Addis Ababa University

Faculty of Informatics

Application of Bayesian Networks (BN) Technology in Predicting Major Factors Behind Poor ART Adherence Trends in Ethiopia - *the case of SNNPR region*

By: Amanuel Biru

March 2009

In Partial Fulfillment *for the Requirements for the* Degree *of Master of Science* in Information Science

DEDICATION

To my Brother,

Dr Estifanos Biru,

With much Love and Affection.

Without you, I would not be here.

ACKNOWLEDGEMENT

This research draws from the efforts of many people around me. Compiling the research work has been tremendous undertaking and there are numerous people whom I want to heartily thank for their hard work, enthusiasm, and dedication.

Dr Rahel Bekele, my Research Advisor, thank you for your insight, your intellect, and your continuous support and belief on my abilities- I could not simply succeed without you! Thank you again for always taking extra time to ask those all-important questions that everyone else seemed to miss. I was fortunate to have you as an advisor.

I am immensely indebted to my beloved parents for giving me unconditional care, love and support throughout my life. Thank you **Aychie**! Thank you **Abaye**!

The encouragement, affection, advice and support I got from my lovely brothers **Estifo, Teshie, Mulat and Dan,** and my sister **Taggie** is immeasurable on any scale and imprinted deep in my heart. Dear brothers and sister, you are true models of sacrificial life. May God bless your days!

I am very much indebted to **Sr. Addisalem** and **Tsehay** who replaced the role of true mother during long stay with them. Your patience, love, encouragement, and unconditional care were critical to attempt this research. Thank you very much!

I am also grateful to my classmates for their dedication and efforts in developing team spirit throughout the course of the study. Special thanks will go to **Behailu**, **Alemayehu** and **Mohammed** with this regard.

Last, but certainly not least, I offer my sincerest gratitude and deepest appreciation to ART Center staff-members at Hossana and Hawassa Hospitals whose cooperation was instrumental in compiling this research. My special thanks extend to Abeba, Rahel and Meskerem (Hossana Hospital ART Center) and Alemayehu (Hawassa Referral Hospital ART Center).

Thank You Jesus! You are the Author and the Finisher of My Life!

TABLE OF CONTENTS

CHAPTER FOUR

CHAPTER FIVE

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ABSTRACT

This research presented the concepts of knowledge breakthrough based on the Bayesian Networks technology to extract valid models of knowledge. The application domain of the research is the health sector, one of the potential sectors to apply Bayesian Network Technology. The research generally aimed at investigating the potential applicability of Bayesian Network technology in developing a model that can support the prediction of ART clients' adherence tends in Ethiopia. The research was conducted in selected hospitals of SNNP Regional Health Bureau.

The methodology used to conduct this research consists of four major steps: Data Source Identification, Data Collection, Data Preprocessing and Model Building/Testing. A total of 1561 records, having 15 attributes each, were used for building, training and testing a Bayesian Network model.

The Bayesian Network learning process was done for complete data (the data for which the training data containing no missing values) which again applied constraint-based approach. This approach performs conditional independence tests on the data. Then it will search for a network that is consistent with the observed dependencies and independencies (applying d-separation concept). Conditional independence relationships among the attributes can serve as **constraints** to construct a Bayesian Network.

The belief network modelling software employed for the purpose of training and testing the BN model was the **Belief Network PowerSoft,** which applies a constraint-based

approach. Three-Phase-Dependency-Analysis (TPDA) in **BN PowerConstructor** was employed in developing the model. The **BN PowerPredictor**, on the other hand, was used to evaluate the prediction accuracy of the model. **BN PreProcessor** was also there to preprocess the data so as to make it ready for model building purpose.

Model testing was implemented in two phases- the first phase without involvement of expert knowledge (i.e., without node ordering, in which, the algorithm learns both structure and parameters), and the second phase by eliciting domain expert knowledge (i.e., involving node ordering, in which, the algorithm learns only the parameters). For both cases, to ensure consistency across the data during the selection of the test and training sets, experiments were carried out by splitting the data into 10 partitions, i.e., a percentage split (10-fold) was used to partition the dataset into training and test data. Each partition, in turn, was used for testing while the remainder was used for training. This process was repeated ten times for the learning algorithm and, at the end, every instance was used exactly once for testing. Finally, the average result of the 10-fold cross validation was considered.

Accordingly, the average predictive accuracy for the model without an **expert** intervention (Experiment I) was **72.80%** at 95% confidence level. According to this model, the adherence (to the medication) of an ART client is **directly** affected by two factors: **Addiction** (drug or alcoholic behaviour) and **loss of job due to ill-health**.

Next, TPDA algorithm (Experiment-II) was implemented by allowing elicitation of expert knowledge. Accordingly, the predictive accuracy of the modified model was **75.9%** which is a better result.

Significant enhancements in prediction and reduction in error rates in the modified model was taken as the indication of the **significance of a domain experts' intervention** during model building.

According to the later model (Experiment II), adherence of an ART client is **directly** affected by six factors: **Addiction** (drug or alcoholic behaviour), **loss of job due to ill-health**, **Residence of a patient**, **Knowledge the client has concerning HIV**, **Employment status of the client**, and **Family dependence (independence).**

From the model developed, it was observed that Bayesian Network is a powerful predictor even in the absence of a domain expert. With a proper intervention of domain experts, it was noticed to perform even better.

ACRONYMS

AIDS	**Acquired Immuno-Deficiency Syndrome**
ART	**Antiretroviral Therapy**
ARV	**Antiretroviral**
BN	**Bayesian Network**
BNJ	**Bayesian Network tool using Java**
CI	**Conditional Independence**
CSA	**Central Statistics Authority**
HAART	**Highly active ART**
HAPCO	**HIV/AIDS Prevention and Control Office**
HIV	**Human Immunity virus**
MoH	**Ministry of Health**
M & E	**Monitoring and Evaluation**
OIs	**Opportunistic Infections**
PI	**Protease Inhibitor**
PLWHA	**People Living with HIV/AIDS**
PMTCT	**Prevention of Mother-to-Child Transmission**
SNNP	**Southern Nations Nationalities and Peoples**
WHO	**World Health organization**
VCT	**Voluntary Counseling and Testing**

CHAPTER ONE

INTRODUCTION

1.1 Background

WIC (2008) revealed that Ethiopia, the second most populous and one of the seriously affected countries in sub-Saharan Africa, has more than 1.3 million people living with HIV, of which more than 250,000 AIDS patients need Antiretroviral Therapy (ART). The Government of Ethiopia has taken measures to reduce the risk of HIV transmission and mitigate the impact of the epidemic on society. One such measure has been the introduction of ART program in 2003 with the goal of reducing HIV-related morbidity and mortality, improving the quality of life of people living with HIV and mitigating some of the impact of the epidemic. In 2005, Ethiopia launched free ART; thereby, over 71,000 were initiated on ART by the end of November 2006 (Federal HAPCO, 2007a). According to Federal HAPCO (2008), the number of PLWHA ever started on ART has increased sharply from 8,276 in 2005 to 24,236 in 2006, to 97,299 in 2007 and to 122,243 at the end of December 2007. At present, Highly Active ART (HAART) has become the breakthrough in treatment of people living with HIV, leading to a reduction in mortality and an improvement in the quality of life. Antiretroviral drugs significantly lowered the rate of HIV transmission from mother to child, and ART has become an integral part of the continuum of HIV care.

Since the free launch of ARV treatment in 2005, several AIDS patients have got access to it and benefited out of it. The government has also launched a national roadmap to cascade the program to individual 'Kebele' levels. Such measures have significantly minimized the accessibility problem which was the most-said issue before few years (MOH Roadmap, 2008). Having somewhat solved the problem of accessibility of the drugs, peoples' perception about HIV has also been changed. Today, with the advent and availability of ARV drugs, AIDS has become a chronic but manageable disease.

However, the era of ART has faced another big challenge- **Poor ART adherence** [(A)] trends of AIDS patients. Even though ART is the single most dramatic development yet in the treatment of HIV (Amico et al, 2004), many have been described as being inconsistent with their treatment regimens, either not taking prescribed medication, taking medications only when they felt up to it, or needing breaks. ART adherence is now widely recognized as critical health promotion behaviour for HIV positive individuals on therapy and it is the "Achilles heel" of successful outcome (Rabkin et al 1999). Adherence to HIV treatment regimen is defined as taking pills in all the prescribed doses at the right time, in the right doses and in the right way[B]. To this end, ART adherence issue has happened to be one of the most critical concerns pertaining to the ART.

According to World Health Organization, adherence is the second strongest predictor of progression to AIDS and death, after CD4 count[C]. Consistently high levels of adherence are also important determinants of virologic and immunologic outcome, AIDS-related morbidity, mortality, and hospitalizations (WHO, 2005a). This makes the ARV treatment one of the most sensitive and challenging treatments in the era of medicine.

[A]*ESAMI (2008)* defines the term adherence as the extent to which a client's behaviour coincides with the prescribed health care regimen as agreed upon through a shared decision making process between the client and the health care provider.

[B] Carter, M (2005). Adherence Information Series for HIV Positive people, 3rd edition, London
[C] CD4 count is ...

Although the minimum threshold of adherence necessary for the clinical effectiveness of HAART remains unclear, available data suggests that patients must take a high proportion (95% or more) of antiretroviral drug doses to maintain suppression of viral replication, that failure rates increase as adherence levels decrease (WHO, 2005a). According to Jerene (2007), for most chronic diseases other than HIV/AIDS, an adherence of 50-75% is sufficient. On the other hand, since treatment outcome (measured virologically) is sensitive to slight changes in adherence, about 95% adherence is recommended for antiretroviral therapy for effective suppression of viral load. WHO (2005b) also underlines that a patient taking less than 95% of his/her doses is at risk for developing viral resistance and ultimately virological failure which may lead to an exposure to Opportunistic Infections (OIs) that may easily endanger the patient's life.

Annual HIV/AIDS Monitoring and Evaluation Report of Federal HAPCO (2007a), however**,** disclosed that currently ART dropout rate ("worst case of poor adherence") of the country has reached to 21.5%. In other words, around 43 out of 200 people taking the drugs likely dropout the medication (these people are a group of people experiencing a "zero percent adherence"). According to the report, in some of the regions of the country, the figure is even much higher than the average**.** One can realize that the aforementioned figure is extremely terrible as the situation has a national implication.

By its very nature, poor ART adherence issue is not simply an issue of an individual who failed to adhere to it. Rather, it is a multifaceted national concern (WIC, 2008). First, as poor adherence can result in ART failure to the individual, HIV-related morbidity and mortality will increase; HIV-related stigma and discrimination will also be aggravated. That will again bring about orphanism of children, and socio-economic and psychological crisis. Second, a poorly adhered AIDS patient can transmit a drug-resistant virus to another person (say, his/her spouse or casual sex partner) via unsafe sex. This may cause ART resistance in another party who has even not started the regimen because of viral exchange between the two. The problem can touch the whole society through such chains. Unless controlling mechanisms are devised, it may end up with **national ART failure**, which is a fear of the ART experts (ESAMI, 2008).

To this end, ART has become a complex health care intervention with a number of challenges. In order to benefit maximally nearly 100% adherence is necessary, therefore, the support of family, friends, health care workers and other care providers is mandatory for the patient to achieve this level of adherence. Failure to comply with treatment results in complications that are usually difficult to manage especially in resource-limited settings. That is why it is said, "The first regimen is the best regimen" (Federal HAPCO, 2007a).

Adherence to HAART may be affected by the same reasons that are associated with adherence to other medications (yet, magnitudes of the consequences differ). According to USAID (2008), these factors could be related to the patient (such as poor physical/mental health, lack of knowledge/having bad attitudes, alcohol/drug dependency, and poor living arrangements), the medication (such as pill burden, side effects and dietary requirement) or health care (such as distance from health facility, lack of follow-up arrangements, etc.). Interventions to improve adherence should take all these reasons into account. Such interventions could be patient-focused (education, reminders, rewards and reinforcement), provider-focused (continuing medical education, cues and instruments, etc.), regimen-focused (decreasing frequency of dosing and pill burden, reduced cost), or the combination of these (Machtinger & Bangsberg, 2006).

At the initial stages of the ART, researches were focused on the biomedical behavior of the virus and the immune response of the human body. Meanwhile, HIV-related morbidity and mortality was increased at an alarming rate before it was eventually realized that socio-economic and cultural aspects influence human behavior and put individuals at high risk of infection (WIC, 2008). Currently, domain experts are convinced that non-clinical factors like sociodemographic, social and behavioral factors can affect a client's adherence conditions drastically (Federal HAPCO, 2007a).

Castro (2005) believes that most previous studies concerning ART adherence have focused on the clinical aspects, which are easier to measure quantitatively but which do not account for the larger social context. Although clinical epidemiological studies are essential to finding associations between drug regimens and adherence—and, depending on the method chosen, to establishing causality—a biosocial approach that combines quantitative and qualitative methodologies is necessary to bridge the current gap in knowledge on adherence to ART. Only by understanding the complicated interplay between the clinical and social factors that affect adherence to ART can we hope to overcome the real causes of non-adherence.

For countries like Ethiopia that take the lion share of the HIV endemic, the introduction of ART has a significant role. Success on ART however can be guaranteed not only by making it available to the people in need of it but also by continuous struggle to make the clients faithful for the prescribed medications. That is why ART experts say, "Adherence is the heart of HAART" (ESAMI, 2008). To this end, this research attempts to predict the major sociodemographic, social and behavioral factors affecting ART adherence applying current computer-based technology. The intention here is putting efforts in integrating the "most neglected" social dimensions of ART clients to the "most said" clinical aspects of them, thereby coming to all-round solutions in relation to adherence- for resource-limited settings.

1.2 Statement of the Problem

Despite its horrible consequences, a number of people have poor ART adherence records. Moreover, several people completely dropout the treatment for some (unproven) reasons (Federal HAPCO, 2007b). To stop such a catastrophic trend, the root causes behind such terrible decisions of the ART clients should be traced. Different stakeholders (including the domain experts) have different assumptions and estimations concerning the reasons behind the current poor adherence trends on ARV treatment in the country. The differences in assumptions might have come from considering the very diverse background of the ART clients; therefore it has become difficult to generalize the root causes unless a sound scientific procedure is adopted. As such, no strong solutions have been devised so far to address the problem.

Annual HIV/AIDS Monitoring and evaluation Report of Federal HAPCO (2007a) also revealed the situation as follows: "*Around a fifth of the patients who initiated ART are lost-to-follow-up …. It is imperative to conduct an operational research to understand "lost-to-follow-up" and devise workable mechanisms to address the problem."*

Federal HAPCO (2007b) on the other hand has reported that Socio-economic and Behavioral factors may hinder an individual's access to ART services and proper use of the accessed services. According to the report, behavior is a product of intricate interplay between genetic and environmental factors. It is shaped by pervasive cultural, religious, social and economic situations. The educational background, the gender, the marital status, the belief, and socio-economic status of the individual are able to determine his/her behavior. Furthermore, adverse behaviors such as drug or alcohol abuse, sexual behavior, etc may determine decision on getting the access to ART services and clinical/drug adherence. Adverse effects of the regimen also can take a part on the behavior of the person. Persons taking ART may see less need to adopt or maintain safe behavior because they may mistakenly believe the medicine is curing them or rendering them non-infectious. Therefore, patients that are going to be put on ART must be adequately assessed in terms of their real life situations and behaviors that act as barriers to adherence in order to provide the necessary support (Federal HAPCO, 2007b). A Guideline for Management of Opportunistic Infections and ART of Federal HAPCO (2007c) no its part put ART clients with poor educational background, family and social problems, alcohol or drug abuse, and ill-health conditions as the ones less likely adhere.

Studies report conflicting evidence about the association between socio-demographic factors and adherence behavior. Some literatures reported that certain socio-demographic variables have influence over adherence to HAART; however, others showed no association (Tadios & Davey, 2006).

Different scholars make different factors responsible for poor ART trends in different contexts. Escobar et al (2003), a research group in Madrid-Spain, reported that low Level of education, Unemployment, Emotional situation, and Abuse of substances including intravenous drugs can highly affect ART adherence of a patient. The group also reported that Age and Anxiety also have significant contributions on ART adherence of a patient. The research recommends special intervention to reinforce adherence for younger patients, patients taking a high number of antiretroviral drugs, those who have a history of intravenous drug use, and those with high anxiety status.

As of the findings of Melchior & Nemes et al (2008), social and cultural factors are more difficult to be overcome in order to achieve treatment adherence than those related to taking medication. According to the research, these dimensions must be faced not only in the health sector, but also on social and political levels. The content analysis of the research classified the difficulties as follows: related to social factors and life styles, including the stigma; related to beliefs about the use of medication; and directly related to the use of medication. All the interviewees of the target group reported having difficulties concerning the stigma of living with HIV/AIDS. The difficulties related to the use of medication were the most important among patients with the best adherence level. Patients with average adherence level presented all three types of difficulties.

According to Alemayehu et al (2008), adherence is common in those patients who have a social support. Patients who are not depressed are two times more likely to be adherent than those who are depressed. However, at the follow up visit, social support and the use of memory aids were found to be independent predictors of adherence. The principal reasons reported for skipping doses in this study were simply forgetting, feeling sick or ill, being busy and running out of medication in more than 75% of the cases.

Another research team in Jimma University, Kebede et al (2008), found that the reasons given for poor adherence were loss of hope in medication, lack of food, mental illness, holy water (religious belief), lack of money for transport, and other illnesses. Taking hard drugs (cocaine, cannabis and IV drugs), excessive alcohol consumption, being bedridden, living outside the hosting town and having an HIV negative or unknown HIV status partner were associated with defaulting ART.

Stewart et al (1999), on its part, argues that unlike some prior studies, demographic variables (Age, Sex, Marital status, Educational Background, etc), quality of life (Employment status, Income, Infrastructure, etc), and treatment variables other than pills per day did not predict adherence. The investigators did find, however, that forgetting doses and embarrassment at taking antiretroviral were significant barriers to adherence.

Various studies including that of Adam et al (2005) have documented that inadequate knowledge and negative beliefs about HIV disease and treatment effectiveness present an important barrier to ART adherence.

Another research tells us that the more complicated the treatment regimen, the more likely it is for one or more doses to be missed (Mehta et al, 1997). The result made forgetfulness a major cause for missed doses, a problem that could be potentially exacerbated in patients whose memory is affected by AIDS-relative cognitive disorder (Corless et al, 2000). However, Gao *et al (2000)* showed that regimen complexity alone was not a significant predictor of patients' medication adherence.

Zambian research team, James, C. & Aniset, *K.* et al (2008), have reported that travel-related factors did not predict adherence. Adherence was higher for those on ART for a longer time in spite of the fact that they were coming from longer Zambian villages. According to the researchers, patients in rural Zambia can achieve adherence rates compatible with good clinical outcomes despite long travel distances. This finding contradicts the prediction of Jimma University researchers.

One can realize from the discussions above that there are lots of ambiguities and controversial views concerning poor ART adherence trends in different contexts. To this end, no consistent associations were found between different factors and adherence behavior. Results obtained from different researches also tell us the existence of significant gaps on where the major problems lie. Different researchers working on similar suspected factors were seen to arrive at different results; some exaggerate one factor; some completely deny its contribution for adherence. As such, the problem remains unsolved.

While the need for proper interventions of all stakeholders in current ART adherence problems is what many people believe in, the intervention of course should be based on understanding of the underlying causes of the problem. Since HIV patients come from diverse backgrounds, it is difficult to single out the exact reason behind their failure to adhere to the medication. Human experts have lots of pitfalls in dealing with such uncertain knowledge. Subjective knowledge, when taken alone, has its own limitations and consequently not effective in decision making. Simple statistical tools may not serve in addressing such problems. The modern trend is to integrate such knowledge to the current technologies to come to better results. One such emerging technology to manage uncertain knowledge is Bayesian Network Model which is proposed and of course applied in this particular research.

Bayesian networks are useful for both inferential exploration of previously undetermined relationships among variables as well as descriptions of these relationships upon discovery. The benefit of such a process is evident in the ability to describe the discovered network in the future. Moreover, Bayesian networks allow the investigation process to combine domain knowledge with statistical data. Considering these advantages, this research attempts to apply Bayesian network technology to predict the major socio-demographic and behavioral factors aggravating the current poor ART adherence trends in Ethiopia. (Bayesian Network Model, a graphical model which is based on probabilities, is discussed in detail in Chapter 3 of this research).

As the world gears toward increasing access to antiviral treatment in the developing world, it is critical to understand factors (motivators and barriers) that influence adherence to Antiretrovirals and apply the lessons learnt in improving existing and new programs. Available research in Ethiopia has shown that current understanding of factors associated with ART adherence is limited, and related literature in the study area is remarkably scarce. Understanding the predictors of adherence in the local context is a forefront agenda in Ethiopia, where little is known and scaling up of ART program is in progress. In view of this, the prospective study was conducted in SNNP Region to determine the socio-demographic, social and behavioral factors associated with ART adherence. It is anticipated that the findings generated from this study will contribute to the knowledge and understanding of non-adherence to ARVs and be useful in developing evidence based interventions that are undertaken to address ARV adherence in Ethiopia. The following section gives us guiding questions for the research.

1.3 Research Questions

The major questions that guided this research work were the flowing:

I. What are the major sociodemographic, social and behavioral factors/ characteristics of an ART client that can contribute for his/her poor adherence?
II. What are the causal relationships among the attributes?
III. How can one build the Bayesian Network based on the attributes?
IV. How can one evaluate the extent of predictive accuracy of the software tool?

1.4 Objective

1.4.1 General Objective

The general objective of this research is to investigate the potential applicability of Bayesian Network technology in developing a model that can support the prediction of ART clients' adherence tends in Ethiopia.

1.4.2 Specific Objectives

- To identify and select Bayesian network modeling software(s) that best fit for the specific application domain.
- To apply standard preprocessing techniques on data**:** Important steps in data preprocessing include cleaning, transformation and reduction
- To Design, Build, Train and Test Bayesian Network Models.
- To evaluate the results predicted by the software tool(s)
- To analyze the outcome of the research
- To forward critical recommendations depending on the findings.
- To generate reports on what was observed

1.5 Methodology Adopted

To achieve the aforementioned objectives, the researcher adopted the following methodologies:

1.5.1 Source Identification and Data Collection

The data for this purpose was collected from the ART service centers of two hospitals (Hawassa and Hossana hospitals) of SNNP Regional Health Bureau. After intentional selection of health facilities, sufficient number of patients' data in the selected facilities was taken. The data of interest were the most historical ones. The data were manual (paper) formats. The researcher must have gone through each individual record to find and integrate relevant features from the data. Detail discussions on data collection phase are made in Chapter 4 of this thesis.

1.5.2 Data Preparing

The tasks performed in this phase include **data entry**, data **cleaning** (includes filling missing values, smoothening noisy data, identifying or removing outliers and removing inconsistencies), **data transformation** (includes data generalization, and new attributes construction), strategies for **data reduction** (include dimension reduction, and data discretization). Features implemented in this research were discussed in detail in Chapter 4 of the thesis.

1.5.3 Build and Train/Test the Model

This research employed one of the best known learning algorithms for Bayesian Network, called Three-Phase-Dependency-Analysis **(TPDA)** for model building and training/testing**.** The algorithm uses mutual information calculation as the quantitative Conditional Independence (CI) test to avoid the exponential complexity on CI tests. This algorithm takes a database as input and constructs the belief network structure as output. The algorithm also assumes that the database attributes have discrete values and there are no missing values in all the records.

The platform used to design, develop and train the model was **Belief Network PowerSoft**. The package in Belief Network PowerSoft includes three applications: **BN**

PowerConstructor (bnpc.exe), BN PowerPredictor (bnpp.exe) and Data PreProcessor (prep.exe) each having different missions. TPDA learning algorithm and Belief Network PowerSoft are discussed in Chapter 4 of this thesis.

1.5.4 Evaluating the Model

Using BN PowerPredictor of BN PowerSoft, the model was validated for its accuracy with and without elicitation of expert knowledge. Cross-validation and confusion matrix were the mechanisms used to evaluate the model's predictive performance. Other mechanisms used to evaluate the model were instant classifier of BN PowerPredictor and CPT tables generated by BNJ. Chapter 4 details the process.

1.6 Scope and Limitations

The scope of this research is to appraise the potential applicability of Bayesian Network technology in supporting the ART adherence improvement efforts in the SNNP Region. The findings of the research work can be (fairly) considered as relevant in the national level at large.

A misused "Confidentiality" issue of the ART data made its accessibility unattainable. The ethical clearance issue at each segment has hampered the progress of the research significantly. Moreover, the outcome of the research is restricted to the data that was collected to conduct the research.

1.7 Data Protection and Privacy Issues

Prior to the conduct of all tasks, attempts were made to address the data protection and privacy issues by explaining the main objectives of the study to the SNNPG Regional Health

Bureau, Health Facilities' administrators, staff members and assistants. The researcher was also introduced as an MSc student with an official support letter from the Dean of the Faculty of Informatics, Addis Ababa University. Because of the sensitive nature of the data, all necessary care was taken to protect the personal data during its collection.

1.8 Organization of the Thesis

This thesis is organized into five chapters. The **first** chapter presents introduction of the research. The **second** chapter discusses ART adherence issues: challenges and opportunities in resource-limited countries' context- to strengthen the background knowledge of the research domain. The **third** chapter deals with literature review on Bayesian Network Technology and its relevance in the research work at hand. Some prominent researches were also presented with an intention of appreciation of the technology. The **fourth** chapter discusses the experimentation part in detail. The **final** chapter is there for conclusion and recommendation purpose. The recommendations were forwarded based on the findings of the research.

CHAPTER TWO

ART Adherence: Challenges and Opportunities

Millions of people have died of HIV during the last 25 years. The highest number of deaths occurred in poor African countries where antiretroviral therapy (ART) was introduced only recently (**Jerene, 2007**). Worst, the implementation of the drugs still remains challenging. Having many others, ART adherence challenge has become the most terrible one. Many people view ART adherence problems from clinical angles only. But ART adherence has a lot to do with multifaceted sociodemographic, economical, and behavioural factors, especially in resource-limited countries. This Chapter discusses the ART adherence challenges and opportunities in the context of resource-poor countries, including Ethiopia.

2.1 Importance of Antiretroviral Therapy

2.1.1 Biological Importance of ART (HAART)

The survival of people diagnosed with HIV/AIDS dramatically improves with access to highly active antiretroviral therapy (HAART). Such therapy employs a combination of antiretroviral agents—protease inhibitors (PIs), nucleoside reverse transcriptase inhibitors, non-nucleoside reverse transcriptase inhibitors, nucleotide reverse transcriptase inhibitors, and fusion inhibitors—to suppress viral replication, and, thus, reduces the likelihood of developing HIV mutations that could lead to the development of drug-resistant viral strains. HAART also prevents further viral destruction of the cellular immune system, thereby, allowing for increases in the level of CD4+ cells, which improves the immunologic response to opportunistic infections (Castro, 2005).

2.1.2 Socioeconomic Importance

Socioeconomic relevance of ART is accepted worldwide. According to the Annual Report of Federal **HAPCO (2007a)**, the primary socio-economic importances of ART are: improvement of quality of life, reduction of HIV-related morbidity and mortality, increase in uptake of HIV testing contributing to prevention as well as care of HIV infection, change in perceptions of HIV society regards AIDS as a treatable chronic disease, raise of the hope of individuals, families and community thereby reducing HIV-related stigma and discrimination, enhancement of affected person's participation in HIV intervention activities, increase in uptake of PMTCT, reduction of transmission of HIV in the population and reduction of incidence of Tuberculosis.

2.2 ART Adherence Issues

2.2.1 Adherence and Drug resistance

A consensus exists that in order to achieve an undetectable viral load and prevent the development of drug resistance, a person on HAART needs to take at least 95% of the prescribed doses on time (Paterson & Swindells et al, 2000). For many people, this means taking a regimen of three antiretrovirals twice per day—on both occasions, they are usually taking several pills (Partners In Health, 2004). An increasing number of studies show that the relationship between adherence and resistance is drug specific (Bangsberg, Moss, & Deeks, 2003), but after introduction of HAART, adherence was not highly correlated with specific drug regimen, according to Friedland (1999).

Given the relevance of treatment adherence to improving life expectancy and preventing the spread of drug-resistant strains, many studies have attempted to predict causes of non-adherence in order to design strategies that reduce the number of missed doses. Except for some factors that have been associated with incomplete adherence in various settings, such as depression (Carrieri et al, 2001), or illegal drug use (Ware et al, 2005), study results are often inconclusive and do not yield comparable results—often due to conceptual and methodological differences among research protocols. Methodologically, there is growing agreement that patients' self-assessments of adherence—through interviews or self-administered questionnaires—show significant correlation with viral load tests (Sethi et al,2003), whereas estimations by their health-care providers often lead to invalid results (Moatti et al, 2004).

Despite the current shortcomings in predicting who is more likely to miss doses of antiretroviral therapy (ART), the issue of adherence has become extremely important when setting priorities for allocating resources to fight AIDS in poor countries, where the majority of people who are HIV positive live. Some commentators have argued that adherence barriers are insurmountable in poor settings, so we should be cautious in delivering ART to these populations—a claim that is not grounded in evidence. In fact, adherence in poor settings is proving to be equal to, or even higher than, adherence in developed countries (Coetzee et al, 2004; Katzenstein et al, 2003; Koenig, 2004; Nemes et al, 2004; Pérez et al, 2004). Furthermore, the argument about insurmountable adherence barriers in poor settings has also been challenged because it creates an unjustifiable double standard (Moatti et al, 2004)—ART is not withheld from wealthy settings on the basis that many patients will skip doses.

2.2.2 Adherence as a Biosocial and Dynamic Phenomenon

According to several scholars, a biosocial approach to adherence relies on the dynamic analysis of the clinical and social course of disease and the continuous interaction of biological and social processes over time (Kleinman et al, 1995; Farmer, 1990; Singer & Clair, 2003; Castro & Farmer, 2005). The study of pathology embedded in social experience captures a series of distal and proximal factors acting together—such as not seeking treatment for undesirable side effects due to lack of money to travel to a health center or purposely missing doses when one is asymptomatic to pretend that AIDS is not a concern. These factors shape the everyday life of patients, while patients internalize them and use them to provide meaning to their disease experience.

Years ago, before the advent of AIDS, anthropologists had noted that the introduction of effective therapy for a particular disease may profoundly alter the social interpretations of that disease (Kleinman et al, 1995; Farmer, 1990; Singer & Clair, 2003; Castro & Farmer, 2005; Lévi-Strauss, 1958). Exposure to AIDS in Haiti, in the 1980s, generated cultural models of its etiology and expected course (Farmer, 1990; Farmer, 1992; Farmer, 1994), which aimed to provide meaning to otherwise unknown phenomena—often locally interpreted in terms of jealousy and curse. Likewise, the social experience of AIDS in rural Haiti is also deeply affected by the advent of effective therapy, as preliminary data suggest that the introduction of quality HIV care can lead to a rapid reduction in stigma, with resulting increased uptake of testing (Castro & Farmer, 2005).

Within the changing context in which disease may take place, adherence level is likely to change as biological and social circumstances, and interpretations of them, unfold. In Brazil, a boy and a girl living in a support house for children and adolescents orphaned by AIDS or living with HIV explained why, upon feeling well, they had started to delay the morning dose of antiretrovirals until the afternoon: “[The schedule] doesn't matter; you can take them any time, as long as you take them” (Abadía-Barrero & Castro, 2005). Until they started HAART, they had both suffered several opportunistic infections and, at times, had been at the brink of death. Their incomplete adherence to the regimen was, in part, a lack of understanding about the importance of not missing doses—but, fundamentally, it was a strategy they had developed as an act of defiance against the rules of the support house, and probably to feel more like the children who were HIV negative who also lived there and were not taking daily medications (Abadía-Barrero & Castro, 2005). By delaying the morning dose, they were providing their lives with a sense of normalcy, which had otherwise been characterized by orphanhood and chronic disease since early age, while acting out in protest.

By any account, these children would be classified as having incomplete adherence or, plainly, as non-adherent. Yet to improve their adherence, how helpful is it to know that they are not taking all their medications on time without understanding why? Could their non-adherence really be understood without analyzing the life trajectories of these children and their social context? Some studies approaching ART adherence within a dynamic framework have relied on a biosocial approach, providing a rich context in which taking medications occurs and evolves (Ware et al, 2005; Desclaux, 2003; Ickovics & Meade, 2002). Other studies, while having observed that the level of adherence is not a static value, have remained mostly biomedical, examining the impact of side effects (Ammassari et al, 2002) or substance abuse (Tesoriero et al, 2003), and have been devoid of the complexity of biosocial interactions and their changing nature, as in the case of the Brazilian children.

The study of life histories of patients—a standard qualitative method of ethnographic research—and of the interactions of social experience with illness episodes allows us to generate associations between the clinical and the social course of disease, including such themes as stigma, health-seeking behavior, or adherence to therapy. These associations, which are often drawn from a small sample of patients, can be validated by larger, statistically representative, epidemiological studies designed to include variables reflecting the social context of patients. For example, the effect of adipose tissue alterations (lipodystrophy)—a common side effect of PIs—on adherence to ART has often been studied without considering local patterns of ideal weight and body shape for women and men at different ages, and existing variations of these ideals related to the social position of patients. This lack of consideration for local patterns may explain why some studies arrive at opposite results—dystrophic weight gain as a barrier to or as an enhancer of adherence (Ammassari et al, 2002). The effect of adherence on weight gain seen in patients on ART, as a result of a reversal of the disease process and general clinical improvements, should also be analyzed in relation to these social ideals. In Senegal, for example, where weight gain is a symbol of good health, such weight gain has been shown to increase adherence (Desclaux, 2003).

Most research studies on adherence to ART share the basic understanding that patients are adherent when they, after agreeing to the recommendations of a health-care provider (World Health Organization, 2003), take the prescribed medications in a timely manner. However, an overemphasis on pill counting as a sum of discrete events limits our understanding of adherence as a complex process embedded in the clinical and social course of AIDS, as the case of the Brazilian children shows. An approach to adherence that combines both biological and social knowledge—a biosocial approach—and that relies on qualitative and quantitative methodologies is more likely to move us closer to a better understanding of adherence and, eventually, to improving adherence to ART.

2.2.3 Poverty and Adherence

Partly because the introduction of ART in poor settings is recent, and partly because biomedical research rarely examines the social context in which patients live, there is a dearth of information on the direct effect of poverty on ART adherence. Most studies conducted in poor settings overlook how direct and indirect economic burdens borne by patients affect their ability to access a steady supply of antiretrovirals and take them on time. Such burdens may include the cost of missing work, the cost of elder or child care during medical visits, the cost of transportation to a health center, the cost of user fees, or the cost of tests and supplies. Although these costs may seem minimal to health professionals and decision makers, bearing these costs often translates into difficult household decisions about who eats, who works, or who goes to school. Taking medications in a timely manner may also require the challenging tasks of obtaining food and safe water, or of readjusting food intake to fit the drug-regimen schedule.

Despite the difficulties in overcoming these obstacles, the inability of a person living in poverty to obtain and take medications after initiating therapy is often labeled "noncompliance" or "non-adherence"—as has often occurred with tuberculosis patients (Castro & Farmer, 2003; Greene, 2004)—and categorized as patient-related characteristics, ignoring social and economic causes or failures on the part of public-health interventions to address those causes (Castro & Singer, 2004). Some studies conducted in Côte d'Ivoire (Delaunay et al, 2005), Senegal (Desclaux, 2004;), and Botswana¨(Nwokike, 2003) show that user fees not only deter people from accessing AIDS care but also create an obstacle to treatment adherence. In other contexts where ART is free, such as Costa Rica, transportation costs have been associated with lower adherence (Stout et al, 2003). The argument that patients would not value free drugs, or that free treatment might be humiliating for patients, are not borne out by higher adherence rates when drugs are free, such as in a comprehensive AIDS program in rural Haiti (Koenig et al, 2004) or in Cuba (Pérez et al, 2004), or when user fees are lowered, such as in Senegal (Desclaux, 2004). Indeed, variations of directly observed therapy (DOT) for the delivery of ART (known as DOT-HAART) have proven useful in introducing complex multidrug regimens in poor settings lacking health-care infrastructures. In rural Haiti, for example, support for patients receiving DOT-HAART from community health workers improves rates of adherence (Koenig, 2004).

As adequate and equitable access to comprehensive AIDS prevention and care are introduced, optimal adherence could be achieved if the multiple causes that shape patients' adherence are analyzed within their social context—including those related to the financing of health-care systems, and particularly cost-recovery mechanisms.

2.2.4 A Biosocial Approach to Causes of Non-adherence

The use of a biosocial framework grounded in the lived experience of people diagnosed with AIDS is essential to understanding adherence, the way adherence changes over time, and the reasons for non-adherence. Often times, particularly in poor settings, these reasons will be found outside the individual responsibility of patients. Addressing adherence may require providing social support to patients, lowering or eliminating user fees, bringing health-care workers closer to patients, opening health centers focused on patients' competing demands to survive, improving drug procurement strategies, or creating mechanisms for lowering the cost of drugs and lab tests. In many cases it will mean improving and investing in primary health care, public hospitals, and referral networks, or in incentives to recruit and retain health-care workers committed to serving their patients. Given this complexity, the possibility of advancing the understanding of the multifaceted causes of non-adherence needs to be analyzed within its larger social, economic, and political context.

According to Castro (2005), the main biosocial variables needed to analyze adherence to ART defined within eight broad categories:

- **Socioeconomic factors:** Poverty and inequality, War, Political violence, Cost of medications, Cost of CD4+ counts, viral load, Transportation costs, Cost of missing days from work, Cost of food and safe water, Costs associated to changes in lifestyle, and other costs.

- **Health-Care System:** Health-care infrastructure, Drug stock shortages, Financing mechanisms (including user fees), and Quality of relationship with health-care providers.

- **Social Capital:** Kinship patterns, Networks of social support, Social status, Homelessness, and incarceration

- **Cultural Models of Health and Disease:** On etiology and transmission, on health-care providers and healers, on cure, on efficacy and toxicity of drugs, and on sick role.

- **Personal Characteristics:** Age, Sex and gender, Ethnic group, Education, Occupation, Household composition, Substance abuse, and Physical disabilities.

- **Psychological Factors:** Self-esteem, motivation, and Mental health conditions.

- **Clinical Factors:** Immunological or clinical stage of HIV disease, Occurrence and severity of opportunistic infections, Side effects (desirable and undesirable), Symptomatology at onset of treatment, and Effect of pregnancy or lactation.

- **Antiretroviral Regimen:** Number of drug regimens per day, Number of pills per regimen, and Therapeutic class composition of drug regimen.

Castro (2005) believes that most studies have focused on the clinical aspects only, which are easier to measure quantitatively but which do not account for the larger social context. Although clinical epidemiological studies are essential to finding associations between drug regimens and adherence—and, depending on the method chosen, to establishing causality—a biosocial approach that combines quantitative and qualitative methodologies is necessary to bridge the current gap in knowledge on adherence to ART. Only by understanding the complicated interplay between the clinical and social factors that affect adherence to ART can we hope to overcome the real causes of non-adherence.

2.3 ART Adherence Challenges and Opportunities in Ethiopia [D]

With multiple causes and far reaching consequences, HIV/AIDS has caused untold human misery and socio-economic crisis over the past 25 years (WIC, 2008). Aggravated by economic, social, cultural and behavioral factors, HIV/AIDS has even gone far to the extent of becoming a development constraint. One does not necessarily be a professional economist to recognize the huge annual cost of productivity lost due to HIV/AIDS and the provision of treatment.

[D] Based on Walta's Report (WIC, 2008)

At the initial stages of the epidemic, research were focused on the biomedical behavior of the virus and the immune response of the human body. Meanwhile, the HIV virus spread at an alarming rate before it was eventually realized that socio-economic and cultural aspects influence human behavior and put individuals at high risk of infection.

According to HIV/AIDS Prevention and Control Office (HAPCO, 2007c) single point estimate 2007, about 1.3 million people live with HIV and some 10,825 people have died of AIDS in this year only. Over 898,350 have lost their parents or at least either parent due to HIV/AIDS. The situation is worse in Addis Ababa as the center of political and economic interaction.

The response against the epidemic gradually evolved from risk focused prevention to care, support and treatment to the infected. The whole process requires identifying people with HIV, putting them on treatment as quickly as possible and scaling up treatment for the people needing the treatment. But this is not as simple as we might think. In this country, HIV treatment is needed by about a million people.

In Ethiopia today more than 250,000 people living with HIV require anti-retroviral treatment but only 145,000 are receiving the treatment, according to sources [(E)].

It can be clearly seen from the above figures that hundreds of thousands HIV positive people requiring ART are missing out on treatment. Obviously, this poses a huge challenge to address. Yet, the greatest challenge is to ensure that people adhere to HIV/AIDS treatment (WIC, 2008).

[E] The "Source" refers to HAAPCO (HIV/AIDS Prevention & Control Office)

Unlike other treatments, HIV/AIDS treatment is a lifetime treatment. It is not the kind of treatment where one gets treated and sometime later everything is gone. ART for HIV is a lifelong treatment. So people need to be comfortable with it; they should be able to talk about their infection which is highly stigmatized under many circumstances. They need to be able to advocate for themselves and their families. And above all, they must adhere to the treatment.

Yet adherence to ART faces a broad array of challenges ahead. One of the major challenges is that in our country, HIV/AIDS is inseparably mixed with sheer poverty both at the national as well as individual levels. This is manifested in the fragility of the health care system that most of the institutions need more trained people, more space, more chain of supply, so and so forth.

The people that are treated need a variety of services that we collectively call 'care and support' ranging from food, transportation, family support to additional treatment if the primary treatment doesn't work.

The other important factor is that about 85 per cent of Ethiopia is really rural than urban (CSA, 2008). So reaching out to people as close to where they can get health care has been a real challenge.

Until very recently, ART used to be provided at a hospital level only. But most hospitals are located in cities and major towns. The vast majority of the rural public had difficulty to access one. Even with the decentralization of ART services at the health center level since recently, many people still have to travel over a long distance to reach a health center (MOH Roadmap, 2007).

As a matter of fact, HIV treatment requires regular journey to health care institutions. Usually people are given drugs for one month to see that they tolerate it. Then they are given for three months supply and eventually for a whole year. And there will be the case for sometime people will have to come back to see the facility to have check ups and to get their drugs. So they have to come back periodically for a lifetime if they have to survive HIV.

ART adherence is also influenced by a wide range of factors such as misconceptions, religious beliefs and practices as well as peoples' risk behaviors. People sometimes give up treatment due to side effects (Saitoh et al, 2005), shift to holy waters, discontinue treatment because of fasting or even forget to take their drugs at the right time. Yet, for the drugs to be effective people should adhere to the treatment 100 per cent; not even 95 percent!

Adhering to ART requires people to stop taking alcohol, smoking and chewing ***'chat'*** as these substances would counter react with the drugs and also increase vulnerability of patients (McNicholl, 2005). Although breaking a habit is not an easy thing to do, people on HIV treatment should avoid substance abuse at any cost if they are to survive the virus.

Many people have died of AIDS long before ART was available. But these days while ART is available in every health center people should not die of their own risk behaviors.

In closing, the primary mission to prevent and control the HIV/AIDS epidemic should focus on ensuring that people who are sexually active understand the risks and prevent themselves as well as others from becoming infected. And the people who are already infected, we have to make sure that they are on treatment, and of course, that they adhere to that treatment.

2.4 Summary

For long, suboptimal treatment adherence has been associated with virologic, immunologic, and clinical failure. In this literature review, however, we looked critically at the issue of adherence, and argued that, to address causes of incomplete adherence, we need to combine both quantitative and qualitative methodologies. These methodologies must be grounded in an understanding of adherence not only as a biological but also as social process that changes with time, and must be framed within an analysis of access to health care and medications.

CHAPTER THREE

BAYESIAN NETWORK TECHNOLOGY

3.1 Introduction

The term "Bayesian" refers to **Thomas Bayes** (1702–1761), who proved a special case of what is now called **Bayes' theorem**. Although Bayes' theorem has been in use for more than two hundred years (Bayesian inference), the **Bayesian interpretation of probability** is more recent. This chapter deals with the concepts of Bayes' theorem, Bayesian Probabilities, Bayesian Network Model, and some Mathematical theories behind the BN Model-that are relevant to the research. It also discusses BN Model as a potential tool in knowledge exploration in the healthcare sector.

3.2 Basics of Bayesian probability

3.2.1 Elements of Probability Theory

Probability theory provides the mathematical rules for assigning probabilities to outcomes of random experiments. Basic elements of probability theory are:

- **Sample space:** set of all possible "elementary" or "finest grain" outcomes of the random experiment.
- **Set of events F:** set of (all) subsets of Ω —an event A C Ω occurs if the outcome ω C A
- **Probability measure P:** function over F that assigns probabilities to events according to the axioms of probability.

Formally, a probability space is the triple **(Ω, F, P).**

3.2.2 Axioms of Probability

A probability measure P satisfies the following axioms:

1. P (A) $\geq$ 0 for every event A in F

2. P (Ω) = 1

3. If A_1, A2 . . . are disjoint events—i.e., $A_i \cap A_j = ø$, for all $i \neq j$ —then

$$P(\bigcup Ai) = \prod_{i=1}^{n} P(Ai)$$

Notes:

- P is a measure in the same sense as mass, length, area, and volume—all satisfy axioms 1 and 3
- Unlike these other measures, P is bounded by 1 (axiom 2)
- This analogy provides some intuition but is not sufficient to fully understand probability theory—other aspects such as conditioning and independence are unique to probability.

3.2.3 Prior and Conditional Probabilities

Conditional probability is the probability of some event *A*, given the occurrence of some other event *B*. Conditional probability is written $P(A|B)$, and is read "the probability of *A*, given *B*".

Joint probability is the probability of two (or more) events in conjunction. That is, it is the probability of both events together. The joint probability of *A* and *B* is written as:

P (A $\cap$ B) or P (A, B).

Marginal probability is then the unconditional probability $P(A)$ of the event A; that is, the probability of A, regardless of whether event B did or did not occur. If B can be thought of as the event of a random variable X having a given outcome, the marginal probability of A can be obtained by summing (or integrating, more generally) the joint probabilities over all outcomes for X. For example, if there are two possible outcomes for X with corresponding events B and B', this means that $P(A) = P(A \cap B) + P(A \cap B')$. This is called **marginalization**.

Conditioning of probabilities, i.e. updating them to take account of (possibly new) information, may be achieved through Bayes' theorem. In such conditioning, the probability of A given only initial information I, $P(A|I)$, is known as the **prior probability.** The updated conditional probability of A, given I and the outcome of the event B, is known as the **posterior probability**, $P(A|B, I)$.

3.2.4 Conditional Independencies

Independence: Two variables, A and B, are independent if their conditional probability is equal to their unconditional probability. In other words, A and B are independent if, and only if, $P(A|B) = P(A)$, and $P(A|B) = P(B)$. In engineering terms, A and B are independent if knowing something about one tells nothing about the other. This is the origin of the familiar, but often misused, formula $P(AB) = P(A) \times P(B)$, which is true only when A and B are independent.

Conditional Independence (CI): A and B are conditionally independent, given C, if $P(A=a, B=b \mid C=c) = P(A=a \mid C=c) \times P(B=b \mid C=c)$ whenever $P(C=c) > 0$. So the joint probability of ABC, when A and B are conditionally independent, given C, is then P(C) X (A | C) X P (B | C). A directed graph illustrating this conditional independence is:

A ←C →B

For example,

P (Adherence| Addiction=Yes, Sex=Female) = P (Adherence| Addiction=Yes)

Here, Adherence is conditionally independent of Sex give Addiction.

3.2.5 Bayes' Theorem and Conditional Independence

Bayes' theorem relates the conditional and marginal probabilities of events *A* and *B*, where *B* has a non-vanishing probability:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}.$$

Each term in Bayes' theorem has a conventional name:

- P (*A*) is the **prior probability** or **marginal probability** of *A*. It is "prior" in the sense that it does not take into account any information about *B*.
- P (*A*|*B*) is the **conditional probability** of *A*, given *B*. It is also called the **posterior probability** because it is derived from or depends upon the specified value of *B*.
- P (*B*|*A*) is the conditional probability of *B* given *A*.
- P (*B*) is the prior or marginal probability of *B*, and acts as a **normalizing constant**.

Intuitively, Bayes' theorem in this form describes the way in which one's beliefs about observing 'A' are updated by having observed 'B'.

We can derive Bayes' theorem from conditional probabilities. To do that, we start from the definition of **conditional probability**. The probability of event *A* given event *B* is:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Equivalently, the probability of event *B* given event *A* is

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

Rearranging and combining these two equations, we find

$$P(A|B)\, P(B) = P(A \cap B) = P(B|A)\, P(A).$$

This lemma is sometimes called the **product rule** for probabilities. Dividing both sides by P (*B*), providing that it is non-zero, we obtain Bayes' theorem:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)\, P(A)}{P(B)}.$$

More generally, where $\{A_i\}$ forms a **partition** of the event space,

$$P(A_i|B) = \frac{P(B|A_i)\, P(A_i)}{\sum_j P(B|A_j)\, P(A_j)},$$

for any A_i in the partition.

Theorems analogous to Bayes' theorem hold in problems with more than two variables. For example:

$$P(A|B \cap C) = \frac{P(A)\, P(B|A)\, P(C|A \cap B)}{P(B)\, P(C|B)}.$$

This can be derived in a few steps from Bayes' theorem and the definition of conditional probability:

$$P(A|B \cap C) = \frac{P(A \cap B \cap C)}{P(B \cap C)} = \frac{P(C|A \cap B)\, P(A \cap B)}{P(B)\, P(C|B)} = \frac{P(A)\, P(B|A)\, P(C|A \cap B)}{P(B)\, P(C|B)}.$$

Similarly,

$$P(A|B \cap C) = \frac{P(B|A \cap C)\, P(A|C)}{P(B|C)},$$

which can be regarded as a conditional Bayes' Theorem, and can be derived by as follows:

$$P(A|B \cap C) = \frac{P(A \cap B \cap C)}{P(B \cap C)} = \frac{P(B|A \cap C)\, P(A|C)\, P(C)}{P(C)\, P(B|C)} = \frac{P(B|A \cap C)\, P(A|C)}{P(B|C)}.$$

A general strategy is to work with a decomposition of the **joint probability**, and to **marginalize** (integrate) over the variables that are not of interest. Depending on the form of the decomposition, it may be possible to prove that some integrals must be 1, and thus they fall out of the decomposition; exploiting this property can reduce the computations very substantially. A **Bayesian network**, for example, specifies a factorization of a **joint distribution** of several variables in which the conditional probability of any one variable given the remaining ones takes a particularly simple form.

3.3 Fundamentals of Bayesian networks

3.3.1 Definition of Bayesian Network

The Bayesian belief network is a powerful knowledge representation and reasoning tool under conditions of uncertainty. A Bayesian belief-network is a directed acyclic graph (**DAG**) with a conditional probability distribution for each node (Pearl, 1988; Neapolitan, 2004). The DAG structure of such networks contains nodes representing domain variables, and arcs between nodes representing probabilistic dependencies. On constructing Bayesian networks from databases, we use nodes to represent database attributes.

More formally, Pearl (1988) represented a Bayesian network by $\boldsymbol{BN} = \langle \boldsymbol{N}, \boldsymbol{A}, \boldsymbol{\theta} \rangle$, where $\langle \boldsymbol{N}, \boldsymbol{A} \rangle$ is a directed acyclic graph (DAG) – each node $\boldsymbol{n} \in \boldsymbol{N}$ represents a domain variable and each arc $\boldsymbol{a} \in \boldsymbol{A}$ between nodes represents a probabilistic dependency between the associated nodes. Associated with each node $n_i \in N$ is a conditional probability table (CPT), collectively represented by $\theta = \{\theta\}$, which quantifies how much a node depends on its parents.

Russel and Norvig (2003) stated that the structure of a Bayesian network is a graphical illustration of the interactions among the set of variables that it models. They explained the full specification as follows:

a. A set of random variables makes up the nodes of the network;
b. A set of directed links connects pairs of nodes. If there is a directed link from node A to node B, A is said to be a parent of B;
c. Each set contains a finite set of mutually exclusive states;
d. Each node A has a conditional probability table P (A| par (A)) that quantifies the effect of the parents on the node. If the variable A does not have any parent, then the table can be replaced by prior probabilities, i.e. P(A);
e. The variables coupled with the directed edges construct a directed acyclic graph (DAG).

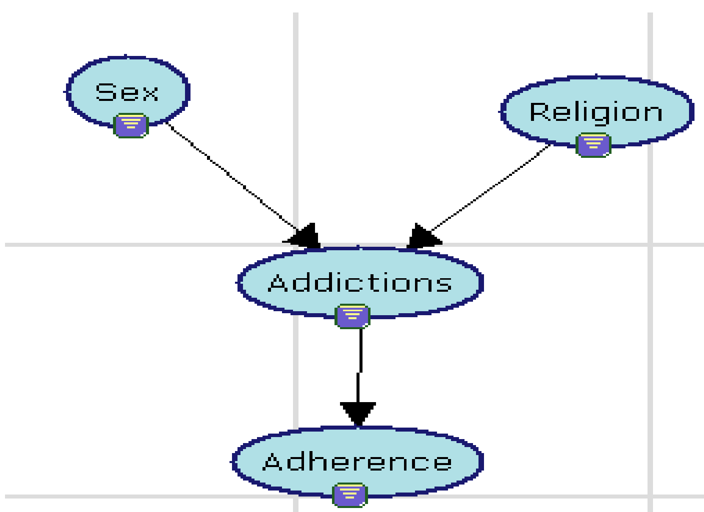


Figure 1: A Bayesian Network with four attributes

Figure 1 shows a Bayesian network consisting of four discrete variables: **Sex, Religion, Addiction,** and **Adherence**. At the quantitative level, the dependence relations are expressed in terms of conditional probability distributions for each variable in the network. Each variable X has a set of possible values called its *state space* that consists of mutually exclusive and exhaustive values of the variable. In Figure 1, e.g., Sex has two states: 'Male and 'Female;' Religion has five states: 'Orthodox', 'Protestant', 'Muslim', 'Catholic' and 'Others'; and Addiction has two states: 'Yes and 'No'; and Adherence has two states: 'Adhere' and 'Drop.' If there is an arc pointing from X to Y, we say X is a *parent* of Y. In Figure 1, Sex and Religion have no parents. However, Addiction has two parents (Sex and Religion) and Adherence has one parent (Addiction).

The following figure shows the screen print of the CPT table generated after the learning process with the four attributes was completed and thereby exported to *BNJ Tools with Java.*

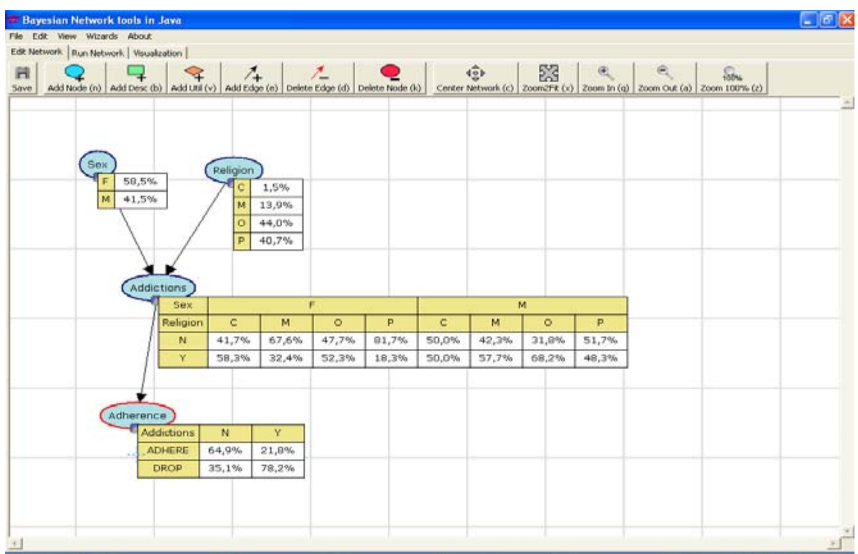


Figure 2: Bayesian Network with CPT table (BNJ Tools with Java interface)

For each variable, we need to specify a table of conditional probability distributions, one for each configuration of states of its parents. Figure 2 shows these tables of conditional distributions—**P (Sex), P (Religion), P (Addiction | Sex, Religion), and P (Adherence|Addiction).**

3.3.2 Semantics of Bayesian Network [F]

A fundamental assumption of a Bayesian network is that when we multiply the conditionals for each variable, we get the joint probability distribution for all variables in the network. In Figure 1, e.g., we are assuming that:

P (Sex, Religion, Addiction, Adherence) = P (Sex) * P (Religion) * P(Addiction | Sex, Religion) *P (Adherence | Addiction),

where * denotes pointwise multiplication of tables. The rule of total probability tells us that:
P (Sex, Religion, Addiction, Adherence) = P (Sex) * P (Religion | Sex) * P(Addiction | Sex, Religion) * P (Adherence | Sex, Religion, Addiction).

Comparing the two, we notice that we are making the following assumptions: P(Religion | Sex) = P(Religion), i.e., Religion is independent of Sex; and P(Adherence | Sex, Religion, Addiction) = P(Adherence | Addiction), i.e., **Adherence is conditionally independent of Sex and Religion given Addiction.** Notice that we can read these conditional independence assumptions directly from the Bayesian network graph as follows. Suppose we pick a sequence of the variables such that for all directed arcs in the network, the variable at the tail of each arc precedes the variable at the head of the arc in the sequence. Since the directed graph is acyclic, there always exists such a sequence. In Figure 1, e.g., one such sequence is Sex-Religion-Addiction-Adherence. Then, the conditional independence assumptions can be stated as follows: For each variable in the sequence, we are assuming it is conditionally

[F] Semantics in this context means- the study of ways of interpreting and analyzing theories of logic independent of its predecessors in the sequence **given its parents**. The essential point here is that missing arcs (from a node to its successors in the sequence) signify conditional independence assumptions. Thus, the lack of an arc from Sex to Religion signifies that Sex is independent of Religion; the lack of an arc from Sex to Religion and from Religion to Adherence signifies that Adherence is conditionally independent of Sex and Religion given Addiction.

In general, there may be several sequences consistent with the arcs in a Bayesian network. In such cases, the list of conditional independence assumptions associated with each sequence can be shown to be equivalent using the laws of conditional independence (Lauritzen et.al., 1990)

Unlike a causal map, the arcs in a Bayesian network do not necessarily imply causality. The (lack of) arcs represent conditional independence assumptions. How are conditional independence and causality related? Conditional independence can be understood in terms of relevance. In our ART Adherence example, Adherence is conditionally independent of Sex and Religion given Addiction. This statement can be interpreted as follows. If the true state of Addiction is known, then in assigning probabilities to states of Adherence, the states of Sex and Religion are irrelevant. In other words, if we know that somebody is addicted, then any knowledge about his/her sex and religion is irrelevant to the probabilities of adherence of the person.

In practice, the notion of direct causality is often used to make judgments of conditional independence. Consider in our adherence example, a situation where Sex directly causes Addiction and Addiction in turn directly causes Adherence, i.e., the causal effect of Sex on Adherence is completely mediated by Addiction. Then it is clear that although Sex is relevant to Adherence, if we know the true state of Addiction, further knowledge of Sex is irrelevant (for assigning probabilities) to Adherence, i.e., Adherence is conditionally independent of Sex given Addiction. Such a connection is called **Serial Connection**. This situation is represented by the Bayesian network:

$$\textbf{Sex} \Rightarrow \textbf{Addiction} \Rightarrow \textbf{Adherence}$$

in which there is no arc from Sex to Adherence. Here, Sex and Adherence are said to be **D-separated** (Direction Separated).

As another example, consider a situation where variable X directly causes variable Y and variable X also directly causes variable Z. Although knowledge of Y is relevant to Z (if Y is true then it is more likely that X is true which in turn means that it is more likely that Z is true), once we know the true state of X, then further knowledge of Y is irrelevant to Z, i.e., Y is conditionally independent of Z given X. Such a connection is called **Diverging Connection**. This situation is represented by the Bayesian network:

$$\mathbf{Z} \Leftarrow \mathbf{X} \Rightarrow \mathbf{Y}$$

in which there is no arc from Y to Z or vice-versa. Finally as a third example, consider the situation where X and Y are two independent direct causes of Z, i.e., X and Y are unconditionally independent. But if we learn something about the true state of Z, then X and Y are no longer irrelevant to each other (if Z is believed to be true and X is false, then it is more likely that Y is true), i.e., Y is not conditionally independent of X given Z. Such a connection is called **Converging Connection**. This situation is represented by the Bayesian net:

$$\mathbf{X} \Rightarrow \mathbf{Z} \Leftarrow \mathbf{Y}$$

in which there is no arc from X to Y or vice-versa. (Nadkarn & Shenoy, 2004). The relationship among the nodes **Sex**, **Religion** and **Addiction** is of converging type.

The three connections explained above let somebody see all the forms in which evidence may be transmitted though a variable. One can decide for any pair of variables in a causal network whether or not they are dependent once knowing the evidence entered into the network.

As we have already discussed above, the main advantage of Bayesian networks is the ability to define the conditional independencies first, before specifying numerically the actual conditional probability distributions. A general conditional independence property of Bayesian networks is that any variable X in the network is conditionally independent of its non-descendents **ND(X)** given its parents **par(X)** (Pearl, 1988). That is, if a variable's parents become known, then any information about nodes that are not on a directed path from X will be irrelevant. This is the so-called directed Markov property of Bayesian networks.

3.4 Learning Bayesian Networks

According to Neapolitan (2004), conditional probabilities can be assessed by the expert, learned from data, or obtained using a combination of both techniques. However, eliciting Bayesian networks from experts can be a laborious and difficult procedure in the case of large networks. To avoid the problem, methods that could learn the DAG (**structure**) as well as the conditional probability distributions from data (**parameter)** were developed. For this research, probabilities were learned from data.

3.4.1 Learning Bayesian Networks from Data

In many fields of studies including biomedicine and health-care, data have been collected and maintained, sometimes over numerous years. Such a data collection usually contains highly valuable information about the relationships between the variables discerned, be it implicitly. If a comprehensive data set is available, a Bayesian network can be learnt from the data, that is, it can be developed without explicit access to knowledge of human experts.

Learning a Bayesian network from data involves the tasks of structure learning, that is, identifying the graphical structure of the network, and parameter learning, that is, estimating the conditional probability distributions to be associated with the network's digraph. In many learning algorithms, the two tasks are performed simultaneously and, as a consequence, are not easily distinguished. To this end, the process of learning Bayesian networks from data takes four different forms, in terms of whether the structure of the network is known and whether the data is complete. These are:

- Unknown network structure and complete data
- Known network structure and complete data
- Unknown network structure and incomplete data
- Known network structure and incomplete data.

Learning with complete data indicates that the training data contains no missing values, while, learning with incomplete data indicates that some piece of information in the data are not known. It is essential to remember here that each of the cases mentioned above has its own learning algorithms.

Algorithms related to incomplete data were not the interest of this research as they were not relevant for the experimentation. Therefore we jump discussions related to incomplete data.

3.4.1.1 Learning with Complete Data-TPDA

One of the approaches to learning a Bayesian network from data is to build upon the use of a dependence analysis (Cheng et al, 1997). A Bayesian network in essence models a collection of conditional dependence and independence statements, through its Markov condition [G]. By studying the available data set, the dependences and independences between the various variables can be extracted, for example, by means of statistical tests, and subsequently captured in a graphical structure. The information-theoretical algorithm of Cheng et al (1997) is an example of an algorithm taking this approach. The algorithm is known as **Three-Phase-Dependency-Analysis (TPDA)**. The algorithm has three subsequent phases termed **drafting**, **thickening** and **thinning**. In the drafting phase, the algorithm establishes, from the data, the mutual information for each pair of variables and constructs a draft digraph from this information. In the thickening phase, the algorithm adds arcs between pairs of nodes if the corresponding variables are not conditionally independent given a certain conditioning set of variables. In the thinning phase, to conclude, each arc of the graph obtained so far is examined using conditional independence tests, and is removed if the two variables connected by the arc prove to be conditionally independent. An optional **orienting** phase (in which age orientation is conducted) is also part of the algorithm.

TPDA learning algorithm tells us which nodes should be joined by arcs. The algorithm works **incrementally** (i.e., at each point, it has a current set of arcs, and is considering adding some new arc, or deleting and existing one). Decisions are based on information flow between a pair of nodes.

In TPDA algorithm:

[G] Any variable X in the network is conditionally independent of its non-descendents **ND(X)** given its parents **par(X).** This condition is known as Markov Condition

- A Bayesian network is viewed as a network of information channels or pipelines
- Each node is a valve that is either active or inactive
- The valves are connected by noisy information channels (arcs)
- Information can flow through an active valve but not an inactive one.

There are, basically, **two** main approaches for learning the structure of a network from complete data: Constraint-based and Score-based.

A. Score-Based

This approach defines a score that evaluates how well the (in) dependencies in a structure match the data, and, search for a structure that maximizes the score. In this technique, the best Bayesian Network is the one that best fits the data.

B. Constraint-Based:

This approach performs conditional independence tests (tests such as **chi-squared test**) on the data. Then it will search for a network that is consistent with the observed dependencies and independencies (applies d-separation concept). Conditional independence relationships among the attributes can serve as **constraints** to construct a Bayesian Network. There are some publicly known algorithms to build Bayesian Network using this approach, TPDA being the prominent one.

According to a preliminary investigation made by Rahel (2005) for the purpose of selecting an appropriate learning algorithm from those which are publicly available for learning with complete data, **TPDA** algorithm was found to perform better. Depending on the finding, the algorithm was also adopted in this research. In what follows, TPDA algorithm with both unknown and known structures (and complete data) is discussed.

3.4.1.1.1 TPDA with Unknown Network Structure and Complete Data

TPDA learning algorithm with unknown structure and complete data takes a database table as input and constructs a Bayesian network structure as output. In other words, for unknown structure learning with complete data, the learning algorithm is given the set of variables in the model and needs to select the arcs between them and to estimate the parameters. According to Rahel (2005), Unknown Structure Learning with Complete data is useful in the following situation:

- In the absence of domain expert
- When we want to get all of the benefits of a Bayesian network model
- When we want to give the expert some indications of what attributes are correlated

Bayesian learning method, with unknown structure and complete data, involves **three** phases:

I. Manual selection of the model variables and their possible values
II. Automatic determination of the structure of the graph based on a Dataset (Here, either **Constrained-based** or **Score-based** approach can be implemented)
III. Automatic calculation of the conditional probability distribution.

Since node ordering is not given as input, this algorithm has to deal with two major problems (Rahel, 2005):

I. How to determine if two nodes are conditionally independent, and
II. How to orient the edges in a learned graph.

As described by Cheng et al (1997), the algorithm has four phases: ***drafting*** (computes the mutual information of each pair of nodes as a measure of closeness and creates a draft based on this information), ***thickening*** (adds edges when the pairs of nodes cannot be *d-separated*)**,** ***thinning*** (each edge of the current graph is examined using CI tests and will be removed if the two nodes of the edge can be *d-separated*) *and* ***orienting edges*** (edge orientation procedure). Even though the mathematical details of each phase are important, we skip it, as its applications are more relevant for this research work.

3.4.1.1.2 TPDA with Known Network Structure and Complete Data

If the network structure is already defined (i.e., if the structure is already known), the algorithm needs to estimate only the parameters (CPT) - using techniques such as Maximum Likelihood Estimation and Bayesian estimation. Accordingly, this algorithm takes as input both a table of database entries and a node ordering, and constructs a Bayesian Network structure as output.

The **first three phases** of this algorithm are the same as the TPDA algorithm described in the previous section. However, the last phase (orienting edges) described above, is not implemented in this algorithm, since the direction of the arcs are decided by the node ordering provided. The main features involved in these three phases, according to the discussions of Rahel (2005), are the following:

I. When direct cause and effect relations are available, it uses them as a basis for generating a draft in phase I.

II. In phase II, the algorithm will try to add an arc only if it agrees with the domain knowledge.

III. In phase III, the algorithm will not try to remove an arc if it is already specified by domain experts.

3.5 Application of Bayesian Networks

3.5.1 Problem Solving in HealthCare

According to recent findings, Bayesian networks and other probabilistic graphical models are beginning to emerge as methods for discovering patterns in biomedical data and also as a basis for the representation of the uncertainties underlying clinical decision making. At the same time, techniques from machine learning are being used to solve biomedical and health-care problems (Lucas, 2007).

Physiological mechanisms in human biology, the progress of disease in individual patients, and hospital work-flow management are just a few of the many complicated processes studied by researchers in biomedicine and health-care. For controlling the ever increasing complexity of these fields, a proper understanding of their processes is important as is the ability to reason about them. The characteristics of the processes vary widely; however, typically only part of all the factors by which they are governed can be observed in practice. The processes, moreover, include the effects of individual as well as random variation. Essentially they are uncertain; the uncertainties involved render an overall understanding hard to achieve and reasoning an overwhelming task. Models capturing these processes and methods for using these models are thus called for to support decision-making in real-life practice. Bayesian networks with their associated methods are especially suited for capturing and reasoning with uncertainty (Pearl, 1988). They have been around in biomedicine and health-care for more than a decade now and have become increasingly popular for handling the uncertain knowledge involved in establishing diagnoses of disease, in selecting optimal treatment alternatives, and predicting treatment outcome in various different areas.

Bayesian networks are also increasingly developed in areas of health-care that are not directly related to the management of disease in Individual patients. Examples include the use of Bayesian networks in clinical epidemiology for the construction of disease models and within bioinformatics for the interpretation of microarray gene expression data. This special issue aims to convey an impression of the current state-of-the-art of the use of Bayesian networks in biomedicine and health-care. By devoting attention to new application areas, it complements what is known about the use of Bayesian networks in building decision-support systems for individual patient care (Artificial Intelligence in Medicine, 2004).

Bayesian networks are increasingly used in biomedicine and health-care to support different types of problem solving, some of which are briefly reviewed here.

Diagnostic Reasoning: Establishing a diagnosis for an individual patient in essence amounts to constructing a hypothesis about the disease the patient is suffering from, based upon a set of indirect observations from diagnostic tests. Diagnostic tests, however, generally do not serve to unambiguously reveal the condition of a patient: the tests typically have true-positive rates and true-negative rates unequal to 100%. To avoid misdiagnosis, the uncertainty in the test results obtained for a patient should be taken into consideration upon constructing a diagnostic hypothesis. Bayesian networks offer a natural basis for this type of reasoning with uncertainty. A significant number of network-based systems for medical diagnosis have in fact been developed in the past and are currently being developed (Heckman et al, 1992; Andreassen et al, 1992).

To assist physicians in the complex task of diagnostic reasoning, a Bayesian network is often equipped with a test-selection method that serves to indicate which tests had best been ordered to decrease the uncertainty about the disease present in a specific patient (Andreassen, 1992). A test-selection method typically employs an information-theoretic measure for assessing diagnostic uncertainty. Such a measure is defined on a probability distribution over a disease variable and expresses the expected amount of information required to establish the value of this variable with certainty. An example measure often used for this purpose is the Shannon entropy. The measure can be extended to include information about the costs involved in performing a specific test and about the side effects it can have. Since it is computationally hard to look beyond the immediate next diagnostic test, test selection is generally carried out non-myopically, that is, in a sequential manner. The method then suggests a test to be performed and awaits the user's input; after taking the test's result into account, the method suggests a subsequent test, and so on.

Prognostic reasoning: Prognostic reasoning in biomedicine and health-care amounts to making a prediction about what will happen in the future. As knowledge of the future is inherently uncertain, inprognostic reasoning uncertainty is even more predominant than in diagnostic reasoning. Another prominent feature of prognostic reasoning when compared to diagnostic reasoning is the exploitation of knowledge about the evolution of processes over time. Even if temporal knowledge is not represented explicitly, prognostic Bayesian networks still have a clear general temporal structure. The outcome predicted for a specific patient is generally influenced by the particular sequence of treatment actions to be performed, which in turn may depend on the information that is available about the patient before the treatment is started. The outcome is often also influenced by progress of the underlying disease itself. Formally, a prognosis may be defined as a probability distribution:

$$P\,(\text{Outcome} \mid \varepsilon, \psi)$$

where **ε** is the available patient data, including symptoms, signs and test results, and Ψ denotes a selected sequence of treatment actions. The outcome of interest may be expressed by a single variable, e.g. modeling life expectancy. The outcome of interest, however, may be more complex, modeling not just length of life but also various aspects pertaining to quality of life. A subset of variables may then be used to express the outcome.

Prognostic Bayesian networks are a rather new development in medicine. Only recently have researchers started to develop such networks, for example, in the areas of oncology (Galan, 2001; Lucas et al, 1998) and infectious disease (Andreassen, 1999; Lucas et al, 2007). There is little experience as yet with integrating ideas from, for example, traditional survival analysis into Bayesian networks. Given the importance of prognostication in health-care, it is to be expected, however, that more prognostic networks will be developed in the near future.

Treatment selection: The formalism of Bayesian networks provides only for capturing a set of random variables and a joint probability distribution over them. A Bayesian network therefore allows only for probabilistic reasoning, as in establishing a diagnosis for a specific patient and in making a prediction of the effects of treatment. For making decisions, as in deciding upon the most appropriate treatment alternative for a specific patient, the network formalism does not provide. Reasoning about treatment alternatives, however, involves reasoning about the effects to be expected from the different alternatives. It thus involves diagnostic reasoning and, even more prominently, prognostic reasoning. To provide for selecting an optimal treatment, a Bayesian network and its associated reasoning algorithms are therefore often embedded in a decision-support system that offers the necessary constructs from decision theory to select an optimal treatment given the predictions

(Andreassen et al, 1999 (a)). Alternatively, the Bayesian network formalism can be extended to include knowledge about decisions and preferences. An example of such an extended formalism is the influence diagram formalism (Shakhter, 1986). Like a Bayesian network, an influence diagram includes an acyclic directed graph. In this graph, the set of nodes is partitioned into a set of probabilistic nodes modeling random variables, a set of decision nodes modeling the various different treatment alternatives, and a value node modeling the preferences involved.

Discovering Functional Interactions: So far we have focused on the use of once constructed Bayesian networks for problem solving in biomedicine and health-care. However, the insight obtained by the construction process itself, in particular when done automatically by using one of the learning methods described above, may also be exploited to solve problems. As the topology of a Bayesian network can be interpreted as a representation of the uncertain interactions among variables, there is a growing interest in bioinformatics to use Bayesian network for the unraveling of molecular mechanisms at the cellular level. For example, finding interactions between genes based on experimentally obtained expression data in microarrays is currently a significant research topic (Friedman et al, 2000). Biological data are often collected over time; the analysis of the temporal patterns may reveal how the variables interact as a function of time. This is a typical task undertaken in molecular biology. Bayesian networks are now also being used for the analysis of such biological time series data (Ramoni et al, 2002).

Support of Clinical Management: According to Lucas (2007), solving health-care problems includes advances in using Bayesian Network technology such as Support of clinical management and discovering regulatory processes.

3.5.2 Related Researches

Different researchers have put their hands on exploring the application of Bayesian network in different real-life problem solving tasks in different domains, including the health sectors. Among many, few but seen to be most relevant (to this particular research) are discussed here.

The paper by Lise et al (2001), which is titled ‘‘Understanding tuberculosis epidemiology using structured statistical models’’, addresses one of the applications of the standard Bayesian network formalism. Statistical relational models are proposed as a means to increase the expressive power of Bayesian networks, and learning the structure and parameters of such models for the exploratory analysis of epidemiological data of patients with tuberculosis was investigated.

In the paper titled ‘‘Using literature and data to learn Bayesian networks as clinical models of ovarian tumors’’, Peter et al (2003) explore the potential of the huge collection of information available on the World Wide Web as prior information for learning Bayesian networks. One of the problems that are often encountered upon learning Bayesian networks for clinical problems is that the available clinical data sets are too small to be exploited. As a consequence, it is usually necessary to extract information from various complementary sources. In this paper, techniques developed in the area of information retrieval are used as a basis for finding relationships among variables from the Web. The applicability of these techniques is studied with the construction of Bayesian networks for the prediction of ovarian tumors in patients.

Another paper, which was done in education sector (and yet, contributed a lot for this research) was that of Rahel (2005) whose research domain was the education sector. In the thesis titled “Computer-Assisted Learner Group Formation Based on Personality Traits”, a mathematical model that can address the group formation problem in cooperative learning, through the mapping of both performance and personality attributes into a student vector space, was developed using Bayesian Network. The researcher reported that the model can serve as a foundation for the application of formal methods in determination of heterogeneous groups based on both performance and personality attributes.

One more paper, titled “Using Bayesian Networks to Manage Uncertainty in Student Modeling”, done by Conati et al (1996), is worth mentioning. In this paper, the researchers described the knowledge structures represented in student models and discussed the implementation of the underlying Bayesian networks. They then demonstrated the possible levels of uncertainty a tutoring system must deal with in making decisions on how to help students and they argued that the Bayesian network representation provides a sound and semantically well-defined way of handling this uncertainty. Using Bayesian networks, a probabilistic student model that provides long-term knowledge assessment, plan recognition, and prediction of students' actions during problem solving was devised. A model that provides assessment of student's understanding and learning during example studying, by monitoring how students read and explain the target examples were also devised.

Having discussed all these, we quit the literature review part. The next Chapter discusses the experimentation part, which is the core element of this research.

CHAPTER FOUR

EXPERIMENTATION

INTRODUCTION

This chapter presents detailed steps undertaken during the experiment. The discussion includes the tasks of data source identification, data collection, data preprocessing, model construction and testing. Critical evaluation of the model is also included.

4.1 Data Source Identification

An important prerequisite to develop Bayesian Network from data is the data itself. This research was based on the data acquired from Hawassa and Hossana Hospitals, two prominent ART service centers under the administration of SNNPG Regional Health Bureau. These health facilities were selected for three main reasons: One- as both of them serve as central sites for most ART clients in the region, they host both urban and rural residents coming from different corners of the region. This has helped the inclusion of marginalized rural ART clients as part of the research. Two- as they are among the pioneers of the service, most historical data were found in the facilities. Since ART adherence is time-dependent, historical aspect of the data was important. Last-unlike many other health facilities, personal records of the ART clients were relatively complete in the aforementioned facilities. This was found to be very important because- as the data are purely in manual format and as the clients having long ART history are limited, it minimized the suffering during collection. Moreover, the quality of record keeping system takes a lion share in the reliability of the results of the research on those records. Because of these reasons, the facilities were choices of the researcher.

In the aforementioned health facilities, detailed personal information of ART clients were filled in **seven** different leaves. The format was imposed by Federal Ministry of Health in 2005, and it has been serving as a standard by all health facilities throughout the country. The first leaf is *Patient Registration Form* (which contains sociodemographic information of the ART client). The second leaf is *Past Medical/ Treatment History Form* (which includes OIs history, Past tests, Negative reactions of the medication, etc). The third leaf is *General Condition/ Physical Exam* (in which, vital signs and symptoms are examined physically). The fourth leaf is *Clinical Review* of the patient. The fifth leaf is *Social Assessment* of the patient (which includes Employment status, Living condition, Confidentiality issues, Spouse and Family conditions, Financial and other concerns). The sixth leaf is *ART Adherence Counselling* guidelines (which includes information of a patient about his/ her Knowledge about HIV and other infectious diseases, Risk behaviours- related to sexual partner, condom using behaviour, addiction to alcohols and drugs, barriers reported by the client, and the general feeling of the client). The last one is *ART Assessment and Plan* (which is directly related to ART eligibility criteria and recommended regimen types). An ART client's personal information includes all these information compiled in one folder.

4.2 Attributes Selection

Although it is the researcher's belief that all the seven leaves mentioned above (Section 4.1) were important when viewed from different angles, only the Sociodemographic, Social Assessment and ART Adherence Counselling parts (Part I, Part V and Part VI) were taken into consideration for this particular research. The intention here was to integrate the "most neglected" social dimensions of ART clients to the "most said" clinical aspects of them, thereby to reach to all-round solutions in relation to adherence. Accordingly, the attribute selection process was focused on the three parts and done very carefully in order not to lose the relevant information there. Domain experts were involved during the selection for more consolation. The attributes selected for the research purpose were depicted in Table 1 below.

Attribute Names	**Modified Attribute Name (if any)**	**Description**	**Selected**
Patient Card Number	PCN	Unique Patient Number	✗
Unique ART Number	UAN	Patient's Unique ART card number	✗
Age	Age	Age of the client when he/she starts ART	✓
Gender	Sex	Sex	✓
Marital status	MaritalStat	Marital status of the client	✓
Level of Education	LevEduc	Education level of the Client	✓
Religion	Religion	Religion of the client	✓
Husband/Wife and Dependent Children at Home	FamilyDependsOn	Whether the client is burdened by responsibility of his/her family	✓
Residence	Residence	Residence of the client	✓
Employment	Employed	Employment status of the client	✓
Disclosure	StatusKnown	Whether the status of the ART client is known (i.e., HIV positivity)	✓
Family Members-Spouse	SpouseStatus	Refers to whether the spouse of the client is HIV positive	✗

Family Members-Children	HasChildren	Whether the client has one or more children	✓
Issues/Concerns Identified	Concerns	Concerns of the ART clients	✗
Understanding of HIV transmission	KnowledgeAbout	Refers to the extent to which the client is knowledgeable about the way of HIV transmission	✓
Has Regular/ Casual sex partner(s)	SexPartner	Refers to the presence/absence of Regular/Casual sex partner	✓
Condom Use	CondomUse	Condom use of the client during last three months	✗
Addictions	Addictions	Refers to whether the client is addicted to tobacco, alcohol, soft drugs or/and hard drugs	✓
Adherence: Concerns/Barriers to ART	Barriers	Refers to different barriers that may hinder the client's adherence	✗
Lost Job due to Current illness	LostJob	Refers to whether a client has lost his/her job due to the current illness	✓
	Adherence	**Refers to whether the client is on the regimen or dropped it out (Target Class)**	✓

Table 1: List of Selected Attributes and their Converted forms

4.3 Data Collection

After the identification of data source and selection of relevant features (attributes) of the data, the next step was collecting the data. As mentioned earlier, the data used for this purpose was available in manual format. Therefore, attributes selected should have been collected and converted into an electronic format. The collection process was one of the most difficult tasks for five main reasons: First- ART patients' records were mixed with other patients' records (usually mixed with MCH and OPD records); therefore each of them must have been searched by patient card number (PCN). Second- beyond the seven leaves which can be taken as De Facto, different papers of clinical reports, follow-up notices, transfer-in documents, etc made the individual records too huge to handle without losing the slices. As the data of interest were most historical, the Size-Time trade-off was the problem. Third- some attributes of interest for this research needed conversion (deriving another attribute) on-spot before they were converted to electronic forms. Because of this reason the researcher must have gone through each record without the help of others. Fourth- the target class (whether the patient is on the regimen or stopped it) was handled in separate log book called "*ART Patient Register*". To get such information, both *Unique ART Number* and *Patient Card Number* were needed because some of the records on the *Register* have only one of the two identifiers and some of them contain both of them. Fifth- some healthcare workers (perhaps, physicians or nurses) rush to the clinical aspects of the clients without having detail information about important non-clinical aspects of the person. Consequently, it was not uncommon to see even empty slices except the name of the patient in the intended pages.

Despite all these challenges, the data with their selected attributes were collected with a maximum care that can be taken. The collected data were entered in Microsoft Excel format to take advantage of easier data manipulation and further ease of accessing the data from BN PowerSoft interface which was used as platform for model building purpose. Accordingly, a total of 1561 data were collected from the two hospitals (792 data from Hawassa Hospital and the remaining 769 from Hossana Hospital) and filled in Excel format. The following table shows distribution of the data with respect to the target class.

Adherence	Number of Records	Records in Percent (%)
ADHERE	1105	70.8
DROP	456	29.2
TOTAL	1561	100

Table 2: Data Distribution w.r.t. Target Class

Two important points were well thought-out during data collection:

1. Relatively most historical data were collected and most recent data were avoided. The reason for this was the assumption that adherence might wane with time. In view of that, the research excluded records of ART clients who started their regimen after January 2007.
2. The research excluded paediatric data as it doesn't contribute for the goal of the research. Consequently, the minimum age of the clients was 16. The records of this kind are usually called as "records of adolescents and adults".
3. Except for reasons mentioned in 1 and 2, and for some vague, inconsistent and incomplete records, all the data in both facilities were collected.

4.4 Data Preprocessing

Once the data were collected and converted into an electronic format, the next vital step to undertake was data preprocessing. Important steps in data preprocessing include cleaning, integration, transformation and reduction. Preprocessing was undertaken with the goal of enhancing the excellence of the data thereby helping to improve the effectiveness of the machine learning process. Some important preprocessing tasks performed under different levels that need detail discussions are explained as follows.

4.4.1 Data Cleaning

The reliability of output of the machine learning procedure highly depends on cleanness of the data. Three data cleaning tasks undertaken in this research were *filling the missing values, removing inconsistencies,* and *removing outliers.*

Some missing values were recognized in the collected data. The reason was data-recording errors during the collection. Such data were corrected immediately by tracing them with their *Patient Card Number (PCN)* and *Unique ART Number (UAN)*. On the other hand, for conditions like a person has "*lost his/her job due to current illness*", the *"Employed"* column was automatically filled as "*NO*" in cases when the field was not filled out.

Inconsistencies within the data were also avoided during data cleaning. For example, a person who was "*working full time*" should not have reported as "*Lost job due to current illness*" at the same time. To be safe in such controversial situations, the data were avoided completely.

One record was observed with a client's age 116. Since there were no other values between ages 75 and 116, the record was considered as an outlier and thus it was removed. Accordingly, the maximum age observed was 75.

4.4.2 Data Integration

Data from two different sources (Hossana and Hawassa Hospitals) were integrated without any difficulty because both ART centers use same format which is the standard of the Ministry of Health (MoH).

4.4.3 Data Transformation

Another important task undertaken under data preprocessing was data transformation. Two data transformation techniques, *data generalization* and *new attribute derivation* techniques were applied in this research. For some attributes, since some details were found to be not as such helpful for the learning purpose, they were transformed into their most general forms. For some others, new attributes were derived from the existing ones. They are discussed as follows.

"Patient Address" in Patient Registration Form refers to the residence of the patient. Health Facilities capture these information (Woreda/ Kifle-Ketema/ Peasant Association, Kebele, House Number, Telephone number) for the purpose of effective follow-up. But these details are irrelevant for Bayesian Network learning process. Possibly, what is important is the person's residence category- Rural or Urban. Therefore, the clients were categorized simply as *rural* or *urban*.

Another attribute, "*Employment*", contains four values: *Working Full time, Working Part-time, Not Working due to ill health* and *Unemployed.* Again this attribute was modified as *Employed,* and its values were assigned simply to be "*Yes*" if a person is working full-time or part-time, or "*No*" otherwise.

Yet another attribute, "Family Members-Children" details the number of children alive/died. For most cases, however, only the presence or absence of children was reported and not other details. Here, from the researcher's point of view, the presence/absence of children was more important (and in fact reliable) as compared to the partially-filled "*number of children*" for Bayesian learning. Therefore, this field was renamed as "*HaveChildren*" and the values for the attribute became "*Yes*" if the client has at least one child, or "*No*" if not.

The "*Addiction*" attribute, whose values include Tobacco addiction, Alcohol addiction, Soft drugs addiction and Hard drugs addiction (and the degree of addiction to each cases) was found to be difficult to make use of it directly. So, the researcher took the information only on being addicted or not. Subsequently, the value was "*Yes*" if a person was addicted of at least one of the aforementioned cases or "No" if he/she was not.

"*Understanding of HIV Diseases*" was another field with five choices, whose intention was to measure the extent to which an ART patient is knowledgeable about the HIV disease. Again, such subjective measures may affect the accuracy of BN model. After a thorough discussion with the domain experts, the attribute was renamed as "*KnowledgeAbout*" and the possible values were assigned to be either "*Good*" or "*Poor*". A person with no or very little knowledge about HIV was assigned as "Poor". Clients with sufficient knowledge about the disease were grouped as "Good".

Two fields, "*Has Regular Sexual Partner*" and "*Has casual Sexual Partner*", were merged to a single field, "*SexPartner*". Here again, casuality was more important than the number of casual partners the person has had. With this assumption, the values of the newly derived attribute were "*Regular*" or "*Casual*".

4.4.4 Data Reduction

A technique of data reduction is applied on data to obtain a reduced representation of the dataset that is much smaller in volume, yet closely maintains the integrity if the original data. Among different strategies of data reduction, *dimensions reduction* and *discretization* techniques were implemented in this research.

Dimension reduction refers to the process of identification and removal of irrelevant, weakly relevant or redundant attributes from the data. For this particular case, "*Patient Card Number (PCN)*" and "*Unique ART Number (UAN)*" were immediately removed after they had served as search keys for target class (adherence condition of a person). The attribute "Concerns" was also removed for two main reasons. First- among its 9 alternative values, some of them are overlapping concepts. For example, a person may reflect his/her concern as "Financial Problem" or "*Dietary Problem*". Here, the first highly affects the second. This can possibly create inconsistency in data. Second, the values of the attribute can be taken as indirect effects of other attributes (such as Employment, Lost-Job, Family-Dependence, etc). Hence its omission. The "*Barriers*" attribute was removed for the same reason. Another attribute, "*Spouse-Status*" was removed since most of the ART clients reported that they "do not know" their spouse's (or, sex partner's) status. It seems this attribute was carelessly filled. Moreover, several clients haven't been asked this question (not ticked at all). Though the attribute might be important, it was removed for the sake of ensuring consistency. One more attribute, "*UseCondom*" was dropped for the very reason that almost all the patients reported that they either use condoms rarely or not at all. Both cases can be taken as poor condom usage trend. In view of this, almost all ART clients have similar value ("No"). Therefore, the attribute has no contribution on BN modelling and hence it was rejected.

After removal of the abovementioned attributes, a total of 15 attributes (including the target class) were held for the model building purpose. (See Table 1)

Another data reduction technique applied in this research was discretization. The collected and cleaned data attributes here constitute discrete (categorical) values except the *Age* attribute whose values were numerical. Since the tool selected to conduct the research was a Bayesian Network tool which accepts only discrete values, the *Age* attribute was discretized into three discrete classes. The task was performed by BN PowerSoft's pre-processor, using equi-width discretization technique. The following table shows the actual values and their corresponding transformations for the *Age* attribute. The research deals with the ART adherence trends of adolescents and adults only; therefore, the minimum age observed was 16 and the maximum age was 75.

Age Range	**Transformed Value**
>=16 and <= 35	YOUNG **(?)**
>35 and <= 55	SENIOR **(?)**
>55	OLD **(?)**

Table 3: Discretization of the *Age* Attribute using BN PreProcessor

Yet, domain experts were not convinced on such discretization results because the technique doesn't consider the qualitative aspect of **Age** attribute, thereby, Age-based inferential results will be vague. To avoid this, the manual discretization technique was recommended by the experts. In this regard, clients were categorized as **Young**, **Middle_Age**, **Senior,** and **Old** based on the recommendations in different literature's[H]. Accordingly, the Age attribute was discretized as follows:

Age Range	**Transformed Value**
>=16 and <= 29	YOUNG
>30 and <= 55	MIDDLE_AGE
>56 and <=75	SENIOR
>75	OLD[I]

Table 4: Discretization of the Age Attribute Manually

The later case was found to be more appropriate. Therefore, the results of this research were based on the manually discretized Age ranges.

4.5 Bayesian Network Modeling Technique

This research employed Three-Phase-Dependency-Analysis **(TPDA)** algorithm**.** The algorithm uses mutual information calculation as the quantitative CI test to avoid the

[H] Age discretization is arguable. The *age range* for the *youth* policy (according to Assessment of *Youth* Reproductive Health Programs in *Ethiopia* [URL: www.fhi.org/NR/rdonlyres/.../EthiopiaAssessRptenyt.pdf *])* is 15-29. But other age ranges were not standardized. Yet, most literatures including “Adolescents & Young Adults Committee” URL: http://www.curesearch.org/our_research/index_sub.aspx?id=1774 agree with the discretization made here.

exponential complexity on CI tests. It takes a database as input and constructs the belief network structure as output. The algorithm also assumes that the database attributes have discrete values and there are no missing values in all the records. TPDA learning algorithm was discussed in Chapter 3 of this research.

The platform used to design, develop and train the model was **Belief Network PowerSoft**. The package in Belief Network PowerSoft includes three applications: **BN PowerConstructor, BN PowerPredictor, and Data PreProcessor.**

Data **PreProcessor** is a simple tool to be used with PowerConstructor and PowerPredictor. It has three functions: discretize data, convert other data formats to *.MDB (MS-Access) format, and partition the training data into **internal training data and internal test data**.

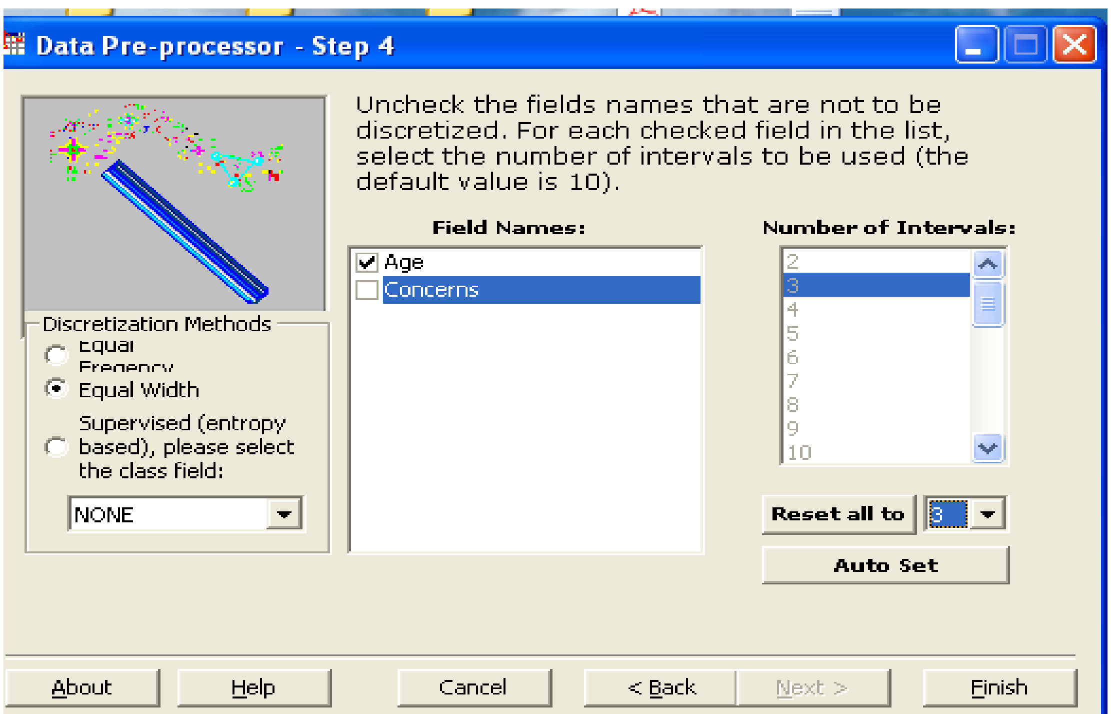


Figure 3: Visualization of Data PreProcessor

[1] As a matter of chance, no Old-aged client was included in the research as the maximum age was 75

PowerConstructor is designed based on three-phase belief network (BN) construction algorithms. It includes a wizard-like user interface and a belief network construction engine. Data set to be used with PowerConstructor, the following conditions must be satisfied: One- if the data set contains continuous fields, Data PreProcessor should be used to discretize it. Two- if the data set is in other formats, we can either choose to convert it to *.MDB format or use the original data. PowerConstructor runs faster on *.MDB data format. The data set does not need to be partitioned.

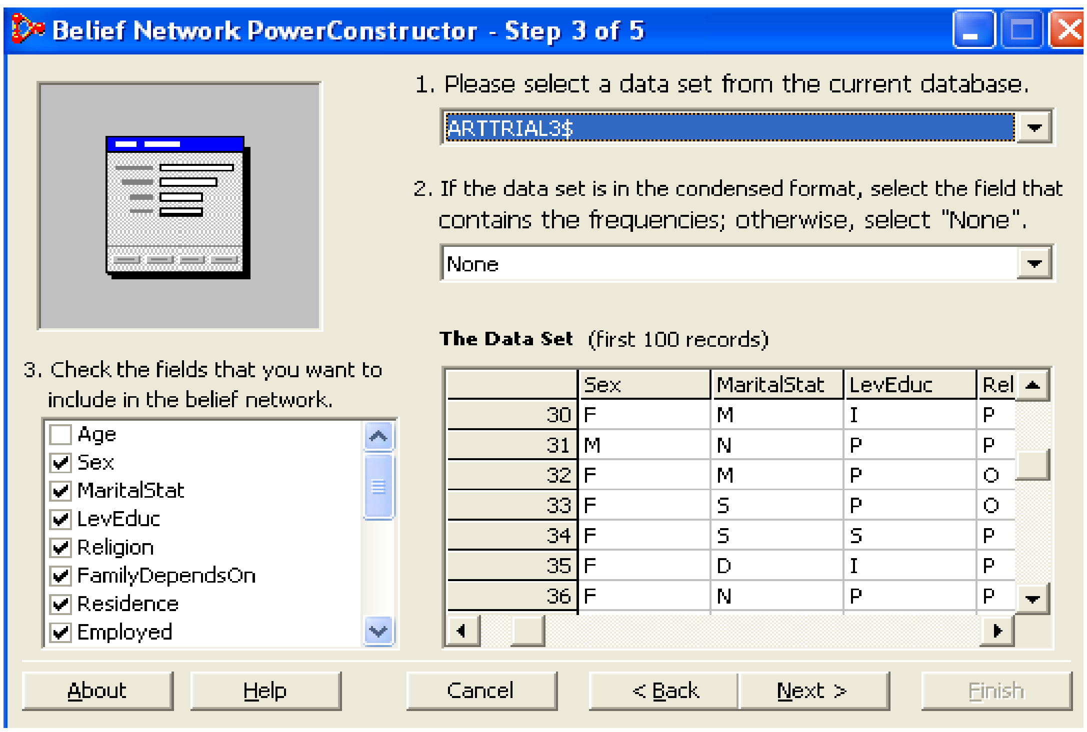


Figure 4: Visualization of BN PowerConstructor

PowerPredictor system is an extension of BN learning system (BN PowerConstructor) to BN based classifier learning and using. It can learn general Bayesian network classifiers and Bayes' multi-net classifiers from training data and use these classifiers to classify new data. The system can also perform feature subset selection automatically.

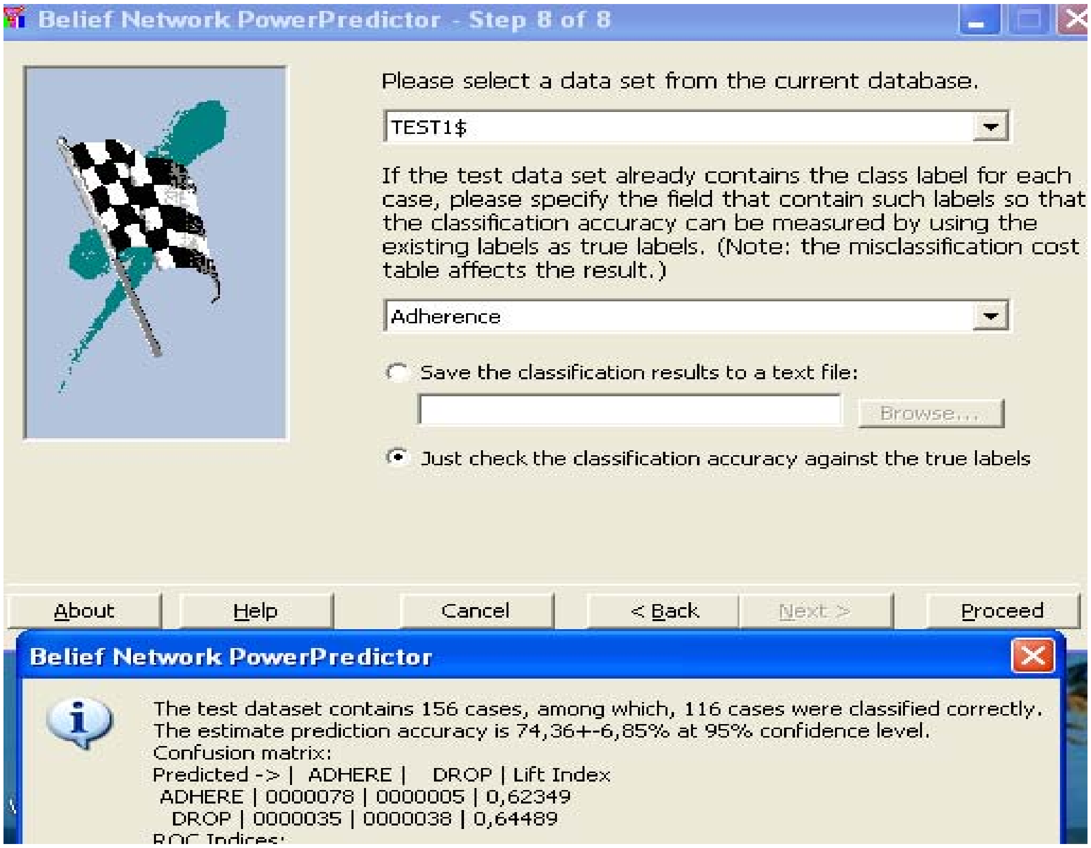


Figure 5: Visualization of PowerPredictor

PowerPredictor is suited to learn a new general BN classifier or Bayesian multi-net classifier (with or without auto feature subset selection) from training data, to modify an existing BN based classifier, and to use an existing BN based classifier to classify new data.

4.6 Model Test Design

Estimating predictor accuracy helps to evaluate how accurately the predictor will label the data on which the predictor has not been trained. Among different methods to evaluate classifier/predictor accuracy, Holdout and Cross-Validation techniques take the lead. The two techniques are based on randomly sampled partitions of the given data. For this particular research, **cross-validation** method was used to evaluate predictive accuracy during training.

4.7 Building Bayesian Network Model

The belief network modeling software employed for the purpose of this experiment was the Bayesian **Belief Network PowerSoft** software package.

Since results might be influenced by the selection of the test and training datasets, experiments were carried out by splitting the data into 10 partitions, i.e., a percentage split (10-fold)[J] was used to partition the dataset into training and test data. Each partition, in turn, was used for testing while the remainder was used for training. This process was repeated ten times for the learning algorithm and, at the end, every instance was used exactly once for testing. Finally, the average result of the 10-fold cross validation was considered.

4.8 Training and Testing the Bayesian Network (TPDA)

4.8.1 Experiment I: TPDA without Elicitation of Expert Knowledge

For the TPDA algorithm, the training data sets were prepared as tables in Microsoft Access Database. The data were learned without node ordering (i.e., without elicitation of domain-expert knowledge). Using 10-fold cross-validation method, all the 10 partitions were tested for their prediction accuracy. The prediction accuracy and respective confusion matrix for each of the three best partitions (First, Second and Fourth partitions) are discussed as follows:

J The advantage of cross-validation method over repeated random sub-sampling is that all observations are used for both training and testing, and each observation is used for testing exactly once. 10-fold cross-validation is commonly used (Wikipedia, 2008)
According to Fung and Crawford (1990), cross-validation technique can avoid overfitting to the data when the dataset is not large enough.

For the first partition, the test set contains 1405 cases, among which 1025 cases were classified correctly. The estimate prediction accuracy was 72.95% at 95% confidence level. Its confusion matrix is as follows.

	PREDICTED		
ACTUAL	ADHERE	DROP	TOTAL
ADHERE	984	62	1046
DROP	318	41	359

Table 5: Confusion Matrix for the First Partition

For the second partition, the test set contains 1405 cases, among which 1026 cases were classified correctly. The estimate prediction accuracy was 73.02% at 95% confidence level. Its confusion matrix is as follows.

	PREDICTED		
ACTUAL	ADHERE	DROP	TOTAL
ADHERE	988	78	1066
DROP	301	38	339

Table 6: Confusion Matrix for the Second Partition

Similarly, for the Fourth partition, the test set contains 1405 cases, among which 1028 cases were classified correctly. The estimate prediction accuracy was 73.17% at 95% confidence level. Its confusion matrix is as follows.

	PREDICTED		
ACTUAL	ADHERE	DROP	TOTAL
ADHERE	975	81	1056
DROP	296	53	349

Table 7: Confusion Matrix for the Fourth Partition

In the same way, the estimate prediction accuracy of the third, fifth, sixth, seventh, eighth, ninth, and tenth partitions were 72.60%, 72.88%, 72.60%, 72.74%, 72.81%, 72.74%, and 72.10%, respectively, at 95% confidence level. Their respective confusion matrices are shown in **Appendix 1** of this research. The following Table summarizes the prediction accuracies and Risky predictions [K] for each case obtained from the 10-fold cross validation:

TEST#	1	2	3	4	5	6	7	8	9	10	Avg.
Pred. Acc. (%)	72.9	73.02	72.60	73.17	72.88	72.60	72.74	72.81	72.74	72.10	**72.80**
Risky Classif (%)	22.63	21.4	22.6	21.07	20.56	20.85	21.63	20.21	20.0	20.99	**21.20**

Table 8: Prediction Accuracies for All Partitions of 10-Fold Cross-Validation (TPDA)

From Table 7 it can be understood that the learning process was consistent within the partitions. The consistency in the test sets has something to do with both the quality of the data and the predictive performance of the BN model. The figure depicted below (Figure 4.4) illustrates the best learned network among the 10-fold experiment.

According to the model in Experiment I, the adherence to the medication of an ART client is directly affected by two factors: **Addiction** (drug or alcoholic behaviour) and **loss of job due to current illness**.

The average predictive accuracy of the model is **72.8%.** This indicates that the model's performance is promising.

[K] **Risky Prediction** here refers to predicting an instance as "Adherent" whereas the instance actually is "Dropper".

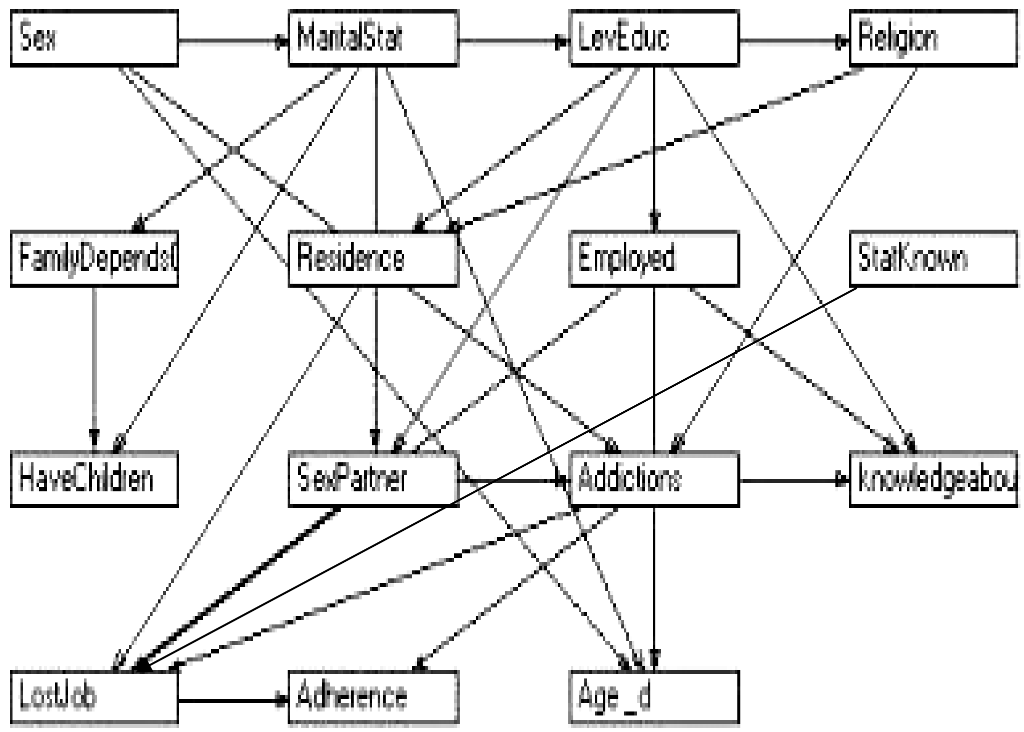


Figure 6: The BN for Best Learned Model in TPDA without Expert Involvement

On the other hand, the average Risky prediction from the partitions is **21.2%.** In other words, 21.2% of the instances were predicted as "**ADHERENTS**" whereas they actually are under the category of "**DROPPERS**". Though the results were encouraging, it was observed that there is a need to enhance the predictive accuracy even more so as to reduce such a risky prediction.

For more enhancement of predictive accuracy of the BN model, domain experts were consulted. In what follows, the **TPDA** based on domain experts' beliefs is presented.

4.8.2 Experiment II: Modifying the Structure of the Model Based on Expert Knowledge

Depending on secondary data alone may have its own drawbacks. As it is learnt from sub-population of ART clients, links might have been established between attributes which are independent in the general population; links may exist where direction should have been opposite from what appears in the network; links may exist where attributes are not directly related; an expert may also expect a variable to have several more parents than actually appearing on the network. This means automatic learning methods alone may not be sufficient. An option considered under such circumstances, was to reinforce the learning using the knowledge of the human/domain expert (Rahel, 2005) [(L)].

In this step, therefore, domain information was elicited from domain experts. The elicitation technique used here was structured technique which is based on a confirmatory approach to data elicitation. In this technique, experts were provided with a list of predefined concepts and were asked to specify the direction and sign between the concepts.

The previously learned model, which was purely based on mutual information concept, was given to domain experts so that they could suggest what they feel. Moreover, they were given a task of rearranging the nodes according to which attribute affects the other (based on their beliefs). The best model found in Experiment-I was modified by the experts as follows:

[L] Here the idea of Rahel (2005) was customized for current work

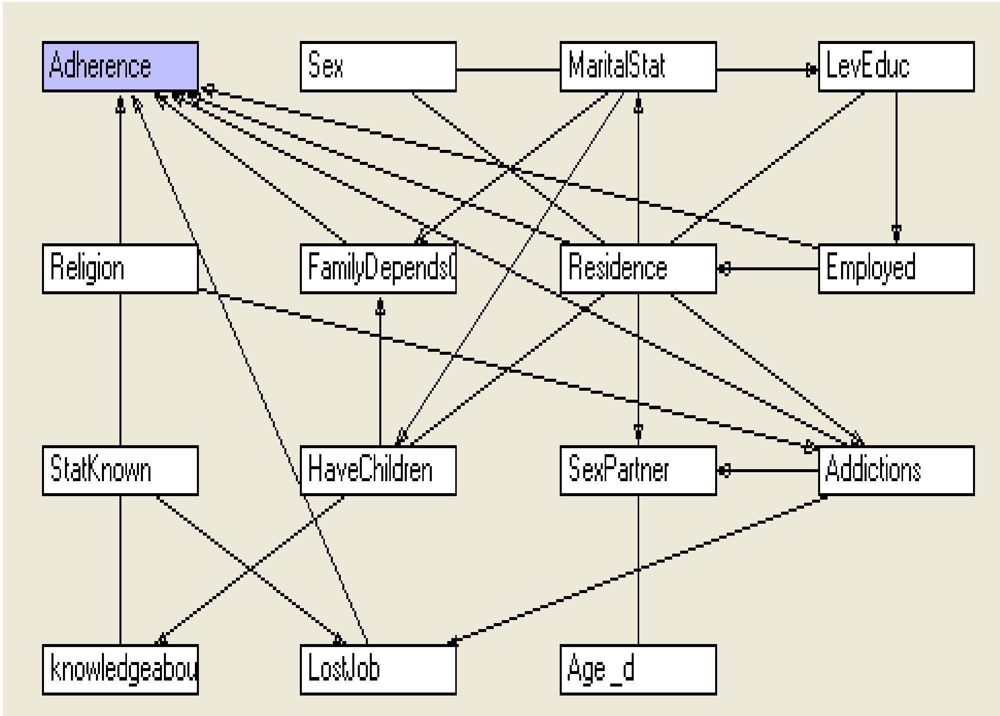


Figure 7: Expert-guided Bayesian Network

Some of the major changes made in the modified network as compared to the original learned networks are the following:

- In the original network we observed that ART adherence had only two parents (*Addiction* & *LostJob*) whereas 5 more attributes are included to have direct influence over adherence.
- In the original learnt network, we have *SexPartner* directly affecting *Addiction*, whereas the reverse is true for the modified case. Similar adjustment was done between the *HaveChildren* and *FamilyDependsOn.*
- *LevEduc* in the previous model was observed to affect four attributes (*Residence, SexPartner, Employed,* and *KnowledgeAbout*). But in the modified model, it was seen to affect only two them (*Employed,* and *KnowledgeAbout*); and, so on.

It took a long-term dialogue among the experts and the researcher in order to come to consensus. The basis for domain experts to come to such a final network was discussed as follows. According to the domain experts:

- Knowledge of HIV and other opportunistic infections has a lot to do with someone's ART adherence. Knowledge avoids negligence; it makes the clients serious and responsible to what they do; it supports them even in their "hard times" not to dropout the treatment; it also increases their commitment of adherence to the intended level. Therefore with some exceptions, it was observed that clients with strong knowledge of HIV and ART adhere better.
- Employment status of a person matters a lot in his/her adherence especially in resource-limited settings. By its very nature, ART needs balanced diet in parallel. Dietary issues again have something to do with financial background of a person. In real practice, employed clients have better financial backgrounds than the unemployed ones. Therefore, the conclusion.
- As the viral load increases, it will put an HIV patient in an ill-health situation. That may create frustration on employers (say, private institutions, NGOs, or even Public sectors) thereby, an ART client may lose his/her job. This again affects the financial status of the person. Moreover, such people usually face psychological & social crisis and likely poorly-adhered.
- People who are addicted of alcohols and drugs are likely less adhered. The reason may be their hopelessness, financial crisis or forgetfulness- which are common on several ART clients.

- ❖ In general, people living in rural villages are likely poorly adhered as compared to the ones living in urban areas. As can be predicted, the reason behind this may be lack of awareness, financial problems or lack of infrastructure such as transportation.
- ❖ An HIV burden will be even worsened if an ART client has family members who depend on him/her. Since he/she needs to support them, commitment of adherence will decrease. Consequently, the adherence records of burdened clients are not satisfactory.

In addition to these, the domain experts argue in the following points. The final consensus was the following:

- ❖ Educational background has a vital role in adherence as it increases awareness/ knowledge of the client. But, literally speaking, it has no direct effect on a client's adherence. In practice, educational background of the person is not directly responsible for his/her poor adherence records
- ❖ Sociodemographic factors such as *Age, Sex, Marital Status* and *Religion* have no direct effect on adherence. But it is believed that they are inputs of the major factors for adherence. For example, if a person is married, he/she is likely burdened of others (spouse or children). Again, people in some religions are abstained from alcohols and drugs. That may help adherence of those groups.
- ❖ Disclosure of the status of a person may affect the person's adherence. But its link to adherence is not direct. Rather, disclosure of status can highly risk the job grant of the person in this context. Loss of job can directly affect adherence (as already discussed above).

As was discussed very briefly, all the fourteen attributes used previously were found to be relevant in building Bayesian Network model. Accordingly, the same attributes were used for training and testing purpose after modification. The steps followed to reach to the prediction were the following:

1. Same dataset used in Experiment-I was trained and tested in each partition of Experiment-II.
2. Via BN Editor of the Belief Network PowerPredictor, the expert knowledge was elicited to the system (node ordering was implemented), and *saved.* Automatically, the Bayesian Network was modified according to the new knowledge.
3. The learning process was undertaken according to the modification.

The following BNJ visualization shows the modified model.

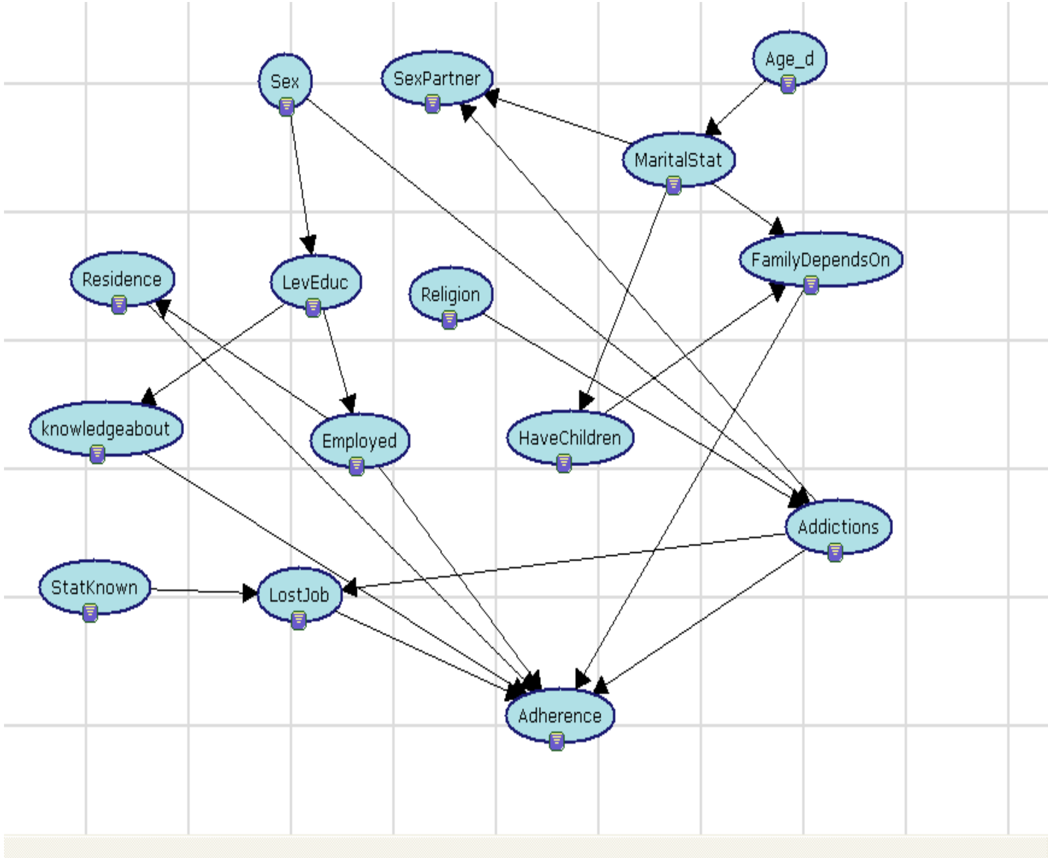


Figure 8: BNJ Visualization of Expert-guided Bayesian Network

Following the aforementioned steps, the prediction results were improved significantly for all partitions of the 10-fold cross-validation. The prediction accuracy and respective confusion matrix for each of the three best partitions (Third, Fourth, and Fifth partitions) are discussed as follows:

For the third partition, the test set contains 1405 cases, among which 1080 cases were classified correctly. The estimate prediction accuracy was 76.9% at 95% confidence level. Its confusion matrix is as follows.

	PREDICTED		
ACTUAL	ADHERE	DROP	TOTAL
ADHERE	927	33	960
DROP	292	153	445

Table 9: Confusion Matrix for the Third Partition

For the Fourth partition, the test set contains 1405 cases, among which 1080 cases were classified correctly. The estimate prediction accuracy was 76.9% at 95% confidence level. Its confusion matrix is as follows.

	PREDICTED		
ACTUAL	ADHERE	DROP	TOTAL
ADHERE	939	37	976
DROP	288	141	429

Table 10: Confusion Matrix for the Fourth Partition

Similarly, for the Fifth partition, the test set contains 1405 cases, among which 1084 cases were classified correctly. The estimate prediction accuracy was 77.20% at 95% confidence level. Its confusion matrix is as follows.

	PREDICTED		
ACTUAL	ADHERE	DROP	TOTAL
ADHERE	926	38	964
DROP	283	158	441

Table 11: Confusion Matrix for the Fifth Partition

Likewise, the estimate prediction accuracy of the first, second, sixth, seventh, eighth, ninth and tenth partitions were 75.10%, 76.60%, 76.70%, 75.40%, 75.40%, 74.80%, and 74.2%, respectively, at 95% confidence level. Their respective confusion matrices are shown in

Appendix 2 of this research. The following Table summarizes the prediction accuracies and the Risky predictions for each case obtained from the 10-fold cross validation:

TEST#	1	2	3	4	5	6	7	8	9	10	Avg.
Pred. Acc. (%)	75.1	76.6	76.9	76.9	**77.2**	76.7	75.4	75.4	74.8	74.2	**75.9**
Risky Classif. (%)	18.3	20.1	20.8	20.5	20.1	20.1	20.4	20.6	19.5	21.3	**20.1**

Table 12: Prediction Accuracies for All Partitions of 10-Fold Cross-Validation after Modification of the BN

By and large, the predictive performance of the model was improved at the elicitation of domain expert knowledge. The following section evaluates the overall performance of the BN model.

4.9 Evaluation of the Model

i. Predictive Performance

The following Table summarizes the predictive performance of Bayesian Network model before and after the elicitation of expert knowledge. The percentage of Risky classification is also presented.

	Experiment-I	**Experiment-II**
Best Prediction	73.17%	77.2%
Worst Prediction	72.10%	74.2%
Average Predictive Accuracy	72.8%	75.9%
Risky Classification (on Average)	21.2%	20.1%

Table 13: Revised form for Performance of BN Model

The predictive performance results obtained from Experiment-I and Experiment-II can be evaluated from the following vital points of view:

- The average predictive accuracy was significantly improved (from **72.8%** to **75.9%**) on knowledge elicitation**.** This can tell us the importance of expert involvement during modelling.
- The **percentage of Risky prediction** (predicting the actual "Droppers" as "Adherents") has also decreased from **21.2%** (on average) to **20.1% %** (on average)**.** This is a very crucial point as the risk of misclassification was reduced thereby.

- Significant enhancements in prediction and reduction in error rates in the modified model can be taken as the indication of the **value of a domain experts' intervention** during model building.
- The best prediction for the model without an expert intervention (Experiment -I) was the one whose predictive accuracy was **73.17%,** and the worst was the one with **72.10%** predictive accuracy at 95% confidence level. This can tell us that the predictive results obtained are less inconsistent among the partitions of 10-fold cross-validation.
- According to Experiment II on the other hand, the best prediction was the one with predictive accuracy **77.2%**, and the worst was the one with **74.2%** at 95% confidence level. This can tell us that the predictive results obtained are again less inconsistent among the partitions.
- In line with the TPDA model without an expert intervention, the adherence to the medication of an ART client is directly affected by two factors: **Addiction** (drug or alcoholic behaviour) and **loss of job due to ill health**. According to the later model (Experiment-II), on the other hand, adherence of an ART client is directly affected by six factors: **Addiction** (drug or alcoholic behaviour) and **loss of job due to current illness**, **Residence of a patient**, **Knowledge the client has concerning HIV**, **Employment status of the client**, and **Family dependence (independence).**
- As was seen, the (average) predictive accuracy of the later model was closer but better than that of the previous one. Moreover, the number of (directly) influencing factors was grown from **2 to 6**. It was also observed that the modified model didn't ignore the two directly influencing factors observed in the previous model (***Addiction*** and ***LostJob***) in its prediction. From these points, the following conclusions were made:

- Bayesian Network is a powerful predictor even in the absence of a domain expert
- With elicitation of domain experts' knowledge, Bayesian network can perform even better.
- There was a constructive relationship between the two phases (Experiment-I and Experiment-II); yet, enhancements of results were observed as domain experts get involved. As such, the domain experts are basically needed for betterment of the results.

ii. Instant Classification

BN PowerPredictor has a facility for creating an instant classification. All the attributes that directly influence the target class appear in the instant classifier window if built properly. Accordingly, all the six attributes affecting adherence directly were appeared in the window as depicted in the figures below. The visualisations show two instances of the instant classifier that takes cases as input and gets class label instantly. Figure 9 shows one of the instances for adherence ("Adhere") and Figure 10 shows an instance for dropout.

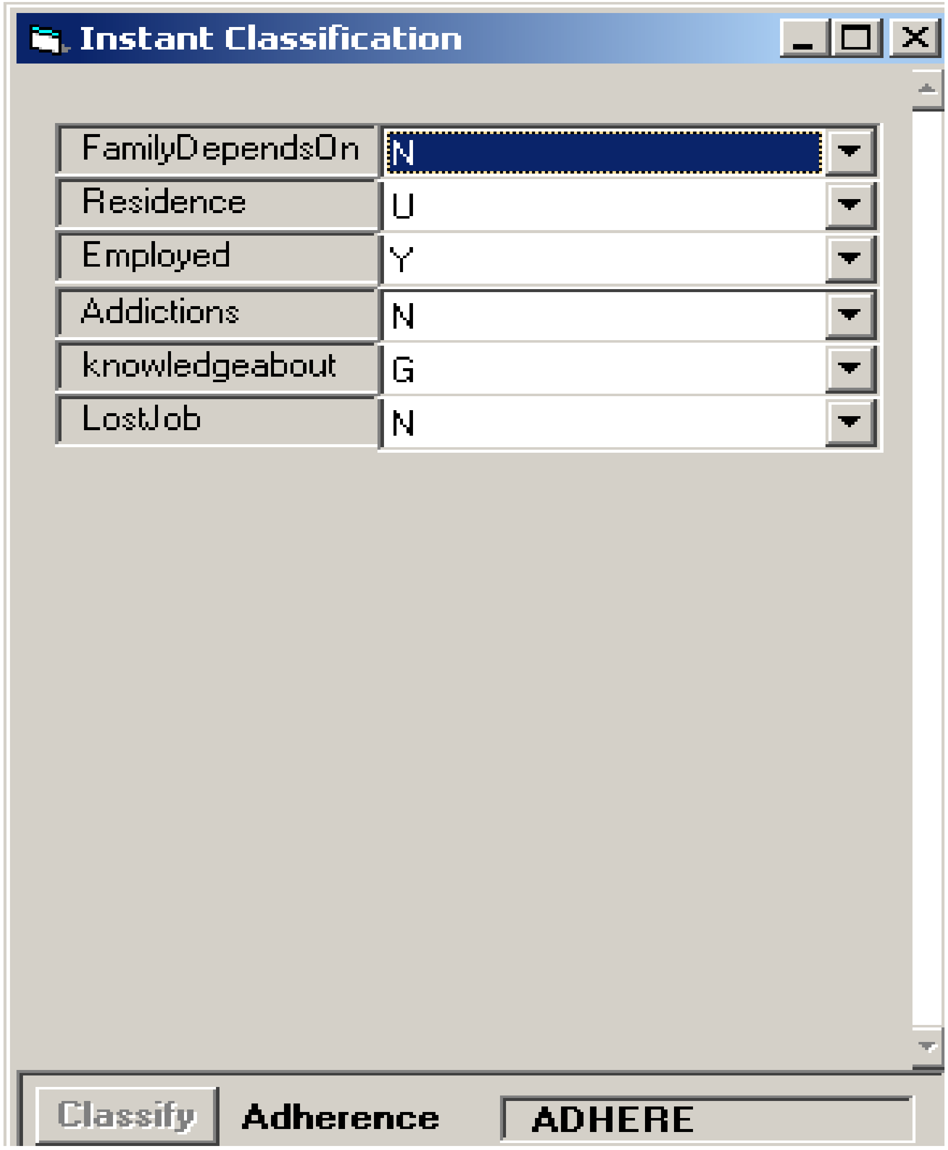


Figure 9: Visualization of an Instant Classification for "ADHERE"

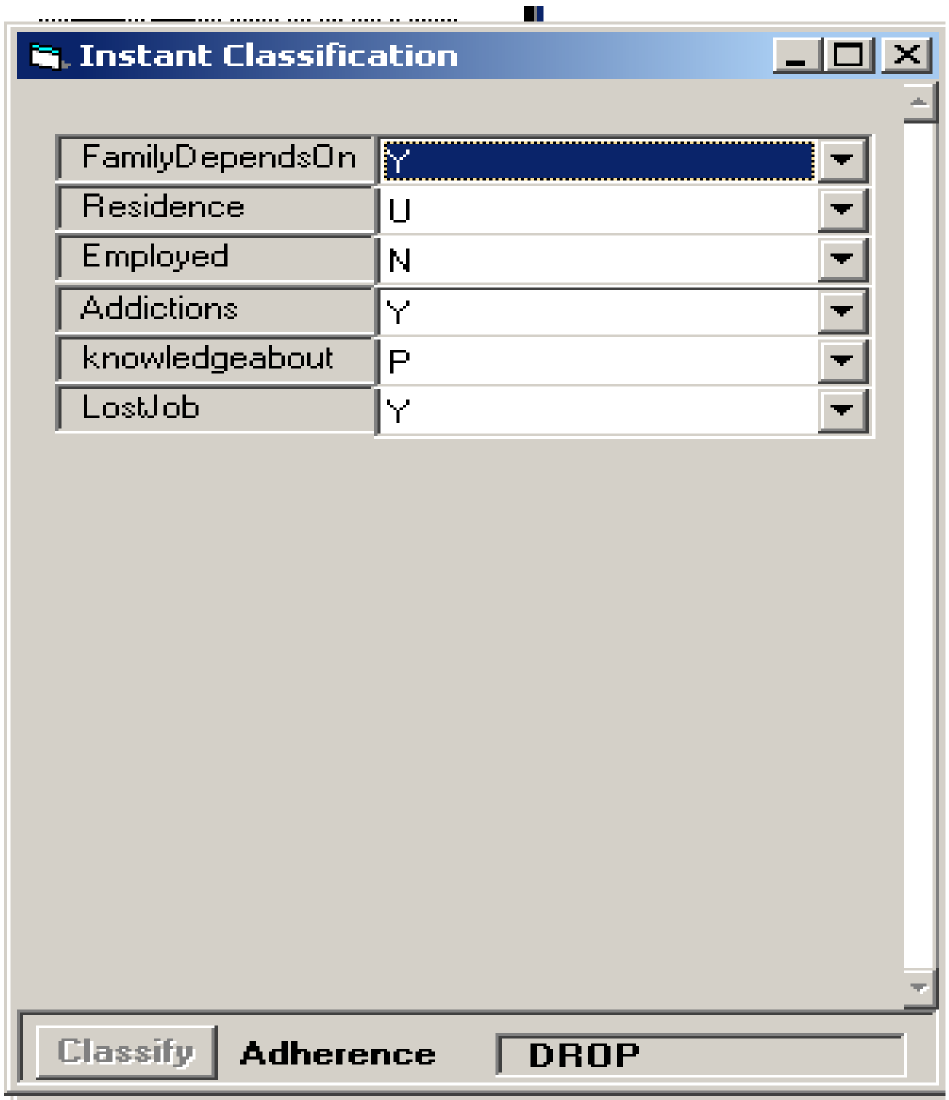


Figure 10: Visualization of an Instant Classification for "DROP"

The instant classifications of such types can be extended further for full-fledged version of an expert system. By doing so, an expert (say, adherence counsellor) can make decisions on what to do immediately, even during client's registration. This aspect is central, as ART adherence counsellors need supporting information for proactive measures.

To this end, the BN model developed can guide further efforts for full-fledged expert system based on the attributes. Moreover, such an analysis can serve to provide insight in evaluating the robustness of the output of the network to possible inaccuracies in the underlying probability distribution.

iii. Conditional Probability Table (CPT)

CPT is another method to evaluate the BN model, especially when we want to check for its robustness. Bayesian Network Tools with Java (BNJ) was employed to visualize the CPT of the BN model developed. To do that, the following steps were carried out:

- The structure learned by the PowerConstructor in BN PowerSoft was transported to BNJ by saving it as ***.net** file.
- The ***.net** file was opened with BNJ.
- The Prior and posterior probabilities for all nodes (on demand) were generated as follows:

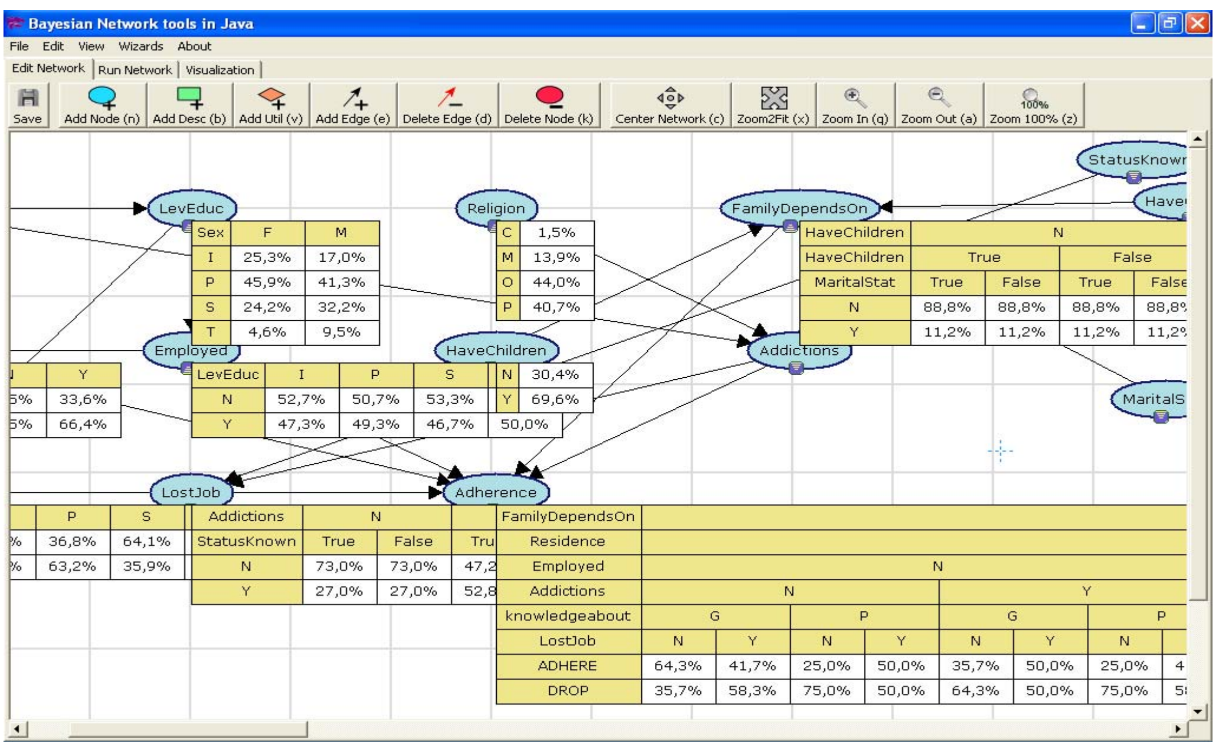


Figure 11: CPT Visualization for All Nodes

In particular, the interest lies on CPT results related to the target class. Therefore, a CPT was generated by BNJ for the six attributes that affect adherence directly, as follows:

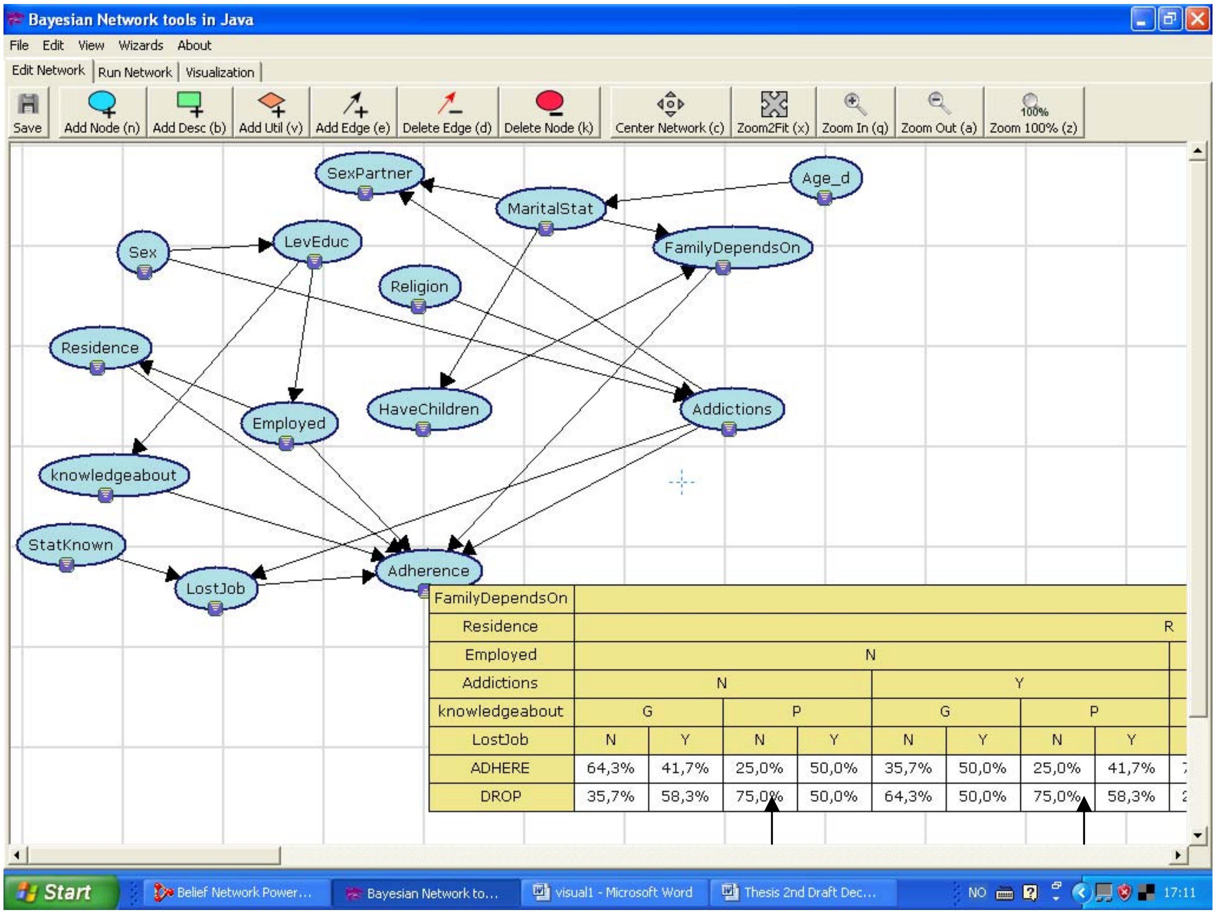


Figure 12: CPT Visualization for the Target Node-Adherence

From the Figure depicted above, we get the joint probability distribution for all variables in the network. Based on CI principle:

P(Adherence | FamilyDependsOn, Residence, Employed, Addiction, KnowledgeAbout, LostJob, StatKnown, HaveChildren, Sex, LevEduc, Religion, SexPartner, MaritalStat, Age_d) =

P(Adherence | FamilyDependsOn, Residence, Employed, Addiction, KnowledgeAbout, LostJob)

The joint probabilities and semantics of some instances are discussed as follows:

A. **P(Adherence = Drop | FamilyDependsOn=No, Residence=Rural, Employed= No, Addiction=No, KnowledgeAbout=Poor, LostJob=No) = 0.75.**

⇒ The probability of poor adherence for an unemployed, rural ART client whose knowledge about HIV is poor- is as high as 0.75 even though he/she is not addicted, not burdened of family responsibility and has not lost his/her job because of ill-health problems.

B. **P(Adherence = Drop | FamilyDependsOn=No, Residence=Rural, Employed= No, Addiction=Yes, KnowledgeAbout=Poor, LostJob=No) = 0.75.**

⇒ The probability of poor adherence for an unemployed, rural, Addicted ART client whose knowledge about HIV is poor- is as high as 0.75 even though he/she is not burdened of family responsibility and has not lost his/her job because of ill-health problems.

Combining A and B can conclude the following:

⇒ A non-addicted rural person has a nearly similar adherence status with the addicted one.

													N			
					R											
	N							Y								
		Y				N				Y				N		
	G		P		G		P		G		P		G		P	
Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	
,0%	35,7%	50,0%	25,0%	41,7%	72,2%	50,0%	73,1%	50,0%	64,3%	50,0%	58,3%	58,3%	50,0%	41,7%	64,3%	4
,0%	64,3%	50,0%	75,0%	58,3%	27,8%	50,0%	26,9%	50,0%	35,7%	50,0%	41,7%	41,7%	50,0%	58,3%	35,7%	5

C. **P(Adherence = Drop | FamilyDependsOn=No, Residence=Rural, Employed= Yes, Addiction=No, KnowledgeAbout=Poor, LostJob=No) = 0.269**.

⇒ The probability of poor adherence for an unburdened, employed, non-addicted ART client- is as low as 0.269 even though he/she is a rural person with poor knowledge of HIV.

							U								
		N								Y					
N				Y				N				Y			
	P		G		P		G		P		G		P		G
Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N
1,7%	64,3%	41,7%	50,0%	35,7%	35,7%	22,7%	88,1%	75,0%	84,4%	50,0%	61,1%	50,0%	72,2%	50,0%	58,3%
8,3%	35,7%	58,3%	50,0%	64,3%	64,3%	77,3%	11,9%	25,0%	15,6%	50,0%	38,9%	50,0%	27,8%	50,0%	41,7%

D. **P(Adherence = Drop | FamilyDependsOn=No, Residence=Urban, Employed= No, Addiction=Yes, KnowledgeAbout=Poor, LostJob=Yes) = 0.773**.

⇒ The probability of poor adherence for an unemployed, urban, addicted ART client whose knowledge about HIV is poor and who lost his/her job because of ill-health problems - is as high as 0.773 even though he/she not burdened of family responsibility.

E. **P(Adherence = Drop | FamilyDependsOn=No, Residence=Urban, Employed= Yes, Addiction=No, KnowledgeAbout=Good, LostJob=No) = 0.119**.

⇒ The probability of poor adherence for an unburdened, employed, non-addicted urban ART client with good background knowledge of HIV- is as low as 0.119. (N.B. This case is an ideal case for adherence)

F. **P(Adherence = Drop | FamilyDependsOn=No, Residence=Urban, Employed= Yes, Addiction=No, KnowledgeAbout=Poor, LostJob=No) = 0.156**.

⇒ The probability of poor adherence for an unburdened, employed, non-addicted urban ART client - is as low as 0.156 even though the client has poor background knowledge of HIV.

U															
N								Y							
N				Y				N				Y			
G		P		G		P		G		P		G		P	
N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y
58,3%	31,2%	50,0%	31,2%	50,0%	16,7%	41,7%	15,6%	58,3%	50,0%	58,3%	41,7%	41,7%	41,7%	43,8%	50,0%
41,7%	68,8%	50,0%	68,8%	50,0%	83,3%	58,3%	84,4%	41,7%	50,0%	41,7%	58,3%	58,3%	58,3%	56,2%	50,0%

G. **P(Adherence = Drop | FamilyDependsOn=Yes, Residence=Urban, Employed= No, Addiction=Yes, KnowledgeAbout=Good, LostJob=Yes) = 0.833**.

⇒ The probability of poor adherence for an unemployed, burdened, urban, addicted ART client who lost his/her job because of ill-health problems - is as high as 0.833 even though he/she has good knowledge about HIV.

H. **P(Adherence = Drop | FamilyDependsOn=Yes, Residence=Urban, Employed= No, Addiction=Yes, KnowledgeAbout=Poor, LostJob=Yes) = 0.844**.

⇒ The probability of poor adherence for an unemployed, burdened, urban, addicted ART client whose knowledge about HIV is poor and who lost his/her job because of ill-health problems - is as high as 0.844.

From the CPT, it could be realized that that the inferences made were meaningful. Moreover, two important points can be considered from the discussions made above:

I. For the best case (an ideal condition for adherence) which was shown in (**E**) above, the adherence of **88.1%** was observed.

II. For the worst case which was shown in (**H**) above, an adherence of only **15.6%** was observed.

These two inferential results are capable of justifying the BN model's strength (robustness) in prediction. Therefore, the predictions made by the BN PowerSoft were considered to be dependable for real-life practice.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1. Conclusion

This research made attempts to investigate the potential applicability of Bayesian Network technology in developing a model that can support the prediction of ART clients' adherence tends in Ethiopia.

Findings of the research tell us that the Bayesian Network model is a powerful predictor even in the absence of a domain expert. With a proper intervention of domain experts, it was observed to perform even better.

On the other hand, for critical real-life aspects such as ART adherence issues, it requires a more precise prediction than the one at hand. Actually this has a lot to do not only with the model but also the data itself. Some potential sources of errors that might have affected the result include:

- Sample size and its representativeness
- Incompleteness of attributes (Clinical aspects missed)
- Loss of vital information during new attribute derivation, attributes discretization, attribute generalization, and so on.

Yet, it was observed that as far as a complete and properly documented data with proper feature selection are exposed to the model, Bayesian Network's performance is dependable for real-life problem-solving purpose.

It is the researcher's belief that encouraging prediction accuracy of the model found in this research can give insight to many more researchers to be done for adapting models of such types for a different purpose in the research domain.

5.2. Recommendations

This research has investigated the potential applicability of Bayesian Network technology in developing a model that can support the prediction of ART clients' adherence tends in Ethiopia. It has also presented the applicability of the technology to predict major sociodemographic, social and behavioural factors behind poor ART adherence trends in the country. It is the researcher's belief that the model developed can help all the stakeholders working on HIV/AIDS prevention and control aspects.

The researcher would like to notify that this research, as an academic exercise, should only be considered as a preliminary effort to assess the applicability of Bayesian network technology for predicting major factors behind poor ART adherence trends in Ethiopia. Therefore, the findings of this research can be considered as a contribution towards more in-depth and comprehensive study in the application of Bayesian Network for prediction activities.

Additionally, the researcher would like to recommend the following points for further studies:

- ❖ **Combined Approach:** This research has dealt with only sociodemographic, social and behavioural dimensions as factors affecting ART adherence. But as the problem is a multifaceted one that needs inclusion of clinical aspects (such as **type of regimen** taken by the client and **CD4 count** of the client), the combined approach may bring about better results. Hence the recommendation.

- **Inference:** Except some visualizations that help to appreciate the inferencing aspects, this research has been focusing on **predicting** aspects only. Further inference mechanisms can increase the quality of the results.

- **Other Algorithms:** This research made use of **TPDA** algorithm for prediction purpose. The prediction results may be even better with other algorithms. The researcher recommends further researches using different algorithms.

- **Primary Data:** This research was based on secondary data. To this end, only the **worst case** of poor adherence (called "dropout") was taken as an indication of poor adherence. But, since missing very few doses has also a lot to do with poor adherence, primary data collection techniques such as DOTs (Directly-Observed-Treatments) in a randomly selected ART clients (without their knowledge) may lead to better results on their adherence trends.

- **A Data with Missing Values:** Several assumptions are usually made to learn structure and parameters from data. One of these is the assumption that each case in the data set specifies a value for every variable discerned, that is, there are no missing values. Unfortunately, for most real-life data sets this property does not hold. Therefore, the researcher recommends Bayesian Network application algorithms for incomplete data.

- **Developing an Expert System:** Since ART adherence issue has critical social, economical and behavioural dimensions that need proactive measures, the researcher recommends development of an expert system that gives an early-warning mechanism during the client's registration time. This can help the service providers what to do even before the client starts the treatment. A full-fledged model of the type proposed in this research can serve as a core module in developing the system.

- **Extending the Application:** The results found from this research were encouraging. Similar researches can be conducted in treatments other than ART that need a high-level of adherence (for example, Tuberculosis).

REFERENCES:

1. **Abadía-Barrero, C., Castro, A. 2005.** *Experiences of stigma and access to HAART in children and adolescents living with HIV/AIDS in Brazil. Soc Sci Med Epub ahead of print.*

2. **Adam, BD., Maticka-Tyndale, E., Cohen, J. 2005.** Adherence practices among people living with HIV. *AIDS care 2005, 15(2):263-274.*

3. **Alemayehu, A. et al. 2008.** ART Adherence Challenges in Southwest Ethiopia.

4. **Amico, K., et al. 2004.** Efficacy of ART Adherence Interventions. *A research Synthesis of Trials*

5. **Ammassari, A., et al. 2002.** Relationship between HAART adherence and adipose tissue alterations. *J Acquir Immune Defic Syndr 31:S140–S144.*

6. **Andreassen, S. 1992.** Planning of therapy and tests in causal probabilistic networks. *Artif Intell Med 1992;4: 227–41.*

7. **Andreassen, S., et al. 1999.** Using probabilistic and decision theoretic methods in treatment and prognosis modeling. *Artif Intell Med 1999;15:121–34.*

8. **Artificial Intelligence in Medicine. 2004.** Bayesian networks in biomedicine and health-care

   URL: http://www.intl.elsevierhealth.com/journals/aiim

9. **Bangsberg, DR., et al. 2003.** High levels of adherence do not prevent accumulation of HIV drug resistance mutations. *AIDS 2003;17(13):1925-32.*

10. **Boaz**, **L. 2001.** Bayesian fluorescence in situ hybridization signal classification.

11. **Castro, A., Farmer, P. 2005**. Understanding and addressing AIDS-related stigma: From anthropological theory to clinical practice in Haiti. *Am J Public Health 95:53–59.*

12. **Castro, A. 2005.** Adherence to Antiretroviral Therapy: Merging the Clinical and Social Course of AIDS. *PLoS Med 2(12): e338 doi:10.1371/journal.pmed.0020338*

13. **Castro, A., Singer, M. 2004.** Unhealthy health policy: A critical anthropological examination. *Walnut Creek (California): Altamira Press. 408*

14. **Cheng, J., et al. 1997.** An algorithm for Bayesian belief network construction from data. *Proc. 6th International Workshop on Artificial Intelligence and Statistics, Florida. 83-90.*

15. **Coetzee, D., et al. 2004**. *Promoting adherence to antiretroviral therapy*: The experience from a primary care setting in Khayelitsha, South Africa. *AIDS 18:S27–S31*

16. **Conati, C., and VanLehn, K. 1996.** Using Bayesian Networks to Manage Uncertainty in Student Modeling

17. **CSA. 2008.** The Central Statistics Authority, Annual Report

18. **Corless, IB., et al. 2000.** *Cognition and HIV adherence.* XIIth International Conference on AIDS. Geneva.

19. **Delaunay, K. 2001**. La mise sous traitement antirétroviral dans l'Initiative*: l'explicite et l'implicite d'un processus de sélection.*

20. **Dempster, A., et al. 1977**. Maximisation likelihood from incomplete data via the EM algorithm. *J R Stat Soc B 1977;39:1–38.*

21. **Desclaux, A. 2004.** Equality in access to AIDS treatment in Africa: Pitfalls among achievements. *In: Castro A, Singer M, eds. Unhealthy health policy: A critical anthropological examination Walnut Creek (California): Altamira Press. 408 p.*

22. **ESAMI- East and South African Management Institute. 2008.** ART Adherence in Resource-Limited Settings

23. **Escobar, I., et al. 2003.** Factors Affecting Patient Adherence to Highly Active Antiretroviral Therapy**.** *The Annals of Pharmacotherapy: Vol. 37, No. 6, pp. 775-781. DOI 10.1345/aph.1C427*

24. **Farmer, P. 1990.** Sending sickness: Sorcery, politics, and changing concepts of AIDS in rural Haiti. *Med Anthropol Q 4:6–27.*

25. **Farmer, P. 1992.** AIDS and accusation: Haiti and the geography of blame. *Berkeley: University of California Press. 338 p.*

26. **Farmer, P. 1994.** *AIDS*-talk and the constitution of cultural models. *Soc Sci Med 38:801–809.*

27. **Federal HAPCO. 2007a.** Annual HIV/AIDS Monitoring and Evaluation Report

28. **Federal HAPCO. 2007b.** Guidelines for Implementation of ART in Ethiopia
29. **Federal HAPCO. 2007c.** Guideline for Management of Opportunistic Infections and ART in Adolescents and Adults in Ethiopia
30. **Federal HAPCO. 2008.** Report on Progress towards Implementation of the UN Declaration of Commitment on HIV/AIDS
31. **Federal MOH Roadmap. 2007.** A Roadmap for Implementation of ART in Ethiopia
32. **Friedland, G., Williams, A. 1999**. Attaining higher goals in HIV treatment: The central importance of adherence. *AIDS 13:S61–S72.*
33. **Friedman, NIR., et al. 2000.** Using Bayesian network to analyze expression data. *J Comput Biol 2000;7:601–20.*
34. **Galan, SF., et al. 2001.** NasoNet: joining Bayesian networks and time to model nasopharyngeal cancer spread. *In: Proceedings of the Eighth International Conference on Artificial Intelligence in Medicine in Europe (AIME 2001).*
35. Gao, X., et al**. 2000.** The relationship of disease severity, health beliefs and medication adherence among patients. *Aids Care* 2000, **12**(4)**:**387-398.
36. **Greene, JA., 2004**. An Ethnography of Non-adherence: Culture, poverty, and tuberculosis in urban Bolivia. *Cult Med Psychiatry 28:401–425.*
37. **Hastie, T., et al. 2001.** The elements of statistical learning: data mining, inference, and prediction. *New York: Springer, 2 001.*
38. **Heckman, DE., et al. 1992.** Towards normative expert systems. I. *The Pathfinder project. Meth Inform Med 1992; 31:90–105.*
39. **Ickovics, JR., Meade, CS. 2002.** Adherence to antiretroviral therapy among patients with HIV: a critical link between behavioural and biomedical sciences. J *Acquir Immune Defic Syndr 2002;31 Suppl 3:S98-102.*
40. ***Jame,s C,.* & Anise, K., *et al. 2008*.** Predictors of Adherence to Antiretroviral Therapy in Rural Zambia*. Epidemiology and Social Science*
41. **Jensen, F.V. 1996.** An introduction to Bayesian network. *London. U.K: University College London Press.*
42. **Jerene, D. 2007.** HIV Antiretroviral Therapy in Ethiopia

43. **Katzenstein, D., et al. 2003).** The evaluation of the HIV/AIDS drug access inititiatives in Côte d'Ivoire, Senegal and Uganda*; how access to antiretroviral treatment can become feasible in Africa. AIDS 17:S1–S4.*

44. **Kebede, D., et al. 2008.** Defaulters from antiretroviral treatment in Jimma University Specialized Hospital; *Southwest Ethiopia*

45. **Kleinman, A., et al. 1995.** The social course of epilepsy: Chronic illness as social experience in interior China*. Soc Sci Med 40:1319–1330.*

46. **Koenig, S., et al. 2004.** *Scaling*-up HIV treatment programmes in resource-limited settings: *The rural Haiti experience. AIDS 18:S21–S25*.

47. **Lam, W., Bacchus, F. 1994.** Learning Bayesian belief networks: an approach based on the MDL principle. *Comput Intell 1994;10:269–93*.

48. **Larranaga, P., et al. 1996.** Structure learning of Bayesian networks by genetic algorithms: a performance analysis of control parameters*. IEEE Trans Pattern Anal Mach Intell 1996;18(9):912–26.*

49. **Lauritzen, S. L., et al. 1990.** Independence properties of directed Markov fields, Networks

50. **Lévi-Strauss, C. 1958.** Anthropologie structurale. Paris: *Plon. 480 p.*

51. **Lise, G., et al. 2001.** Understanding tuberculosis epidemiology using structured statistical models

52. **Lucas, P. 2007.** Bayesian Analysis, Pattern Analysis and Data Mining in Health Care.

53. **Lucas, PJF., et al. 1998.** Computer-based decision-support in the management of primary gastric non- Hodgkin lymphoma. *Meth Inform Med 1998;37:206–19*.

54. **Machtinger, E., Bangsberg, DR. 2006.** Adherence to HIV antiretroviral therapy. HIV INSite Knowledge Base Chapter.

    URL: *http://hivinsite.ucsf.edu/InSite?page=kb-03-02-09#S1X 2005*

55. **McNicholl, I. 2005.** Adverse Events of Antiretroviral Drugs. University of California San Francisco*. Retrieved on 20. 09.2008*

56. **Mehta, S., et al. 1997.** Potential factors affecting adherence with HIV therapy. *AIDS 1997, 11:1665-1670*.

57. **Melchior, R.** & **Nemes, M., et al. 2007.** Challenges of treatment adherence by people living with HIV/AIDS in Brazil. *Rev. Saúde Pública vol.41 suppl.2 São Paulo Dec. 2007*

58. **Moatti, JP., et al. 2004.** Drug resistance and adherence to HIV/AIDS antiretroviral treatment: against a double standard between the north and the south. *AIDS 2004;18 Suppl 3:S55-61*.

59. **Nadkarni, S., Shenoy, P. 2004.** A Causal Mapping Approach to Constructing Bayesian Networks. Decision Support Systems, Vol. 38, Issue 2, pp. 259--281.

60. **Nemes, M., et al. 2004.** ART adherence in Brazil. *AIDS 18:S15–S20*.

61. **Neapolitan, R.E. 2004.** *Probabilistic* reasoning in expert systems*: Theory and algorithms, John Wiley & Sons, 1990.*

62. **Nevin, Z., et al. 2002.** Latent variable discovery in classification models.

63. **Nwokike, JI. 2003**. Baseline data and predictors of adherence to antiretroviral therapy in Maun General Hospital, Botswana. *Antivir Ther 8:S396.*

64. **Partners in Health. 2004.** The PIH guide to the community-based treatment of HIV in resource-poor settings. Boston: Partners in Health. Available: *http://www.pih.org/library/aids/PIH_HIV_Handbook_Bangkok_edition.pdf*.

65. **Paterson, DL., et al. 2000.** *Adherence* to protease inhibitor therapy and outcomes in patients with HIV infection. *Ann Intern Med 2000;133(1):21-30*.

66. **Pearl, J. 1988**. *Probabilistic Reasoning* in Intelligent Systems: Networks of Plausible Inference. *Morgan Kaufmann, San Mateo, CA.*

67. **Pérez, J., et al. 2004**. Approaches to the Management of HIV/AIDS in Cuba. Perspectives and Practice in Antiretroviral Treatment (Case Study). *Geneva*: WHO. URL: http://www.who.int/hiv/pub/prev_care/en/cuba.pdf.

68. **Peter, A., et al. 2003.** *Using literature and data to learn Bayesian networks as clinical models of ovarian tumors.*

69. **Rabkin, JG., Chesney, MA. 1999**. Treatment adherence to AIDS medications: the Achilles heel of the new therapeutics. *In Psychosocial and Public Health Impacts of New HIV Therapies. Edited by: Ostrow D, Kalichman S. New York: Kluwer/Plenum; 1999:61-68.*

70. **Rahel, B. 2005.** Computer-Assisted Learner Group Formation Based on Personality Traits. *Dissertationsschrift zur Erlangung des Grades eines Doktors* der Naturwissenschaften am Fachbereich Informatik der Universität Hamburg

71. **Ramoni, M., et al. 2002.** Bayesian clustering by dynamics. *Mach Learn 2002;47:91–121.*

72. **Russel, S., Norvig, 2003.** *Artificial intelligence:* a modern approach, 2nd ed. NJ: Prentice Hall.

73. **Saitoh, A., et al. 2005.** Myelomeningocele in an infant with intrauterine exposure to efavirenz. *J Perinatol.* ***25****, 555-556 PMID 16047034*

74. **Sethi, AK., et al. 2003.** Association between adherence to ART and human immunodeficiency virus drug resistance. *Clin Infect Dis 37:1112–1118*.

75. **Shachter, RD. 1986.** *Evaluating influence diagrams.* Oper Res 1986;34(6):871–82.

76. **Silvia, A., et al. 2000.** A comparison of learning algorithms for Bayesian networks: a case study based on data of emergency medical services

77. **Singer, M., Clair, S. 2003.** Syndemics and public health: Reconceptualizing disease in bio-social context. *Med Anthropol Q 17:423–441*.

78. **Spirtes, P., et al. 1996.** *Causation, Prediction, and Search.* URL- http://hss.cmu.edu/html/departments/philosophy/TETRAD.BOOK/book.html,1996

79. **Stewart, KE., et al. 1999.** *Prospective prediction of ARV adherence: the role of patient-perceived barriers*. In Program and abstracts of the 39th Interscience Conference on Antimicrobial Agents and Chemotherapy; San Francisco. Abstract 587.

80. **Stout, B., et al. 2003.**. *Determining* risk factors associated with non-adherence in HIV patients in Costa Rica. 2nd International AIDS Society Conference on HIV Pathogenesis and Treatment; 2003 July 13–16 July; Paris, France. Abstract no. 675. URL: http://www.iasociety.org/abstract/show.asp?abstract_id=9974.

81. **Tadios, Y., Davey, G. 2006.** Antiretroviral treatment adherence and its correlates among people living with HIV/AIDS on highly active antiretroviral therapy in Addis Ababa, Ethiopia.

82. **Tesoriero, J., et al. 2003.** Stability of adherence to highly active antiretroviral therapy over time among clients enrolled in the treatment adherence demonstration project. *J Acquir Immune Defic Syndr 33:484–493*.

83. **USAID. 2008.** ART Adherence Counselling. HIV/AIDS Care and Support Program

84. **Ware, N., et al. 2005**. Adherence, stereotyping and unequal HIV treatment for active users of illegal drugs. *Soc Sci Med 61:565–576.*

85. **WIC. 2008.** ART and adherence: opportunities and challenges. *Walta Information Center*.

86. **WHO. 2003.** Scaling-up antiretroviral therapy in resource-limited settings. *Treatment guideline for a public health approach.*

87. **WHO. 2005a.** Interim WHO Clinical Staging of HIV/AIDS and HIV/AIDS Case Definitions for Surveillance. Africa Region.

88. **WHO 2005b.** Summary Country Profile for HIV/AIDS Treatment Scale-Up

LIST OF APPENDICES

Appendix 1: Confusion Matrices for Experiment I

Appendix 1 (a): Confusion Matrix for the Third Partition

	PREDICTED		
ACTUAL	ADHERE	DROP	TOTAL
ADHERE	978	67	1045
DROP	318	42	360

Appendix 1 (b): Confusion Matrix for the Fifth Partition

	PREDICTED		
ACTUAL	ADHERE	DROP	TOTAL
ADHERE	992	92	1084
DROP	289	32	321

Appendix 1 (c): Confusion Matrix for the Sixth Partition

	PREDICTED		
ACTUAL	ADHERE	DROP	TOTAL
ADHERE	987	92	1079
DROP	293	33	326

Appendix 1 (d): Confusion Matrix for the Third Partition

	PREDICTED		
ACTUAL	ADHERE	DROP	TOTAL
ADHERE	985	79	1064
DROP	304	37	341

Appendix 1 (e): Confusion Matrix for the Third Partition

	PREDICTED		
ACTUAL	ADHERE	DROP	TOTAL
ADHERE	944	98	1042
DROP	284	79	363

Appendix 1 (f): Confusion Matrix for the Third Partition

	PREDICTED		
ACTUAL	ADHERE	DROP	TOTAL
ADHERE	949	102	1051
DROP	281	73	354

Appendix 1 (g): Confusion Matrix for the Third Partition

	PREDICTED		
ACTUAL	ADHERE	DROP	TOTAL
ADHERE	982	97	1079
DROP	295	31	326

Appendix 2: Confusion Matrices for Experiment- II

Appendix 2 (a): Confusion Matrix for the First Partition

	PREDICTED		
ACTUAL	ADHERE	DROP	TOTAL
ADHERE	902	93	995
DROP	257	153	410

Appendix 2(b): Confusion Matrix for the Second Partition

	PREDICTED		
ACTUAL	ADHERE	DROP	TOTAL
ADHERE	929	47	976
DROP	282	147	429

Appendix 2 (c): Confusion Matrix for the Sixth Partition

	PREDICTED		
ACTUAL	ADHERE	DROP	TOTAL
ADHERE	917	45	962
DROP	282	161	443

Appendix 2(d): Confusion Matrix for the Seventh Partition

	PREDICTED		
ACTUAL	ADHERE	DROP	TOTAL
ADHERE	926	59	985
DROP	287	133	420

Appendix 2 (e): Confusion Matrix for the Eighth Partition

	PREDICTED		
ACTUAL	ADHERE	DROP	TOTAL
ADHERE	930	57	987
DROP	289	129	418

Appendix 2 (f): Confusion Matrix for the Ninth Partition

	PREDICTED		
ACTUAL	ADHERE	DROP	TOTAL
ADHERE	913	80	993
DROP	274	138	412

Appendix 2(g): Confusion Matrix for the Tenth Partition

	PREDICTED		
ACTUAL	ADHERE	DROP	TOTAL
ADHERE	920	64	984
DROP	299	122	421

Appendix 3: Questions Introduced to Domain Experts

A. During Data Collection

1. Do you think sociodemographic, social and behavioural factors can affect an ART client's adherence?
2. If your answer to question 1 is "**Yes**", which of the below- mentioned sociodemographic, social and behavioural factors do you think can likely affect an ART client's adherence? (Underline the ones you believe so.) **[List of Potential attributes were given to the Experts]**
3. Please suggest on how to categorize values of the attribute "*Understanding/knowledge of HIV Diseases*" without loss of much information.

B. After Experiment-I was conducted

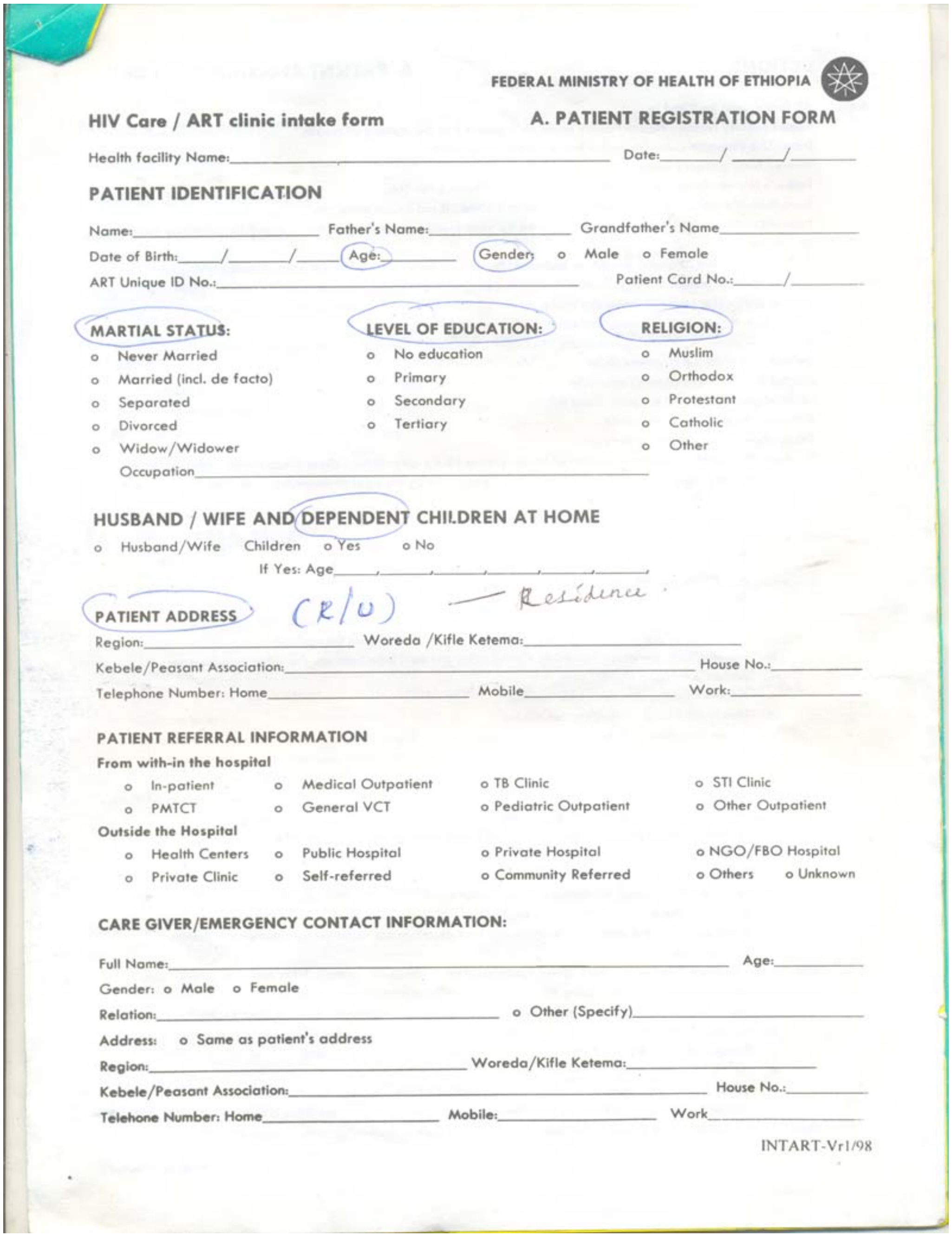
4. Do you think there are dependencies among the factors you already selected?
5. Bayesian Network tool (called "BN PowerConstructor") has already delivered the following dependency structure. Do you agree on this structure? **[The best Learned Model was given to the domain experts]**
6. The machine-learned structure may have lots of pitfalls. Please feel free to modify the structure in such a way you think correct.

Appendix 4: Domain Experts' Qualification

	Sex	Qualification
Expert 1	M	MD, MPH, PhD
Expert 2	M	MPH
Expert 3	F	MPH

Appendix 5: ART Registration Forms used by the Researcher

Form 1



FEDERAL MINISTRY OF HEALTH OF ETHIOPIA

HIV Care / ART clinic intake form

A. PATIENT REGISTRATION FORM

Health facility Name:__________ Date:____/____/____

PATIENT IDENTIFICATION

Name:________ Father's Name:________ Grandfather's Name________

Date of Birth:____/____/____ Age:____ Gender: o Male o Female

ART Unique ID No.:________ Patient Card No.:____/____

MARTIAL STATUS:	LEVEL OF EDUCATION:	RELIGION:
o Never Married	o No education	o Muslim
o Married (incl. de facto)	o Primary	o Orthodox
o Separated	o Secondary	o Protestant
o Divorced	o Tertiary	o Catholic
o Widow/Widower		o Other

Occupation________

HUSBAND / WIFE AND DEPENDENT CHILDREN AT HOME

o Husband/Wife Children o Yes o No

If Yes: Age___,___,___,___,___,___

PATIENT ADDRESS (R/U) — Residence

Region:________ Woreda /Kifle Ketema:________

Kebele/Peasant Association:________ House No.:________

Telephone Number: Home________ Mobile________ Work:________

PATIENT REFERRAL INFORMATION

From with-in the hospital

o In-patient	o Medical Outpatient	o TB Clinic	o STI Clinic
o PMTCT	o General VCT	o Pediatric Outpatient	o Other Outpatient

Outside the Hospital

o Health Centers	o Public Hospital	o Private Hospital	o NGO/FBO Hospital
o Private Clinic	o Self-referred	o Community Referred	o Others o Unknown

CARE GIVER/EMERGENCY CONTACT INFORMATION:

Full Name:________ Age:________

Gender: o Male o Female

Relation:________ o Other (Specify)________

Address: o Same as patient's address

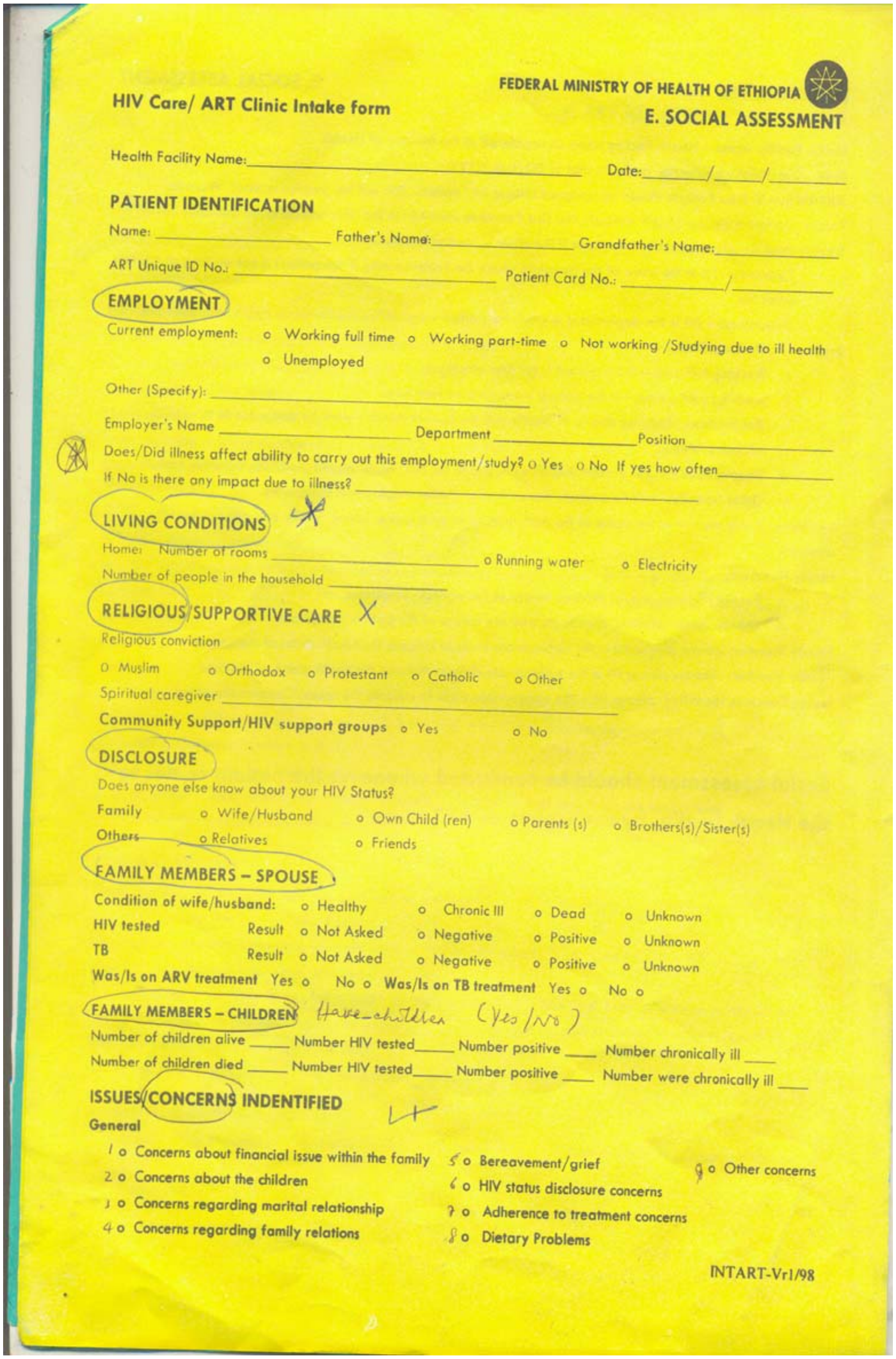
Region:________ Woreda/Kifle Ketema:________

Kebele/Peasant Association:________ House No.:________

Telehone Number: Home________ Mobile:________ Work________

INTART-Vr1/98

Form 2



FEDERAL MINISTRY OF HEALTH OF ETHIOPIA

HIV Care/ ART Clinic Intake form

E. SOCIAL ASSESSMENT

Health Facility Name:______________________ Date:____/____/____

PATIENT IDENTIFICATION

Name: ______________ Father's Name:______________ Grandfather's Name:______________

ART Unique ID No.:______________ Patient Card No.:________/________

EMPLOYMENT

Current employment: o Working full time o Working part-time o Not working /Studying due to ill health o Unemployed

Other (Specify):______________

Employer's Name ______________ Department ______________ Position ______________

Does/Did illness affect ability to carry out this employment/study? o Yes o No If yes how often______________

If No is there any impact due to illness? ______________

LIVING CONDITIONS

Home: Number of rooms ______________ o Running water o Electricity

Number of people in the household ______________

RELIGIOUS/SUPPORTIVE CARE X

Religious conviction

o Muslim o Orthodox o Protestant o Catholic o Other

Spiritual caregiver ______________

Community Support/HIV support groups o Yes o No

DISCLOSURE

Does anyone else know about your HIV Status?

Family o Wife/Husband o Own Child (ren) o Parents (s) o Brothers(s)/Sister(s)

Others o Relatives o Friends

FAMILY MEMBERS – SPOUSE

Condition of wife/husband: o Healthy o Chronic Ill o Dead o Unknown

HIV tested Result o Not Asked o Negative o Positive o Unknown

TB Result o Not Asked o Negative o Positive o Unknown

Was/Is on ARV treatment Yes o No o Was/Is on TB treatment Yes o No o

FAMILY MEMBERS – CHILDREN Have children (Yes/No)

Number of children alive ____ Number HIV tested____ Number positive ____ Number chronically ill ____

Number of children died ____ Number HIV tested____ Number positive ____ Number were chronically ill ____

ISSUES/CONCERNS INDENTIFIED

General

1 o Concerns about financial issue within the family
2 o Concerns about the children
3 o Concerns regarding marital relationship
4 o Concerns regarding family relations
5 o Bereavement/grief
6 o HIV status disclosure concerns
7 o Adherence to treatment concerns
8 o Dietary Problems
9 o Other concerns

INTART-Vr1/98

Form 3

FEDERAL MINISTRY OF HEALTH OF ETHIOPIA

HIV Care/ ART Clinic **F. ART ADHERENCE COUNSELING**

Health Facility Name:______________________ Date:____/____/______

PATIENT IDENTIFICATION

Name: ______________ Father's Name:______________ Grandfather's Name:______________

ART Unique ID No.: ______________________ Patient Card No.: ________/________

HEALTH EDUCATION & KNOWLEDGE

- o Attended HIV related health education session(s) in the past
- o Attended HIV related counseling session(s) in the past

Understanding of HIV disease:	o NA	o -	o +	o ++	o +++
Understanding of HIV transmission:	o NA	o -	o +	o ++	o +++
Understanding of prophylaxis and treatment of OI:	o NA	o -	o +	o ++	o +++
Understanding of ART medication adherence:	o NA	o -	o +	o ++	o +++

RISK – BEHAVIOR

1. Regular
2. Casual
3. NO

- o Has regular sexual partner
- o Has casual sexual partner(s) – Number of casual partners in last 3 months o 1 o 2 o 3 o >3

Condom use o NA o Never o Rarely o Sometimes o Mostly o Always o No response X

Addictions

1. Tobacco	o NA	o -	o +	o ++	o +++	
2. Alcohol	o NA	o -	o +	o ++	o +++	
3. Soft Drugs	o NA	o -	o +	o ++	o +++	e.g., Khat, Shisha, pills, etc
4. Hard Drugs	o NA	o -	o +	o ++	o +++	e.g., cocaine, morphine, i.v.-drugs, etc.

Adherence: Concerns/barriers to ART:

1 o Stigma (family and friends will find out) 4 o Depressed/anxious 4

2 o Afraid of medications (side effects; "poison") 5 o Will forget to take medications

3 o Doubt that medications will work 6 o Other ______________

GENERAL FEELING

Since your last visit, have you had any problems or complaints? o No o Yes

Have you been hospitalized? o No o Yes

How has your appetite been since your last visit? o Not Asked o Good o OK o Poor

How has your strength been since your last visit?

o Normal o Weak, but not in bed o Very weak, often in bed o Extremely weak, mostly in bed

X How many days have you been too sick to work? ________ o Lost job due to current illness

Evaluator's impression about mental condition

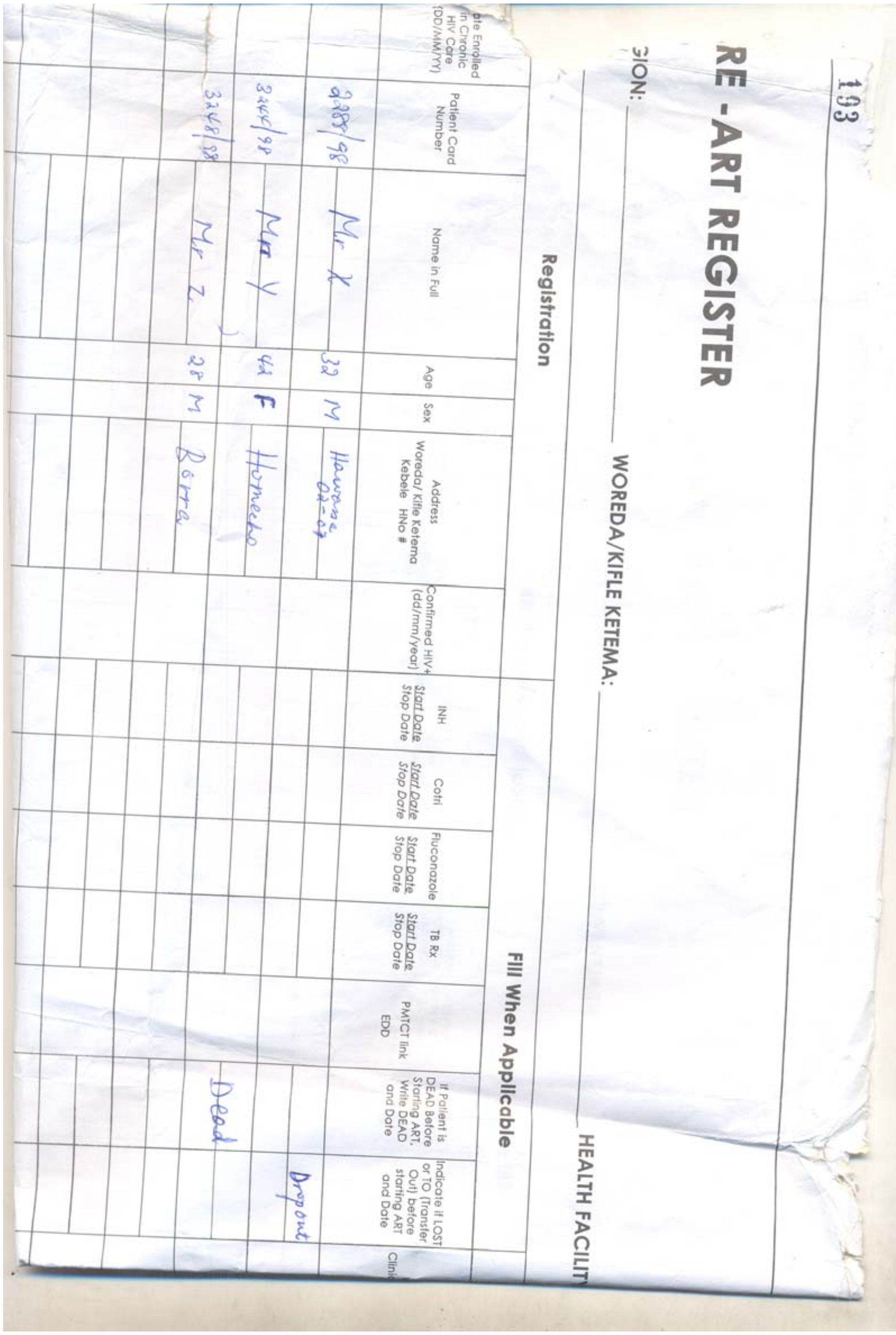
o At ease o Confused o Depressed o Anxious o Suicidal

APPROPRIATE REFERAL

o Physician o Pharmacy o Social Services o Laboratory o Community Based Organizations

INTART-Vr1/98

Form 4



193

PRE -ART REGISTER

REGION: ______ WOREDA/KIFLE KETEMA: ______ HEALTH FACILITY

Registration							Fill When Applicable							
Date Enrolled in Chronic HIV Care (DD/MM/YY)	Patient Card Number	Name in Full	Age	Sex	Address Woreda/ Kifle Ketema Kebele HNo #	Confirmed HIV+ (dd/mm/year)	INH Start Date Stop Date	Cotri Start Date Stop Date	Fluconazole Start Date Stop Date	TB Rx Start Date Stop Date	PMTCT link EDD	If Patient is DEAD Before Starting ART, Write DEAD and Date	Indicate if LOST or TO (Transfer Out) before starting ART and Date	Clinic
	2288/98	Mr X	32	M	Hawzna 02-07								Drop out	
	3446/98	Mrs Y	42	F	Hormeko							Dead		
	3248/98	Mr Z	28	M	Borra									

Appendix 6: Sample Data of the ART Clients

Microsoft Excel - ART data NOV24_9

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Age	Sex	MaritalSta	LevEduc	Religion	FamilyDe	Residenc	Employed	StatKnow	HaveChil	SexPartn	Addiction	knowledg	LostJob	Adherence
2	YOUNG	F	M	I	P	N	R	N	N	N	R	Y	G	N	ADHERE
3	MIDDLE_A	F	D	S	O	Y	U	N	Y	Y	C	Y	P	Y	DROP
4	MIDDLE_A	M	W	S	P	N	R	Y	N	N	C	N	G	N	ADHERE
5	MIDDLE_A	F	W	I	M	Y	R	N	Y	Y	C	N	P	Y	ADHERE
6	MIDDLE_A	F	N	P	O	N	U	N	N	N	C	Y	P	Y	DROP
7	MIDDLE_A	F	M	S	P	N	U	Y	Y	N	R	Y	G	N	DROP
8	YOUNG	M	N	I	P	N	U	N	Y	N	NO	N	P	N	ADHERE
9	YOUNG	M	N	I	C	N	U	N	Y	N	C	N	P	N	ADHERE
10	YOUNG	F	M	S	P	N	R	N	Y	Y	R	N	P	N	DROP
11	MIDDLE_A	M	M	S	P	Y	U	Y	Y	Y	R	Y	G	N	ADHERE
12	MIDDLE_A	F	W	I	M	Y	U	N	N	Y	C	Y	P	N	ADHERE
13	YOUNG	M	N	I	M	N	R	N	Y	N	NO	N	P	N	DROP
14	SENIOR	F	W	I	M	Y	R	Y	Y	Y	C	N	P	N	ADHERE
15	YOUNG	F	W	P	M	Y	U	N	N	Y	C	N	G	N	ADHERE
16	YOUNG	F	M	P	P	Y	U	Y	Y	Y	R	Y	G	N	ADHERE
17	MIDDLE_A	F	M	I	M	Y	R	N	Y	Y	C	Y	P	N	DROP
18	MIDDLE_A	F	W	S	P	Y	U	N	Y	Y	R	N	P	Y	DROP
19	MIDDLE_A	M	M	S	P	N	U	Y	N	N	R	Y	G	N	ADHERE
20	YOUNG	M	N	P	O	N	U	N	N	N	C	Y	G	Y	DROP
21	YOUNG	M	M	S	P	Y	U	N	Y	Y	C	Y	G	Y	ADHERE
22	SENIOR	M	D	I	O	Y	R	Y	Y	Y	C	N	P	N	DROP
23	MIDDLE_A	F	M	S	O	Y	U	Y	N	Y	C	N	P	N	ADHERE
24	SENIOR	F	M	I	P	Y	R	N	Y	Y	C	N	P	Y	ADHERE
25	MIDDLE_A	F	M	I	M	Y	R	Y	Y	Y	R	Y	P	N	ADHERE
26	YOUNG	F	N	I	M	N	R	N	Y	N	NO	N	P	N	DROP
27	MIDDLE_A	F	W	I	P	Y	R	N	N	Y	C	N	P	N	ADHERE
28	MIDDLE_A	F	M	S	P	Y	U	Y	Y	Y	R	N	G	N	ADHERE
29	MIDDLE_A	F	M	P	O	Y	U	N	N	Y	R	N	G	N	ADHERE
30	YOUNG	M	N	I	O	N	U	N	Y	N	C	N	P	N	DROP
31	SENIOR	F	W	I	P	Y	R	N	Y	Y	R	N	P	N	ADHERE
32	MIDDLE_A	M	M	S	P	Y	U	Y	Y	Y	R	Y	P	N	DROP
33	MIDDLE_A	M	M	T	P	N	U	N	Y	N	C	N	G	Y	ADHERE
34	YOUNG	M	N	P	O	N	R	N	N	Y	C	Y	G	N	ADHERE
35	YOUNG	F	N	I	M	N	R	N	Y	N	NO	N	P	N	DROP

TRAIN9 TEST9