded into the system and rarely being removed from the system. Excessive dimensions may call for removable from the system, in the event that these are found to be counterproductive. This would involve reversing the flow of the DMG module to undo the change made, however this is only simple and feasible if it is being reversed immediately and is already addressed by the module. However a situation where the reversal is after several intervals of data loading into the data warehouse calls for further analysis.

The concept of updating the dimensional module can also be moved to include a primary dimensional model (D). The new additions to D are created as copies to versions of D for different stakeholders.

$$D = \{D_1, D_2, D_3 \dots D_N\}$$

This would always leave the underlying model D unchanged and research on the different iterations and their effectiveness can be analyzed over time.

ACKNOWLEDGEMENT

We would like to thank the assistance of Sida for sponsoring this research and the collaborating Institutions of Blekinge Institute of Technology and College of Engineering Design and Art at Makerere University.

REFERENCES

1. C. D. Otine, S. B. Kucel and L. Trojer, "Dimensional Modeling of HIV Data Using Open Source," *World Academy of Science Brazil*, vol. 63, no. 1, pp. 156-160, 2010.
2. WHO, "Antiretroviral therapy," 2011 11 2011. [Online]. Available: <http://www.who.int/hiv/topics/treatment/en/index.html>. [Accessed 20 11 2011].
3. WHO, "Patient Monitoring Guidelines for HIV Care and Antiretroviral Therapy," 2006.
4. C. H. Hinkin, T. R. Barclay, S. A. Castellon, A. J. Levine, S. R. Durrasula, S. D. Marion, H. F. Myers and D. Longshore, "Drug use and Medical Adherence amongst HIV-1 infected," *AIDS Behaviour*, vol. 11, no. 2, pp. 185-194, 2007.
5. D. L. Paterson, J. Mohr, M. Brester, E. N. Vergis, C. Squier, M. M. Wagener and N. Singh, "Adherence to Protease Inhibitor Therapy and Outcomes in Patients with HIV Infection," *Ann Intern Med*, vol. 133, pp. 21-30, 2000.
6. M. I. Hwang and H. Xu, "The Effect of Implementation Factors on Data Warehousing Success: An Exploratory Study," *Journal of Information, Information Technology and Organisations*, vol. 2, pp. 1-14, 2007.
7. M. Rifae, K. Kainmehr, R. Alhaji and M. J. Ridley, "Data Warehouse Architecture and Design," in *Information Reuse and Integration IEEE Conference*, 2008.
8. R. Kimball and M. Ross, *The Data Warehouse Toolkit: The complete guide to dimensional modelling*, 2 ed., NY: John Wiley & Sons Inc., 2002.
9. L. Viking, "Dimensional Modeling Basics," 22 03 2006. [Online]. Available: <http://www.sqlmag.com/article/data-modeling/dimensional-modeling-basic>. [Accessed 3 4 2012].
10. D. Riazati, J. A. Thom and . X. Zhang, "Drill Across & Visualization of Cubes with Non-Conformed Dimensions," Sidney, 2008
11. N. Gerard, "Dimensional Modeling - Dimension," gerardnico.com, 07 10 2011. [Online]. Available: http://gerardnico.com/wiki/data_modeling/dimension. [Accessed 25 05 2012].
12. O. M. G. OMG, "The CWM Multidimensional Metamodel Description," 2005. [Online]. Available: <http://help.sap.com/javadocs/NW04S/SPS08/bi/cwm/org/omg/cwm/resource/multidimensional/package-summary.html>. [Accessed 25 05 2012].
13. S. Jamil and R. Ibrahim, "Performance analysis of indexing techniques in Data warehousing," in *2009 International Conference on Emerging Technologies*, 2009.

4.9 Summary of Papers

Paper I: In this paper the concept of knowledge discovery and data warehousing concepts in health care was introduced. A survey of the status of health information systems in Uganda revealed the need for electronic medical records that facilitate deeper analysis and critic of ART therapy. The study revealed the emergence of a plethora of mobile applications used in health, with a focus on advocacy and information sharing. Expert results from studies can easily be disseminated across to stakeholders and care providers with the mobile infrastructure.

Paper II: The dimensional model for ART is presented, as a star schema centered around two fact tables. Different open source dimensional modeling tools were reviewed and used with their limitations highlighted. A data warehouse dimensional model focused on ART management for HIV patients is produced centered on prescription and periodic medical checkup and testing on the patients.

Paper III: This focused on the development of a data warehouse. Open source was adopted as the technology taking into consideration the cost of proprietary software in resource constrained settings. The data warehouse reporting mechanism caters for decision support to policy and strategy indicating usage statistics of ART and requirement estimates in the future.

Paper IV: This focused on the analysis and definition of system requirements through collaboration of all stakeholders involved in the research. This is important to ensure that the system is accepted by the end users. The success of data warehouse systems is dependent on the level of stakeholder acceptance, since their long term success depends on incremental aggregation of data from the different dimension areas. The involvement of all stakeholders in defining requirements and through the system development also ensures that training on using the system becomes easier in future. 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.

Paper V: This focused on the human resources for the system sustainability in terms of training and bringing new stakeholders onboard on how to use the system. It looks at enabling stakeholders in an interdisciplinary research to learn from each other and de-

velop intelligent product manuals for the data warehouse and data mining system. The research adopted the use of collaborative WEB2.0 technology of the wiki to facilitate this interaction between the different stakeholders involved in the research.

Paper VI: This focused on ensuring data warehouse growth through the automatic addition of new dimensions. This means the expansion and growth of the data warehouse and related analysis capabilities are placed in the hands of the stakeholders. A dimensional manager tool was developed to ensure to manage the process of adding new dimensions to the data warehouse. This moves the data warehouse from the previously predominantly read only mode of operation to some limited write capabilities.

Chapter 5

ADDITIONAL RESULTS

5.1 Introduction

This chapter presents additional results that have not been captured in the papers in chapter Fel! Hittar inte referenskälla.. These relate to the depiction of different aspects of the system, samples of the wiki systems for users, decision support for supply chain estimates of the usage of ARV. The last two sections are on the regression and classification algorithm for CD4 count to viral load estimation and treatment failure prediction.

5.2 Additional Results

5.2.1 Overview of System Architecture

The system architecture is made up of three main components. The Data Store, the data mining platform and the front end user interface. The front end user interface developed using PHP presents the interface between the data store and the data mining platform. The data store is the data warehouse containing the integrated data from the different health care source systems. The data platform was implemented in the open source platform Octave (Eaton 2012). A simple interaction is given in Figure 5-1.

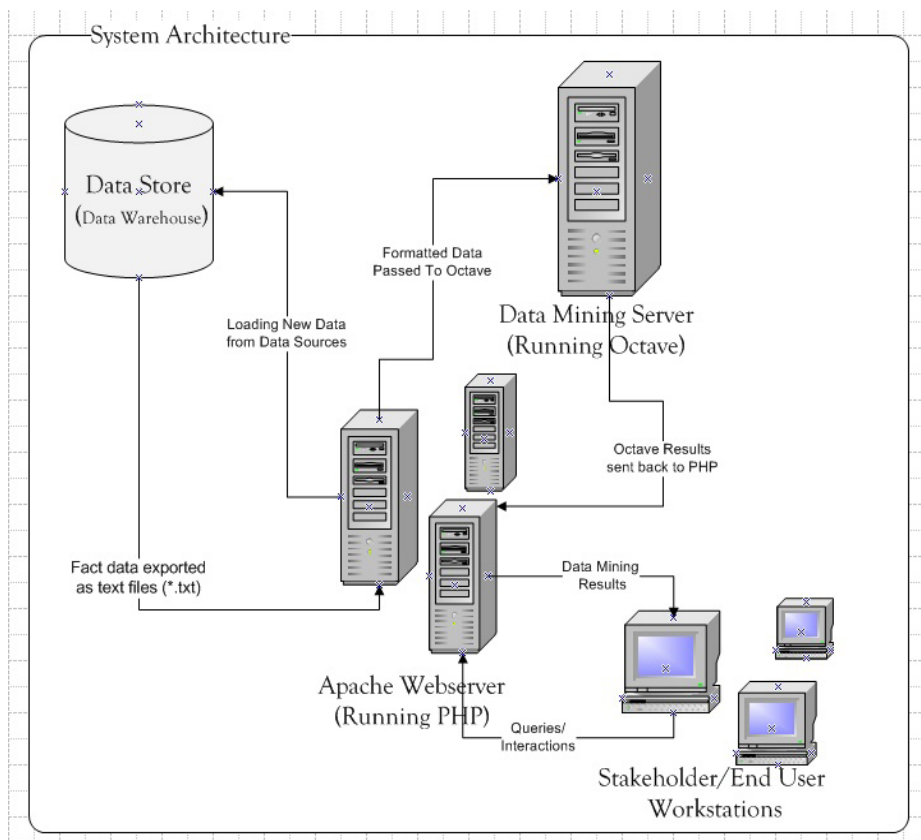


Figure 5-1: System Architecture

The Apache Webserver running PHP acts as the communication link between the data warehouse, the data mining server running octave and the different stakeholders using the system. The PHP server is used to extract a given subset of data from the fact table in the data warehouse; this is then formatted in text files readable by the octave data mining tool. Octave server then can perform the different operations on the data passed to it including running the different regression algorithms for forecasting and classification in addition to optimization of the values being generated from the fact table (these operations are examined in later sections of this chapter).

After Octave has completed running the operations the results are then passed back to the PHP server which then displays it them to the end user. The optimization and machine learning functions are developed and written in octave and only take queries from PHP which in turn captures this from developed user interfaces. The optimization values generated from the machine learning algorithms in the octave server when passed to the PHP server are recorded in the data warehouse for use in new estimate queries by the expert end users.

PHP Script depicting an Interaction between the Data Warehouse, Octave and the Webserver

```
<?php
/*
```

Export data from mysql database to a text file separated by commas

Octave will read from the exported file and upon execution of the script by Octave, the file will automatically be deleted to free disk space and for security reasons

Note use the “D:\xampp\tmp” folder and then the file will be deleted/unlinked.

```
*/
    $path_to_temp_dir = “D:\\xampp\\tmp\\”; // Octave files are handled
in this folder during passing PHP functions
    $path_to_octave=”C:\\Octave\\3.2.4_gcc-4.4.0\\bin\\octave”; //Specify
the location of the Octave executable
    $filename = round(time() / 10 * rand(1,10));
    $export_filename = round(time() / 20 * rand(1,10));
    $filename = $path_to_temp_dir . $filename;

    /* Function to pass Octave expression
    to .m octave file via PHP
    */

function octave_calc($expression){
    global $filename;
    global $path_to_octave;
    if ($expression){
        $script_octave = fopen($filename . “.m”, “a+”);
        fputs($script_octave , $expression . “\n”);
        fclose($script_octave);

        $ans = exec($path_to_octave . “ -q “ . $filename . “.m”); //
passthru() function can be used instead
        //$ans = explode(“=”, $ans);
        //$ans = $ans[1];

        //$fr = $ans;
        unlink($filename . “.m”);
    }

    return $ans;
}
```


//Get the parameters from the HTML/PHP form

```
$y=stripslashes($_POST['yparam']);
$X=stripslashes($_POST['xparam']);
$theta=stripslashes($_POST['theta']);
/* Test parameters
$y=10;
$X=30;
$theta=0.8;
*/
// From here the calculations are made using the GNU Octave
//Octave functions which will be rewrite in an .m file and then executed
against the octave.
//$command="function [J, grad] = costFunction(theta, ".$X.", ".$y.");

$command="m=length($(".y."));
$grad="zer=zeros(size($(".theta.")));

$ngrad="ngrad=1/m* sum((-".$y." .* log($(".X.*".$theta."))) -
(1-".$y.").*log(1-($(".X.*".$theta.")))));";
$ngrad.="\\n";
$ngrad.="n1grad=1/m * (($(".X.*($(".X.*".$theta.)-".$y.")));";
$ngrad.="\\n";
$ngrad.="ans=givens(1,1);";
$ngrad.="\\n";
$out1="D:\\csv1.txt";
$out2="D:\\csv2.txt";
$out3="D:\\csv3.txt";
$ngrad.="csvwrite($(".out1.",ngrad);";
$ngrad.="\\n";
$ngrad.="csvwrite($(".out2.",n1grad);";
$ngrad.="\\n";
$ngrad.="csvwrite($(".out3.",ans);";
```



```

/****
    Rewrite the Octave function in a temporary file and run it via command
line on Octave and return results
    ***/

    $m = octave_calc($command); // calculate m if it is a scalar
    $cgrad = octave_calc($grad); //calculate ser f it is a scalar and pass it to the
function/equation directly
    $final_result= octave_calc($ngrad); //rewrite the ngrad function in an
octave file (.m) and execute it while saving results in files.

/****
Extract octave results from a text file separated by commas.
It extracts line by line and the split and trimmed to independent values.
If it a n x n matrix, line one is split and extracted to values.

**/
    $csvfile=fopen("D:\csv3.txt","r");
    $count=0;
    echo "The Matrix is:<br>";
    while(!feof($csvfile))
    {
        $csvred=fgets($csvfile);
        // Obtain matrix rows
        $lines=explode(",",trim($csvred));

        // Obtain independent values in each column
        foreach($lines as $line)
        {
            $count++;
            $line=trim($line);
            echo "<br>Result: ".$line;
            /*if (strlen($line))
                continue;
            $array[]=explode(",",trim($line)); */
        }
    }
    fclose($csvfile);

```



```
/****
```

1 - Automate results file creation and deletion (csvX.txt) files to free space upon displaying results.

Soln: Use memory or temporary file or
feed results into a temporary database tables and upon using the
results,
free the table too or keep it for reference

```
*** /
```

```
/******
```

```
DATABASE EXPORT
```

```
*** /
```

//In PHP, this code will export data from the database and load it to a text file:

```
$username ="root"; //db username
$password=""; //db password
$dbhost="localhost"; //db server name
$dbname="DWH"; //specify the database
$d="D:\sqldata\data\//"; // The export directory, it should be created at
the installation of this script
$export_filename = $d. $export_filename;
$selected_fields="number"; /* use * to select all (default), otherwise enter
fields separated by commas */
$selected_table="patient"; //Table whose data is to be exported
$con=mysql_connect($dbhost,$username,$password) or die ("The
connection to the server failed!");
mysql_select_db($dbname,$con) or die ("The connection to the database
failed!");
// $dataFile=fopen($export_filename.".txt", 'r+');
$export_query="SELECT ".$selected_fields." FROM ".$selected_table. "
INTO OUTFILE '".$export_filename.".txt' FIELDS TERMINATED BY ',' ";
//echo $export_query;
$result=mysql_query($export_query,$con);
/* Do
Octave calculation
via PHP
*/
//octave_calc($ngrad);
```



```

if($result){

    echo "<br>The data export tool was successful.";

}
else
{

    echo "<br>The data export tool was not successful.";

}
mysql_close($con);

//fclose($dataFile); //close the export file after executing the Octave data.
//unlink($export_filename."txt"); //Delete the export file to free disk
space

//END OF SCRIPT//

?>

```

5.5.2 Sample Screen shots

Sample system screen shots are provided in the sections below. This represents a selection of a few system stakeholders.

The system limits access through authorized and preapproved users managed by system administrators.

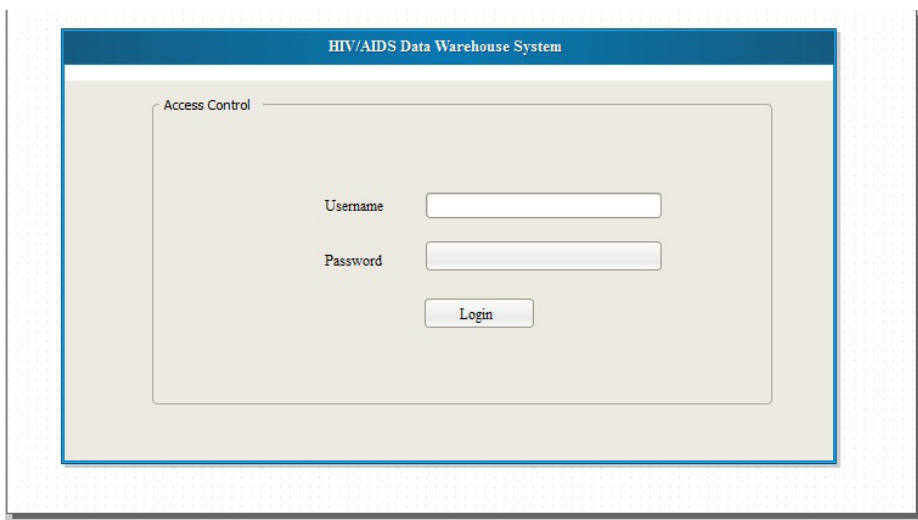
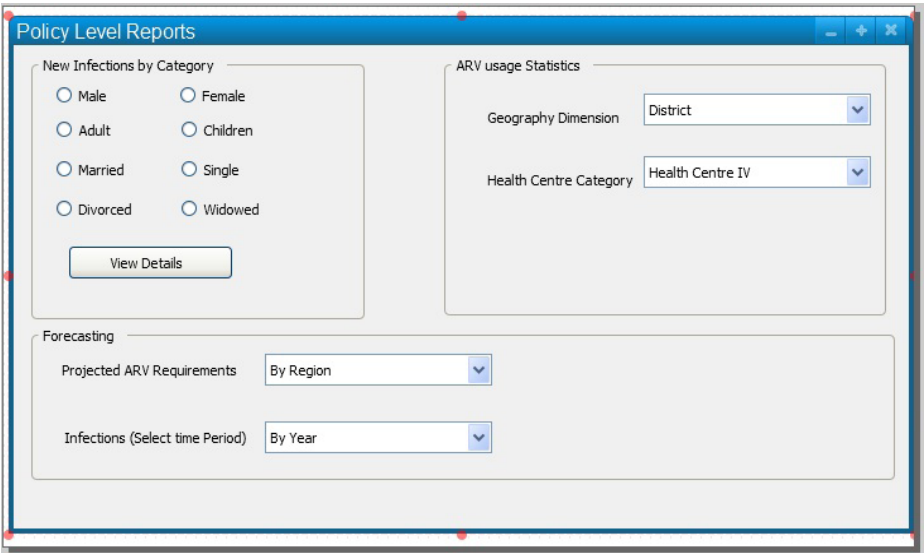


Figure 5-2: Security Options

The reporting tool provides summaries of reports for policy stakeholder. Categorized projections of HIV infections can be viewed as summarized reports, in groups by gender, marital status, occupations and age group segments (Figure 5-3). The usage statistics by geographical dimensions can be generated by selecting the appropriate region to display the uptake (Figure 5-3). The Medical report module provides the option for assigning treatment options and the using the regression and treatment classification to analyze treatment failure (Figure 5-4).



Policy Level Reports

New Infections by Category

☐ Male ☐ Female

☐ Adult ☐ Children

☐ Married ☐ Single

☐ Divorced ☐ Widowed

ARV usage Statistics

Geography Dimension:

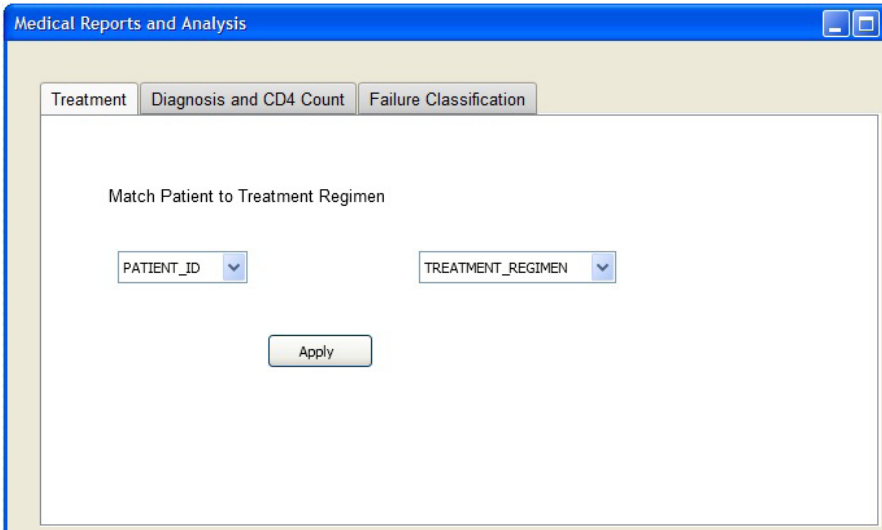
Health Centre Category:

Forecasting

Projected ARV Requirements:

Infections (Select time Period):

Figure 5-3: Patient Reporting Screen and ARV Statistics



Medical Reports and Analysis

Treatment Diagnosis and CD4 Count Failure Classification

Match Patient to Treatment Regimen

PATIENT_ID TREATMENT_REGIMEN

Figure 5-4: Patient Medical Report and Regimen Options

Dimensional manager tool allows for more control of the data warehouse architecture to lie with the stakeholders. The tool enables different stakeholders to add dimensions to the data warehouse model by providing a series of sequential steps to add the dimensions.

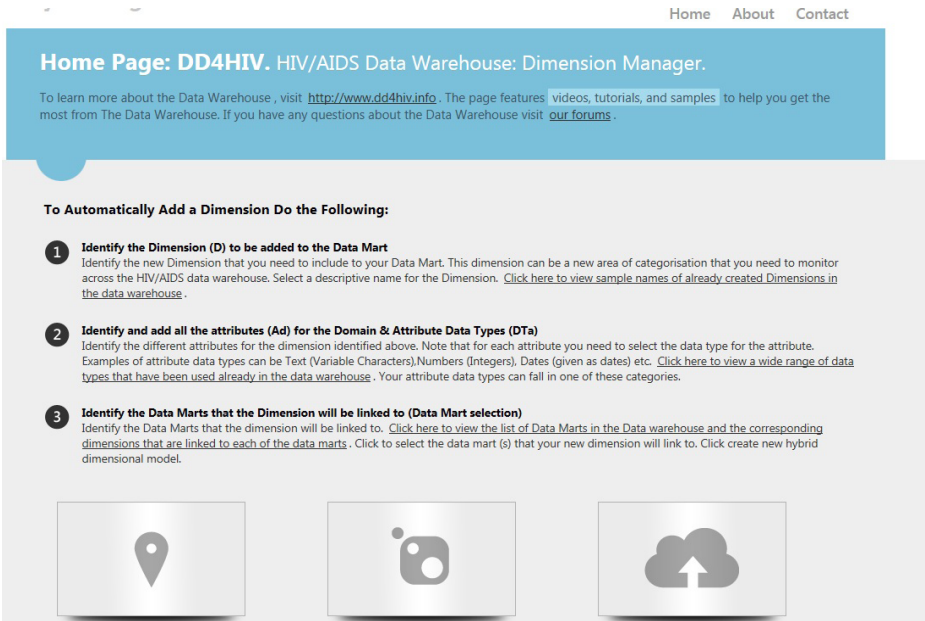


Figure 5-5: Dimensional Manager for Ensuring Adding Data warehouse Dimensions

The dimensional manager tool was implemented in close collaboration with the stakeholder collaboration wiki tool. The implementation of the wiki tool was done on a registered domain and selected users invited to collaborate on using the system and sharing experiences. Figure 5-6 presents a snapshot of the wiki system.

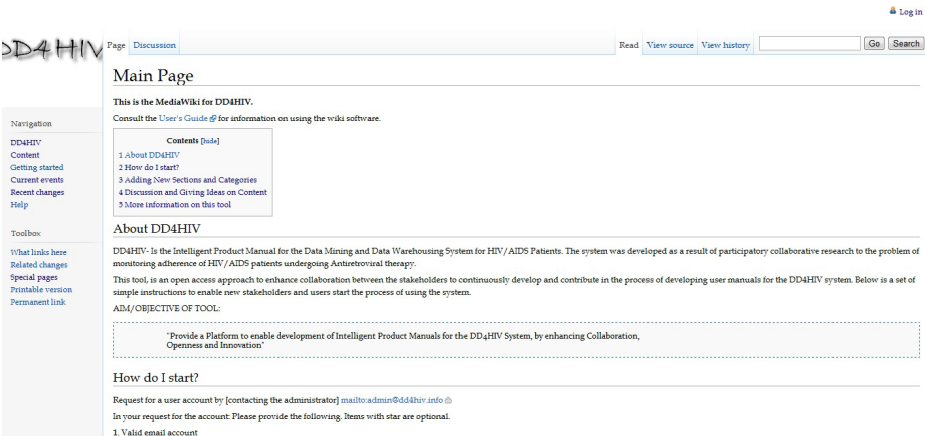


Figure 5-6: Wiki System Screen shot

5.2.3 Materialized Views

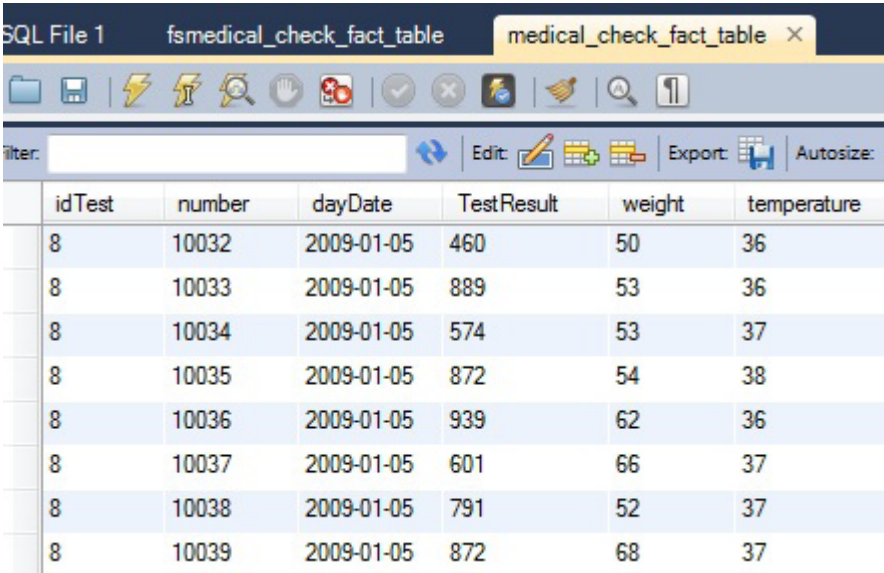
Materialized views were used to facilitate quick access to data required during training the regression algorithm. This enabled calculation and storage of feature averages and standard deviations. These figures were then later used for feature scaling of CD4 entries.

An example was in the analysis done on the medical check fact table. Feature scaling was used because of the sometimes incomparable units of the measures in the medical check fact table. This occurs when measuring different features such as CD4 count results, patient weight, temperature, HB test results, and viral loads. Feature scaling was done to ensure that the machine learning algorithms used standardized measures. The normalization formula for a feature variable X was adopted in the data warehouse

$$X = \frac{X - \mu}{\sigma} \quad \mu = \text{mean and } \sigma = \text{standard deviation}$$

Given the number of entries in the fact table for each patient, the ever changing mean of the feature and the standard deviation given new additions into the data warehouse (especially during ETL) a materialized view was adopted. This was refreshed at each load of the data warehouse with new data using procedures in the database. The snapshots (Figure 5-7 and Figure 5-8) below present snapshots of the data warehouse before and after applying feature scaling to 3 features in the fact table. The test result refers to the result of a CD4 count test carried out on clients.

We also present a section of the procedure that updates the scaled fact table for every update during loading of the data warehouse.



idTest	number	dayDate	TestResult	weight	temperature
8	10032	2009-01-05	460	50	36
8	10033	2009-01-05	889	53	36
8	10034	2009-01-05	574	53	37
8	10035	2009-01-05	872	54	38
8	10036	2009-01-05	939	62	36
8	10037	2009-01-05	601	66	37
8	10038	2009-01-05	791	52	37
8	10039	2009-01-05	872	68	37

Figure 5-7: Medical Check Fact Snapshot Table Prior to Feature Scaling

SQL File 1		ROUT_MCNormalized		fsmedical_check_fact_table		
Filter:		Edit:		Export:		Autosize:
	idTest	number	dayDate	TestResult	weight	temperature
	8	10032	2009-01-05	-1.30112	-1.50234	-1.30483
	8	10033	2009-01-05	0.722621	-1.01333	-1.30483
	8	10034	2009-01-05	-0.763341	-1.01333	-0.062134
	8	10035	2009-01-05	0.642427	-0.850332	1.18057
	8	10036	2009-01-05	0.958488	0.453683	-1.30483
	8	10037	2009-01-05	-0.635973	1.10569	-0.062134
	8	10038	2009-01-05	0.260322	-1.17634	-0.062134
	8	10039	2009-01-05	0.642427	1.43169	-0.062134

Figure 5-8: Medical Check Fact Snapshot post applying Feature Scaling

The ROUT_MCNormalized is an example of a procedure that is used to achieve the normalization of features in the medical checkup fact table. The routine itself uses data from another stored view that keeps the updated averages and standard deviations of the different measures or features from the fact table. This is then used to normalize the specific features as indicated by the normalization formula presented above.

At each run of the routine the fact table is truncated as there are changes in the values due to changes in the averages and standard deviations. This information is kept in views due to the performance related overheads that would otherwise result from carrying out the normalization at each stage when the analysis of the fact table is being carried out.


```

-----
-- Routine DDL Maintain View for Patient/Clinical Medical Check
-- (ROUT_MCNormalized)
-- Note: Patient Medical Checkup Feature Normalization
-----

DELIMITER $$

CREATE PROCEDURE `ROUT_MCNormalized`(
OUT rc INT
)
BEGIN
    -- Capture the Mean and standard deviations of features from a view;

    SELECT @AvgOfTest:=AvgOfTest, @AvgOfWeight:=AvgOfWeight,
    @AvgOfTemp:=AvgOfTemp, @StdTest:=StdTest,@StdWeight:=StdWeight,
    @StdTemp:=StdTemp,
    @AvgOfPVLoad:= AvgOfPVLoad,
    @StdPVL:= StdPVL,
    @AvgOfviralLoad:= AvgOfviralLoad,
    @StdviralLoad:=StdviralLoad
    FROM `DWH`.`featurescaledmedicalcheck`;

    TRUNCATE TABLE `DWH`.`fsmedical_check_fact_table`;

    INSERT INTO fsmedical_check_fact_table

    SELECT idTest, number, dayDate,((TestResult - AvgOfTest)/StdTest),
    ((weight - AvgOfWeight)/StdWeight),
    ((temperature - AvgOfTemp)/StdTemp)

    FROM DWH.medical_check_fact_table;

    SET rc = 0;

END

```


5.3 Matrix implementation of Linear Regression Algorithm

A regression data mining formula was developed to provide care givers with estimate of patient viral load progression basing on the features being monitored during routine therapy.

$$h_{\theta}(x) = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n \quad \text{Equation 2: Projection formula for HIV patient viral load}$$

Using the above regression equation based on x features of a patient parameterized by theta, the optimization for calculation of the viral load for each subsequent patient was then carried out using the regression algorithm cost formulae in Equation. The optimization iterations (500 iterations) minimize the value of $J(\theta)$ below to establish the theta parameters that best fit the data. These parameters are then used in Equation 2.

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 \quad \text{Equation 3 Cost Formula for finding optimization values}$$

Where $x^{(i)}$ represents the different features in the medical_check_fact_table and $y^{(i)}$ represents the projected viral load of the patient on antiretroviral therapy and m is the number of patients in the data warehouse that we are using train the algorithm.

The above was implemented using the open source tool octave (Eaton 2012) using vectors and matrices. Using the matrix and vector approach the Equation 2 ($h_{\theta}(x)$) could then be represented as the transpose vector θ multiplied by a matrix X of the features in the fact table, as represented below. This avoids for loops in the algorithm and results in more efficient code that executes the complex operation involving may variables in fewer steps with fewer lines of code.

$$X = \begin{bmatrix} x_0 & x_1 & x_2 & \dots & \dots & x_n \end{bmatrix} \quad \text{and} \quad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix} \quad \text{Equation 4: X and } \theta \text{ as Matrices}$$

A matrix multiplication of X and θ can then give us Equation 2.

Training the cost formula (Equation 3) to get parameters of θ the regression formula (Equation 2) that provides the minimal value for ($\min J(\theta)$), over 500 iterations provided a cost graph depicted below (Figure 5-9). This was generated from a *MedicalCheck* matrix exported from the data warehouse of size 732X5. The 732 training examples were from entries of 60 patients measuring CD4 counts and viral loads over a period of 4 years.

$$\text{The vector } \theta \text{ values that gave the minimal values of } J(\theta) \text{ was } \theta = \begin{bmatrix} -1.6773e - 007 \\ -7.3284e - 001 \\ -9.3970e - 002 \\ 6.2127e - 002 \\ -5.0825e - 002 \end{bmatrix}$$

as a 5row column vector (R^5). Figure 5-9 indicates the convergence of theta values to the best fit for the data supplied from data warehouse.

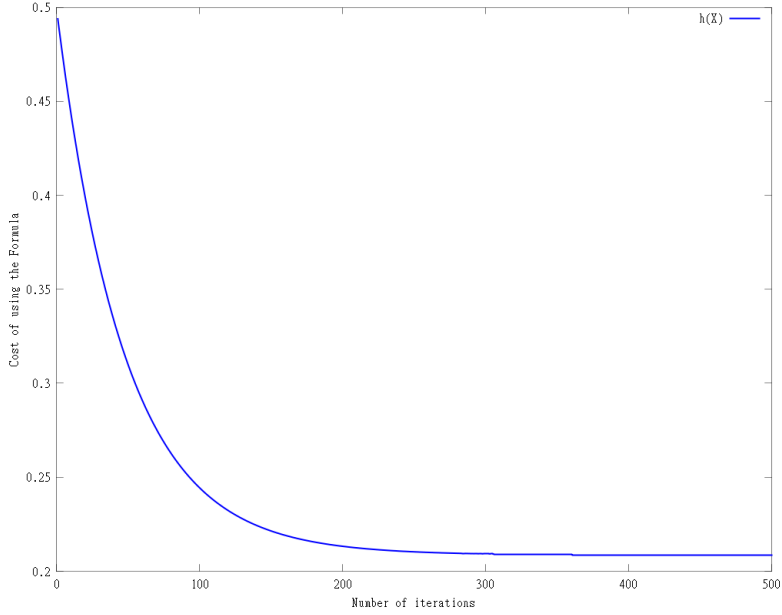


Figure 5-9: Graph of Cost of using Algorithm & Iterations

Basing on the generated column vector R^5 estimates of patient viral load could be done depending on the features provided from the database.

5.4 Supervised Learning for HIV Treatment Failure

A data mining logistic regression model was adopted for treatment failure. This used

the logistic regression formula using the sigmoid function $h_{\theta}(x) = \frac{1}{(1+e^{\theta^T x})}$.

This was used to check the case for whether a patient suffered with treatment failure or not. In summary this cost function below was used forming the classification of two classes for ART treatment failure. In this case $y=1$ when a patient under ART undergoes treatment failure and $y=0$ when a patient does not exhibit this.

$$\text{Cost}(h_{\theta}(x), y) = \begin{cases} -\log(h_{\theta}(x)) & \text{if } y = 1 \\ -\log(1 - h_{\theta}(x)) & \text{if } y = 0 \end{cases} \quad \text{Cost Equation 5: Cost Function for Classification}$$

This was then summarized into the cost equation and the equation for minimization below with regularization of the parameters to avoid the problem of over fitting.

$$J(\theta) = -\frac{1}{m} \left[\sum_{i=1}^m y^i \log h_{\theta}(x^i) + (1 - y^i) \log (1 - h_{\theta}(x^i)) \right] + \frac{\gamma}{m} \sum_{j=2}^n \theta_j^2 \quad \text{Equation 6}$$

During the training of the algorithm developed the objective was to find parameters of theta that minimized the cost function above and defined a decision boundary for

the algorithm. Training the algorithm with a matrix 1560*10 generated after exporting the features from the data warehouse. More features were generated from each data point using feature mapping to ensure better fitting of the data and more interesting polynomial terms.

The column vector θ of size R was established as

$$\theta = \begin{bmatrix} -1.432039 \\ 0.123241 \\ -0.365403 \\ 0.357051 \\ 0.174900 \\ -1.458001 \end{bmatrix}$$

The decision boundary and the classification generated is provided below.

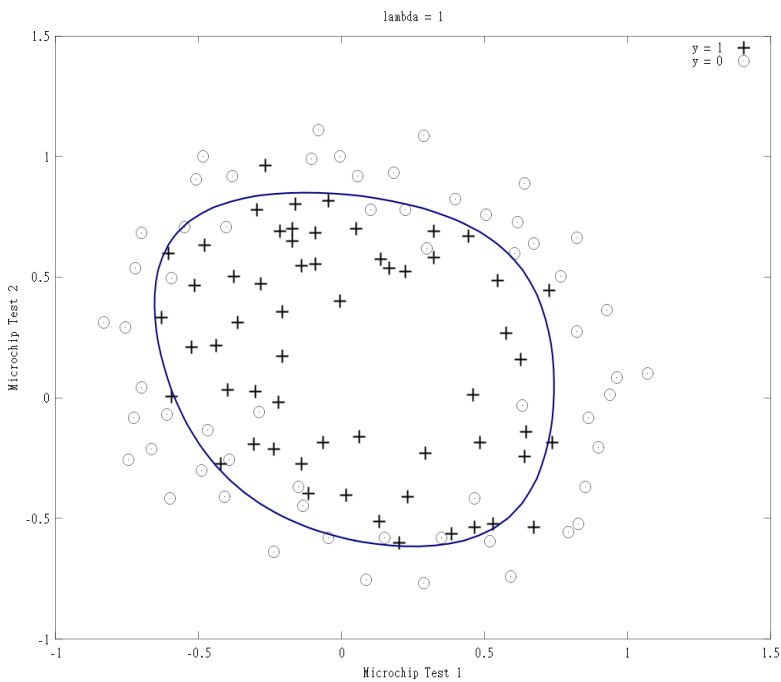


Figure 5-10: Classification for ART treatment failure

This basing on features can then be used to predict the patient treatment failure based on the two different clusters formed above. The ‘+’ indicating those who are progressing normally through treatment and ‘o’ indicating the cases of treatment failure and anomalies in therapy.

The issue of false negatives (a patient is undergoing treatment failure but algorithm classifies them as ‘+’) was considered and managed by making the decision boundary diffuse with a bias on the positive and rejecting ‘+’s that fall on thee boundary. This is

an area also for further research on how to deal with scenarios of cases that fall on the boundary defined.

Part III

In this part of the thesis the results of the study are discussed specifically highlighting meaning of the results, contributionsof the study, the challenges faced and areas for further research.

CHAPTER 6

DISCUSSIONS

In this chapter the meaning of the results identified from the outcome of the study is discussed.

6.1 Summary of Main Finding

- The need for electronic medical records that facilitate deeper analysis and critic of ART was identified. The study revealed the emergence of a plethora of mobile applications used in health, with a focus on advocacy and information sharing. This forms a good foundation for the dissemination of results from knowledge engineering systems to different stakeholders.
- Standardization of data collection, aggregation and formatting is crucial to success of knowledge engineering systems. This is facilitated by collaboration and the involvement of different stakeholders including government, health care providers and researchers.
- The adoption of open source innovative tools in the health care information systems is underway. However the adoption is still at the level of information distribution as opposed to analysis of information. These innovative tools and systems can form sustainable, affordable and effective architectures for expert health information systems to manage ART in the country.
- Star-schema based dimensional models are suitable for building incremental systems such as HIV data warehouses, as the architecture facilitates continuous addition of new dimensions of study and analysis to the model.
- Participatory action research approach through stakeholder involvement in knowledge engineering systems results in high quality systems that are accepted and have a high usability.

- The study on the HIV data warehouse identified and trained 2 potential data mining algorithms. Using vector manipulation of the data warehouse data and machine learning,
 - Linear regression can be applied to optimize the estimation of HIV viral load RNA levels based on captured CD4 count measures with 83% level of accuracy.
 - ART Classification can be done for treatment failure using regularized logistic linear regression. This forecasts treatment failure with confidence level of 83.7%
- Knowledge engineering solutions can be used to support both operational level decisions on ART as well as strategic level decisions for ART needs planning by different categories.

6.2 Conditions for Knowledge Engineering Systems in ART

The full scale adoption of techniques such as data warehouse and the resultant benefits such as improved analysis, machine learning and decision support requires: a gradual adoption of electronic mechanisms of managing patient records to support the technology. Adoption of EMR requires standardization of data collected, formatting and aggregation techniques. This also calls for appropriate IT policies and e-health policies to govern the standardization process and MOH playing an oversight role in ensuring stakeholders conform to these standards.

The ICT infrastructure is under development and already plays and will continue to play important roles in the process of mainstreaming IT health information systems in health care. The mobile sector has penetrated many regions of the country. This can be used to facilitate the dissemination of results of knowledge engineering systems. The lack of knowledge of innovative technologies available is a factor in slow uptake of knowledge engineering in health systems. Knowledge such as effective and affordable open source implementations, encryption and security solutions to ensure privacy and protection of patient information is important and has been tested and shown to work in this research and context. This is forming tested proof that they work and providing a framework of how these innovative tools can be made to work in managing ART.

6.3 Dimensional Model for HIV/AIDS

This presents a starting point for monitoring ART through analysis of therapy information in a data warehouse. The presented data warehouse dimensional model for ART therapy is a first attempt at trying to categorize ART monitoring dimensions in a data warehouse. The process of generating this particular dimensional model adds to the process model proposed by Fayyad et. al(1996) to include data warehousing, machine learning and data mining to provide decision support.

Furthermore this model can be used as tier 1 architecture in building dimensional models that monitor different aspects of ART therapy as the architecture adopted is one that uses the “bus architecture”. New processes monitored during ART of interest to data mining can thus be modeled and joined to model. Relationships with other interesting illnesses can also be modeled. The adoption of an open source implementa-

tion assists to have manageable costs in implementation and affordability for resource constrained settings with resources available for software licenses.

6.4 Viral Load from CD4 Count History

The results provide an expert tool to estimate patient RNA viral load from testing of patient CD4 count values over a period, facilitates monitoring of therapy since RNA viral load indicates the therapy direction. Increments in viral load can be interpreted to mean, treatment is not working optimally or the patient is not adhering to treatment. This points the health care provider to intervene and counsel the patient basing on what the predictions indicate. The algorithm based on continues updates and recalculations of optimal values of parameters θ for the equation ensures that the system gets even more precise and optimal values with increased usage and loading of new data for the data warehouse to learn from.

The prediction also moves more resources away the expensive viral load testing to the only the necessary cases while focusing the manageable CD4 count testing to all routine and normally progressing clients. This improves the time dedicated by a given human resource in providing routine monitoring on a patient, including the time to analyze past tests results and actually carry out an actual viral load test versus estimation and prediction of the viral load, based on CD4 count history.

6.5 Patient Treatment Failure

The treatment failure classification is important as it enables the care givers to predict the treatment failure based on characteristics that are displayed by the patient. Treatment failure means the change of ART regimen, and should be addressed as soon as identified. Complete consensus and guidelines still lacks in clearly defining the point of treatment failure. Three 3 domains of clinical, immunological (based on CD4 count) and virological failures (based on viral load) being expressed with the differences in context of the research not allowing for one clear cut guideline to address all.

The regularized regression algorithm used ensures that we the RNA level of $>10,000$ copies/ml and CD4 count of <100 recommended by WHO are factored in defining the treatment failure class. The expert tool presented therefore adapts the three views during classification for treatment failure within the context of Uganda.

6.6 User interaction and participation in fostering learning, growth and sustainability

The task shifting policy has been put in place to help address the limitations in human resource needs in different health care disciplines including ART. The participatory approach to the research helps to complement the policy of task shifting by fostering learning, growth and sustainability between and amongst all stakeholders. Aside from stakeholders learning from each other it is also important to ensure sustainability by

ensuring that both new medical processes and dimensions can be added to the dimensional model without the need to overhaul the basic underlying system architecture. This guarantees the growth and development of the data warehouse as new users learn from others and through additions of interesting dimensions to the model can obtain even more interesting models linking new medical processes.

The wiki platform and the dimensional manager toolkit on the wiki, ensures these two issues. The issue of learning and capacity development for the stakeholders and the growth of the data warehouse in terms of addition of new interesting dimensions. Avoiding white elephant scenarios that plague knowledge engineering systems such as data warehouses and ensuring high end products that have high usability.

6.7 Decision Support at Policy and Operational Level

At the policy level the decision support is important in defining good strategic direction for policies. In ART the supply chain has been identified as a problem area for ART (Bamuturaki 2008; Windisch et al. 2011), this has a direct influence on the quality of therapy that health centers and care givers are able to provide. The reporting and summarization functions for the policy makers in the system present the opportunity for proper planning for needs of ART based on the number of people on treatment and their requirements. Addressing ART therapy needs such as planning for the needs by region and health center and dealing with the current problems of stock-outs in the centers the system offers a good basis for decision support to stakeholders at policy level.

6.8 Contributions

6.8.1 To Science and Academia

This research and study has:

- Identified data mining techniques that can be leveraged for HIV/AIDS patient monitoring in Uganda. These are regression and classification through regularized logistic regression.
- Developed a matrix implementation of linear regression and regularized linear for viral load projection from CD4 count test.
- Identified the need for electronic data medical records to support the data needs for knowledge engineering systems.
- Adapted knowledge discovery process by Fayyad (1996) to include data warehousing, decision making and knowledge generation.
- Developed a dimensional model for an HIV/AIDS data warehouse focusing on patient monitoring through Medical checkup and ARV prescription.
- Reviewed open source dimensional modeling tools for data warehousing and developed a framework for their use in dimensional modeling for knowledge engineering purpose in ART.

- Developed open source HIV/AIDS data warehouse architecture for ART monitoring. Providing a common framework for data warehouse structure across different ART providers, based on MOH and WHO accepted guidelines and policies.
- ART adherence monitoring flow chart.
- Online PAR. Fostering intelligent product manual development
- PAR adoptions to Data warehouse Knowledge engineering systems: Better system user requirements through user involvement.

The licentiate thesis (Otiye 2011) titled *“Participatory Approach to Data Warehousing in Health Care: Uganda’s Perspective”* was published in 2011 under the ISBN 978-91-7295-204-1.

Six scientific papers were prepared and presented to various conferences and journals. These papers helped to increase the institutions visibility in the academia.

6.8.2 To Policies and Strategies

The research and study is linked to a number of policies and strategies for development as are highlighted below:

- a) MDG Goal 6: The study contributed to the MDG goal 6 of combating HIV/AIDS, malaria and other diseases. This is by contributing to the goal of providing universal treatment for HIV/AIDS by improving HIV/AIDS information systems and monitoring of therapy at both individual level and strategic level. The framework for linking knowledge engineering practices to monitoring ART directly links to this.
- b) The research is directly linked to the proposed IT policy by government of Uganda. The policy’s mission is the adoption and harnessing of IT for national development and governance through development of a knowledge based economy. This study contributes towards this mission by providing a framework of adopting affordable and available software based tools that can be used to develop knowledge bases for decision support at both strategic and operational levels of ART management.
- c) The research is linked to 3 sections of the (2010/11-2014/15) national development plan (NDP) (GOU 2010c) for Uganda. The first objective linked to the research is the increase of access to quality social services specifically the communicable diseases and HIV/AIDS. The research is in line with the goal of promoting science, technology, innovation and ICT to enhance competitiveness by its result in the framework for the adoption of knowledge engineering in HIV ART management. By proposing a framework that also caters for human resource and learning the research contributes to another NDP objective of enhancing human capital development in ART management.

6.8.3 To ART stakeholders

The core ART care providers in the country have a system that they can use to manage therapy to beneficiaries and a platform that they collaborate and share experiences and learn. At the strategic level stakeholders can project and forecast needs of the ART by the different categorizations provided in the data warehouse.

6.9 Challenges and Weakness

Some challenges encountered during study

Open source versus proprietary challenge

The context of this research guided the selection of the software tools employed towards the open source based softwares due to the cost implications in terms of licenses and maintenance. Open source tools such as MySQL, Postgres, DB professional, WEKA and Octave were used in different areas of the research including data warehouse management, modeling, and data mining. While open source software licenses ensure the access to source codes, some tools have limitations in documentations in general to documentations between different versions and upgrades. This affected selected areas of implementation. For example an initial tool selected for dimensional modeling stopped being supported by higher versions of the data warehouse management system. When dealing with multiple products with multiple contributors then ensuring harmony in the final product is a challenge. These can be ensured by having standards available, clear documentations and forum where stakeholders discuss and learn openly. These are the principles of open source collaboration and are very close to the key foundations of participatory action research.

There are also mysteries and misconceptions that many still have with open source due to its history. Stakeholders interviewed associating it to command based systems that are not user friendly and needing very high technical expertise to use and even some believing they are inferior products compared to proprietary software products. Changing this mindset required evidence based demonstration and presentations with stakeholders to gain their support. This is the work of the researcher in this participatory action research setting to guide the group to gain consensus around such research issues.

Research or Project dilemma challenge

During the research a clear distinction had to be made between the research and a consultancy based project. This can be answered by the works of Baskerville & Pries-heje (2000) and Avison et al. (2001) who discuss the characteristics of participatory based research. Key elements of participatory action research such as the motivation, generation of knowledge through and by participants rather than talk down of researchers experiences coupled with change management provide insights on key differences between the two. The motivation for the research is primarily knowledge generation through solution of a problem that affects the stakeholders.

Weaknesses in adopting supervised machine learning algorithms

The two machine learning algorithms of linear regression and classification by logistic linear regression required using selected data from the data warehouse fact table ART processes as the training set for the algorithm. It is necessary for the algorithm to generalize the information from the training examples in a best fit manner or the most optimal solution. The final algorithm is then tested against some sample data to

establish its accuracy in predicting the feature as expected. The research split the data warehouse fact table into two at a ratio of 70 to 30 with 70% of the columns returned being used to train the algorithm and selected random 30% of columns being used to check the accuracy of the predictions. This has a potential weakness in a supervised learning algorithm as the test sample could potentially exhibit the same characteristics as the training set leading to inaccuracies in the prediction. These could be due to noise in the sample data which could be due to errors in loading data into the database, or capturing test cases from the health facility.

For the case of logistic regression this also the potential of the algorithm overfitting the training set provided and failing to generalize adequately and form a good prediction algorithm.

CHAPTER 7

CONCLUSIONS AND RECOMMENDATIONS

7.1 Introduction

This study developed a framework for adopting knowledge engineering in information systems for monitoring ART in HIV/AIDS patients. The study has; identified the status of information systems in the country, developed open source tools for modeling HIV data in resource constrained settings, developed an ART monitoring dimensional model, and tested two data mining algorithms to estimate viral load from CD4 count tests and thereby classify treatment failure.

The study was driven by a need to addresses the challenge of monitoring ART in a resource constrained setting, with high ratio of patients to care givers and need for clear strategic decision support systems for ARV needs and supply management. ART access and management has had a special focus in the different government health policies, national development plans and global millennium development goals. Given the knowledge engineering solutions proposed, the resource constrained context called for a manageable and sustainable solution in form of software costs, because of this an open source solution was adopted.

The study sort to answer the following research questions:

- i. What is the status of information systems in health care in Uganda?
- ii. What are the constraints in developing an HIV patient data warehouse using open source software?

- iii. What are the tools for modeling HIV data in resource constrained settings?
- iv. How can open source data warehouses be deployed for knowledge discovery in HIV?
- v. What dimensional model and data mining technique in an HIV/AIDS data warehouse can be used for in monitoring patient adherence to therapy?
- vi. What is the importance of user engagement and participation in developing cost effective knowledge engineering systems for HIV/AIDS patient care?

To answer the research questions a participatory action research methodology was adopted in building a data warehousing and data mining system. The general theoretical literature on this subject and specifically in the Ugandan context is inconclusive on open source frameworks for dimensional models that could be used as a basis for monitoring therapy across different ART centers in the country.

7.2 Conclusions

The empirical findings were summarized in respective chapters and sections and will be referenced by the appropriate chapter/heading numbers. This section will synthesize the empirical findings to answer the questions posed by the research, linking the conclusions to the research question.

Status of information system in health care

The chapter/section 2.7 on theoretical analysis and standards presented the status of health information systems in Uganda. The country's health care system is based on a tiered system.

- a) **Strength, weakness and opportunities:** The uptake of IT systems in health care for expert decision support in ART is low but their use in basic systems such as information dissemination is noticeable. The tiered health care structure does already support the information flow bottom up and vice versa. This presents a mechanism of communicating policies and guidelines to the lower levels when required.
- b) **National infrastructure backbone:** There is a high potential for an effective knowledge based system with data fed through the tiered health care system, with effective distribution through the government IT infrastructure backbone.
- c) **Human resource needs:** The Chapter/Section 2.8 presents the human resource needs of the country to manage the numbers in need of care basing on the high doctor patient ratio. This reflecting a health care system that requires mechanisms to manage the high numbers of ART and the shortage of human resource and other resources.

Tools for modeling HIV data in resource constrained settings

The chapter on results and specifically Chapter/Section 4.3 and 4.4 address this. The paper on dimensional modeling of HIV data in open source and knowledge discovery in health care using data mining set address the issues raised by this research. The chapter/ section 5.3 and 5.4 present equations developed through machine learning using the tool octave to forecast patient viral load from CD4 count tests and to map treatment failure. Interaction between different knowledge engineering systems can be

achieved and shared through architectures that ensure centralized aggregation of health data in data warehouses and light protocols such as HTTP or HTTPS.

a). **DB professional and data architect:** These are highlighted as tools that can be used to model HIV data. The development of a dimensional model generated from these around two processes of ART and their being open source is an indication of their relevance in a resource constrained setting.

b). **Matrix implementation in Octave:** This tool accepted data from a data warehouse and is used to find optimal parameters for the two data mining equations for estimation of viral load and ART treatment failure mapping. The tool's use in estimation of new patient viral load basing on learned features is testament to the presence of a modeling tool for HIV in a resource constrained setting.

Constraints to developing an HIV patient data warehouse using open source

The scientific papers contained chapter/sections 4.5, 4.6 and 4.8 present the implementation of an HIV data warehouse in Uganda, the importance of engaging stakeholders in requirements and mechanisms of ensuring growth of established dimensional models in ART. Knowledge engineering systems in open source should factor in staff training needs. Training techniques that involve group practical group learning and sharing techniques such as participatory action research can address the training, growth and maintenance constraints faced by such systems. The constraints due to infrastructure continue to be addressed through improved connectivity country wide. The approach of using the architecture presented in chapter/section 5.2.1 and Figure 5-1 can improve access to the system for the locations covered by the current infrastructure. The webserver approach using basic HTTP and HTTPS enabled devices can be adopted at minimal costs to ensure wide access to the knowledge engineering system.

Deploying Open source data warehouses for Knowledge Discovery in HIV

The Paper III (chapter/section 4.5) presented the implementation of an open source data warehouse including the architecture of the system and the technical challenges in implementation. Open source DBMS systems enable health care organizations to have access to software at manageable license costs. The chapter 5 on additional results (chapter/section 5.2.1) presented overview architecture of the system showing the interaction between the open source data warehouse, user front end the developed data mining algorithms. The webserver can be used to enable interaction between different open source knowledge engineering tools encapsulating the complex interactions from the user. Furthermore this approach opens up the possibility of integration with simple devices such as phones and sms.

User engagement and participation in developing Open Source data warehouse systems

The significance of ensuring quality information systems through involvement of stakeholders is first addressed in Paper IV (chapter/section 4.6). Furthermore ensuring that users collaborate in developing intelligent manuals for systems (Paper V chapter/

section 4.7) and in extending defined dimensional models (Paper VI chapter/section 4.8) indicate the important role played by stakeholders in the process of developing knowledge engineering systems.

- a) **Involve stakeholders in defining requirements:** Knowledge engineering systems such as data warehouse systems are dependent on dimensions from different sections of the domain under study. This is evidenced from the categories of stakeholders who provided input in defining the requirements. Adopting a research methodology that brings all these stakeholders in a harmonious mix has been shown to be beneficial. Participatory action research methodology provides this specific framework.
- b) **Incremental growth of system through stakeholder addition of dimensions:** This ensures that stakeholders together contribute to both the growth of the system and learning the features of the system. Inmon (1996), Kimball & Ross (2002) considered the father of data warehouse systems present them as non-volatile data sources which are fixed with underlying model changes managed and overseen by IT specialists in consultation with stakeholders. In the context of the research given the limited resources moving the stakeholders into limited dimensional model management roles and granting them reasonable control is important. It fosters learning and moves systems into the realm of the intelligent, that of self-governing and self-running.

Dimensional Model and Data Mining techniques for monitoring HIV patients during ART

Paper II presenting dimensional modeling of HIV Data in open source (chapter/section 4.4), the matrix implementation of linear regression algorithm (chapter/section 5.3) and the supervised learning for HIV treatment failure (chapter/section 5.4) provide insight to the last research question.

Dimensional model for HIV in open source: This star schema model presented basing on the ART functions of medical checkup and prescription presents a foundation for an ART management system. The literature on HIV data warehouse models is limited; furthermore the conformed bus architecture adopted by this model ensures growth through addition of new processes. This leaves the door open on the extension capability for the addition of new ART processes that may need to be studied.

Supervised learning with linear and regularized logistic regression: These two algorithms and related optimal parameters provide a framework for monitoring ART. The solution presented is one that constantly learns and recalculates the optimal values with new additions to the data warehouse.

7.3 Recommendations

Recommendations to Policy Makers

- a) The standardization of data capture by the data ART data collection points would shorten the time required before the data can be added to the data warehouse and form training sets for the data mining algorithms. All stakeholders must conform to these standards to data capture.
- b) Implementation of phased role out of electronic medical records (open source versions of EMR are available and test cases can be implemented).

- c) Adopt dimensional model for ART focus on medical check process and prescription, these link to the crucial processes in ART. Interesting processes of the ART can be subsequently added to the model and researched on.

Areas for further research

A potential area for further research is the testing of unsupervised learning algorithms on the data warehouse data. This would potentially discover other clusters of treatment failure and move away from the current 2 cluster decision boundary used in this research. This would involve some 1 versus many algorithms and the use of artificial neural networks.

More detailed study of projections made by the algorithm that fall on the decision boundary needs to be made. This would provide more insights on cases of false negatives and false positives for values that fall on the boundary defined.

Improvement to the architecture in Figure 5-1 can be examined to include the mobile sms platform that has enjoyed success in the country. The mobile approach would further address the infrastructure constraints faced.

Potential Improvements of the Research

The data warehouse supervised learning algorithm validation could be improved by matching data warehouse data validation techniques versus real time patient CD4 count tests and viral load tests. By validating the algorithm using a sections of the fact table results randomly selected and excluded from the training set, the potential inaccuracies and noise inherent in the data warehouse values could have affected the accuracy levels displayed during validation.

REFERENCES

- A.D.A.M, M.E., 2011. Acquired immune deficiency syndrome. Available at: <http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0001620/>.
- Aken, J.E.V., 2001. Mode 2 Knowledge Production in the field of Management Mode 2 Knowledge Production in the field of Management. , (December).
- Avison, D., Baskerville, R. & Myers, M., 2001. Controlling action research projects. *Information Technology & People*, 14(1), pp.28–45.
- Bamuturaki, M., 2008. Uganda edges closer to AIDS treatment for all. *Bulletin of the World Health Organization*, 86(6), pp.423–425.
- Baskerville, R. & Pries-heje, J., 2000. Grounded action research : a method for understanding IT in practice. *Accounting Management and Information Technologies*, 9(1999), pp.1–23.
- Bond, C., 1986. AIDS lays waste to Uganda. *New African*, 226.
- Breslin, M., 2004. Data Warehousing Battle of the Giants: Comparing the Basics of the Kimball and Inmon Models. BI Best Practises. Available at: <http://www.bi-bestpractices.com/view-articles/4768> [Accessed March 1, 2008].
- Cabena, P. et al., 1998. *Discovering Data Mining: From Concept to Implementation* NY., Prentice Hall.
- CapacityPlus, 2012. Uganda Launches National Health Workforce Information System Built on iHRIS. Available at: <http://www.capacityplus.org/Uganda-Launches-National-Health-Workforce-Information-System> [Accessed July 1, 2012].
- Carswell, J. & Lloyd, G., 1987. Rise in prevalence of HIV antibodies recorded at an antenatal booking clinic in Kampala, Uganda. *AIDS* (London, England), 1(3), pp.192–193.
- Catley, C. et al., 2006. Predicting high-risk preterm birth using artificial neural networks. *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society*, 10(3), pp.540–9.
- Chambers, R., 1990. Rapid but relaxed and participatory rural appraisal: Towards applications in health and nutrition. Available at: <http://opendocs.ids.ac.uk/opendocs/bitstream/handle/123456789/98/rc405.pdf?sequence=1>.
- Cios, K.J. & Moore, G.W., 2001. Medical data mining and knowledge discovery: Overview of Key Issues. In *Medical data mining and knowledge discovery*. New York: Physica-Verlag Heidelberg.
- Connolly, T.A., 2002. *Database Systems: A Practical Approach to Design, Implementation and Management*, Addison-Wesley.
- Dick, B., 2002. Action research: action and research. Available at: <http://www.aral.com.au/resources/aandr.html> [Accessed May 12, 2012].
- Eaton, J.W., 2012. About GNU Octave.
- Er, O., Temurtas, F. & Tanrikulu, a. Ç., 2008. Tuberculosis Disease Diagnosis Using Artificial Neural Networks. *Journal of Medical Systems*, 34(3), pp.299–302.
- Ewen, E.F. et al., 1999. Data Warehousing in an Integrated Health System ; Building the Business Case. *DOLAP 98*, pp.47–53.
- Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P., 1996. Knowledge Discovery and Data Minin: Towards a Unifying Framework. In *Proc 2nd Int Confon Knowledge Discovery and Data Mining Portland OR*. pp. 82–88.
- Fayyad, U., Piatetsky-shapiro, G. & Smyth, P., 1996. From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence*, 17(3), pp.37–54.
- GOU, U., 2010a. *Government of Uganda UNGASS Country Progress Report 2010*, Kampala.
- GOU, U., 2010b. *ICT Policy Final Draft Feb 2010*,
- GOU, U., 2010c. *National Development Plan (2010/11 - 2014/15)*, Kampala.

- Gibbons, M. et al., 1994. *The New Production of Knowledge The Dynamics of Science and Research in Contemporary Societies*, SAGE Publications Ltd.
- HISP, 2012. DHS2 (District Health Information System 2). Available at: <http://dhis2.org/node/6>.
- Hand, D., Mannila, H. & Smyth, P., 2001. *Principles of Data Mining*, MA: MIT Press Cambridge.
- HealthChild, U., 2012. *Testimony on text messages*, Youtube.
- Heatwole, A., 2011. Digitizing Uganda's Health Services: UNICEF Uganda's New Mobile Program. Available at: <http://mobileactive.org/blog/ugandahealthmanagement> [Accessed February 1, 2012].
- Henriquez, K., 2009. Text to Change: Spreading the Message to Stop the Virus. ICT 4 Uganda. Available at: <http://ict4uganda.wordpress.com/2009/03/31/text-to-change-spreading-the-message-to-stop-the-virus/> [Accessed April 3, 2011].
- Hippel, E.V., 2005. *Democratization of Innovation.pdf*, MIT Press Cambridge.
- Hladik, W. et al., 2008. The estimated burden of HIV/AIDS in Uganda, 2005-2010. *AIDS* (London, England), 22(4), pp.503–10.
- Huang, M.-J., Chen, M.-Y. & Lee, S.-C., 2007. Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis. *Expert Systems with Applications*, 32(3), pp.856–867.
- Inmon, W., 1996. *Building the Data Warehouse*, John Wiley & Sons Inc.
- Inmon, W.H., 1996. The Data Warehouse and Data Mining. *Communications of the ACM*, 39(11), pp.49–50.
- Jonsdottir, T. et al., 2006. The feasibility of constructing a Predictive Outcome Model for breast cancer using the tools of data mining. *Expert Systems with Applications*, 34(1), pp.108–118.
- Kasozi, E. & Agencies, 2012. Experts approve pill to prevent Aids virus. *Monitor Publications*.
- Keim, D.A., 2002. Information Visualization and Visual Data Mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1), pp.1–8.
- Kimball, R. & Ross, M., 2002. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling* 2nd ed., NY: John Wiley & Sons Inc.
- Kriegel, H.B., 2007. Future trends in data mining. *Data mining and Knowledge Discovery*, 15, pp.87–97.
- Krishnaswamy, A., 2004. Participatory Research: Strategies and Tools 1. , 20(June 2003), pp.17–22.
- Kusiak, A., Dixon, B. & Shah, S., 2005. Predicting survival time for kidney dialysis patients: a data mining approach. *Computers in biology and medicine*, 35(4), pp.311–27.
- Larose, D.T., 2005. *Discovering Knowledge in Data: An Introduction to Data Mining*, Wiley Interscience.
- Livesley, N. et al., 2008. *Private-for-Profit HIV / AIDS Care in Uganda : An Assessment Private-for-Profit HIV / AIDS Care in Uganda : An Assessment*,
- Low-Beer, D., 2002. HIV-1 incidence and prevalence trends in Uganda. *The Lancet*, 360(9347), pp.1788–1789.
- Lundin, J., 1998. Artificial Neural networks in outcome prediction. *Ann Chir Gynaecol*, 97(2), pp.128–130.
- Lutalo, I.M. et al., 2009. Training needs assessment for clinicians at antiretroviral therapy clinics: evidence from a national survey in Uganda. *Human resources for health*, 7, p.76.
- MIT, T.R., 2001. 10 Emerging Technologies That Will Change the World. MIT. Available at: <http://www.technologyreview.com/infotech/12265/> [Accessed January 1, 2007].
- MOH, G.U., 2009. *National Antiretroviral Treatment Guidelines for Adults , Adolescents , and Children*, Kampala.
- Maimon, O. & Rokach, L., 2005. *The Data Mining and Knowledge Discovery Handbook*, Springer.

- McNiff, J., 2002. Action Research for Professional Development, Concise advice for new action researchers. Available at: <http://www.jeanmcniff.com/ar-booklet.asp> [Accessed December 7, 2007].
- Mitchell, T.M., 1998. The Discipline of Machine Learning. , (July).
- Nalugoda, F. et al., 1997. HIV infection in rural households , Rakai. *Health Transition Review*, 7(2), pp.127–140.
- Nowotny, H., Scott, P. & Gibbons, M., 2003. “Mode 2” Revisited: The New Production of Knowledge. In Kluwer Academic Publishers, pp. 179–194.
- Noya, H.M., 2005. Applying Data mining Techniques in the Development of a Diagnostic Questionnaire for GERD. *Dig Disc Si*, pp.1871–1878.
- OS, I., 2009. The Open Source Definition (Annotated). *Open Source Initiative Website*. Available at: <http://www.opensource.org/docs/definition.php> [Accessed May 4, 2012].
- Orozova-Bekkevold, I. et al., 2007. Maternal vaccination and preterm birth: using data mining as a screening tool. *Pharmacy world science PWS*, 29(3), pp.205–212.
- Otine, C., 2011. *Participatory Approach to Data Warehousing in Health Care : Uganda's Perspective*, Blekinge Institute of Technology.
- Otushabire, T., 2011. Uganda: Government to Introduce Information System. *Monitor News Paper*.
- O'Brien, R., 1998. An Overview of the Methodological Approach of Action Research. *Practice*, 2006(2001), pp.1–22.
- Palpanas, T., 2000. Knowledge discovery in data warehouses. *ACM SIGMOD Record*, 29(3), pp.88–100.
- Pence, G.E., 2007. *Medical Ethics: Accounts of the Cases that Shaped and Define Medical Ethics* 5th ed., NY: McGraw-Hill Ryerson.
- Rajagopalan, B. & Isken, M.W., 2001. Exploiting Data Preparation to Enhance Mining and Knowledge Discovery. *IEEE Transactions on Systems, MAN and Cybernetics*, 31(4), pp.460–467.
- Raoufy, M.R. et al., 2011. A novel method for diagnosing cirrhosis in patients with chronic hepatitis B: artificial neural network approach. *Journal of medical systems*, 35(1), pp.121–6.
- Riel, M., 2010. Understanding Action Research. *Research Methods in the Social Sciences*, 17(1), pp.89–96.
- Roe, B. & Doll, H., 1995. Prevalence and incidence. *Journal of Clinical Nursing*, 13(3), p.188.
- Sewankambo, N. et al., 1994. Demographic impact of HIV infection in rural Rakai district, Uganda: results of a population-based cohort study. *AIDS*, 8(12), pp.1707–1713.
- TTC, U., 2012. Text to Change - Projects. *Website*. Available at: <http://www.texttochange.org/project-list> [Accessed May 3, 2012].
- Tan, K. et al., 2003. Evolutionary computing for knowledge discovery in medical diagnosis. *Artificial Intelligence in Medicine*, 27(2), pp.129–154.
- UIA, I.A., 2009. ICT Sector Profile. Available at: http://www.ugandainvest.go.ug/index.php?option=com_k2&view=item&id=107:ict-sector-profile&Itemid=317 [Accessed May 4, 2012].
- UN, U.N., 2010. *The Millenium Development Goals Report 2010*,
- UNAIDS, 2010. *Global Report UNAIDS Report on the Global AIDS Epidemic 2010*,
- Vararuk, A., Petrounias, I. & Kodogiannis, V., 2008. Data mining techniques for HIV/AIDS data management in Thailand. *Journal of Enterprise Information Management*, 21(1), pp.52–70.
- WHO, 2006. *Patient Monitoring Guidelines for HIV Care and Antiretroviral therapy (ART)*,
- Wadsworth, R., 1998. What is Participatory Action Research. Available at: <http://www.scu.edu.au/schools/gcm/ar/ari/p-ywadsworth98.html> [Accessed December 7, 2007].

- Wasan, S.K., Bhatnagar, V. & Kaur, H., 2006. The impact of data mining techniques on medical diagnostics. *Data Science Journal*, 5(October), pp.119–126.
- Weir, R., Peng, T. & Kerridge, J., 2003. Best Practice for Implementing a Data Warehouse : A Review for Strategic Alignment. In *Proc of the 5th International Workshop on Design and Management of Data Warehouses*. pp. 5–14.
- Windisch, R. et al., 2011. Scaling up antiretroviral therapy in Uganda: using supply chain management to appraise health systems strengthening. *Globalization and health*, 7(1), p.25.
- Wright, P., 1998. Knowledge discovery in databases: Tools and Techniques. Available at: [http://www.mariapinto.es/ciberabstracts/Articulos/Knowledge Discovery.htm](http://www.mariapinto.es/ciberabstracts/Articulos/Knowledge%20Discovery.htm) [Accessed May 13, 2012].
- Xiang, G. & Min, W., 2010. Applying data cleaning in Changqing Oilfield Company ' s data warehouse. In *Second IITA International Conference on Geoscience and Remote Sensing*. pp. 605–607.