A MODELFOR PREDICTING NON-ADHERENCE AMONG PRE-EXPOSURE PROPHYLAXIS (PrEP) CLIENTS AT SUBA REGION.

BY

FRED BOSIRE NYATIKA

MASTER OF SCIENCE IN DATA ANALYTICS

KCA UNIVERSITY

2018

A MODELFOR PREDICTING NON-ADHERENCE AMONG PRE-EXPOSURE PROPHYLAXIS (PrEP) CLIENTS AT SUBA REGION.

BY

FRED B. NYATIKA

A DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF MSC. DATA ANALYTICS IN THE FACULTY OF COMPUTING AND INFORMATION MANAGEMENT AT KCA UNIVERSITY

NOVEMBER, 2018

DECLARATION

I declare that this dissertation is my original work and has not been previously published or submitted elsewhere for award of a degree. I also declare that this contains no material written or published by other people except where due reference is made and author duly acknowledged.

Student Name: **Fred Bosire Nyatika** Registration No: **15/02936**

Sign:  Date: **09/11/2018**

I do hereby confirm that I have examined the master's dissertation of

Fred B. Nyatika

And approved it for examination.

Sign: _____ Date: _____

Dr. Henry Mwangi

Supervisor

Sign: _____ Date: _____

Mrs. Rachael Kibuku

Supervisor

ABSTRACT

The immune system protects the body against diseases or any foreign body that is harmful to the body, it is the body's natural defense against illnesses. The threats that the immune system attacks include viruses, bacteria and parasites. Human Immunodeficiency Virus (HIV) is a virus that is responsible for the Acquired Immunodeficiency Syndrome (AIDS). AIDS is a set of symptoms that occur after HIV infection, it is when the body's immune system is too weak to fight off infection(Rachel Nall 2016). There has been numerous attempts by the government and indeed the foreign donors to try and contain the epidemic. The use of *test and treat* has had a great impact by ensuring that those who are tested HIV positive are put on care immediately and with good adherence then we expect to have low transmission rates. The latest approach by the world health organization is the recommendation to use *test and treat* combined with putting the most at risk for HIV infection on the daily pill of Pre-Exposure Prophylaxis(PrEP) which reduces the risk of getting infected by HIV by more than 90%(CDC 2018a). PrEP is very effective if taken consistently failure to which may lead to HIV infection. It has been observed that there still exist a number of clients who are becoming HIV positive even after being put on PrEP. The objective of the study was to develop a model to predict non-adherence to PrEP among PrEP enrolled clients in Suba region of Homabay County. The study utilized data that had been collected from the most at risk population who are enrolled in PrEP for the period between the years 2016 to 2018. Data mining on the data played an important role in coming up with the model.

Keywords: Data mining, PrEP, non-adherence

ACKNOWLDEGEMENT

I acknowledge the support that I have received from my supervisors guiding me through to where I have reached with my dissertation. I specifically acknowledge one Mr. Placide Ntwali Nsengiyumva and my employer KEMRI SEARCH study who made it possible for me to start my MSc program. I also acknowledge my colleagues at work who have offered me encouragement to reach where I am today. I can't forget my family who have given me ample time to write my proposal then to research progress and finally to the dissertation and above all my Father in heaven who has guided and taken care of me for this far.

DEDICATION

This dissertation is dedicated to my wife Loyce Moraa and my son Brandon Bosire, who made the environment conducive for me to reach this far. Lastly I dedicate this work to my able parents Samson Nyatika and Joyce Arasa for their continuous prayers thus far.

ACRONYMS AND ABBREVIATIONS

AIDS:	Acquired Immune Deficiency Syndrome
ART	Anti-Retroviral Therapy
ARV	Anti-RetroViral
HIV	Human Immunodeficiency Virus
PrEP	Pre-Exposure Prophylaxis
WHO	World Health Organization
CD4	Cluster of Differentiation 4
CDC	Centers for Disease Control
USAID	United States Agency for International Development
NACC	National AIDS Control Council
UNAIDS	Joint United Nations Programme on HIV/AIDS
PLHIV	People Living with HIV
PBMC	Peripheral Blood Mononuclear Cells

OPERATIONAL DEFINITION OF TERMS

1. **Serodiscordant:** Couples in a sexual relationship where one of them is HIV positive and the other one is HIV negative. They are also known as mixed serostatus.
2. **Serostatus:** refers to whether an individual has HIV or not. Sero refers to blood serum. It is either one has a positive serostatus meaning that the person has HIV or one has a negative serostatus meaning that the person has no HIV.
3. **Boda boda:** motorcycle taxis commonly found in East Africa. Motorcycle taxis are present in Africa and beyond but the term boda boda is specific to East Africa, in Kenya it is also known as piki piki.
4. **HIV Prevalence:** proportion of the population tested and found to be HIV positive at a specific point in time. It is normally expressed in percentage of the population.
5. **Immunodeficiency:** a state in which the body's immune system's ability to fight infectious diseases is compromised or it is completely absent. Primary immunodeficiency is as a result of genetics and the secondary immunodeficiency is due to environmental factors such as HIV/AIDS or malnutrition.
6. **Seroprevalence:** overall occurrence of a disease or a condition within a defined population at one time as measured by blood tests also known as serologic tests.
7. **Incidence:** total number of individuals that acquire HIV during a particular time period such as a month or a year.
8. **Pre-Exposure Prophylaxis (PrEP):** it is a way for people who are HIV negative but at a substantial risk of getting HIV to prevent getting the virus by taking a pill every day. The pill is called truvada and it is a combination of two medicines (tenofovir and emtricitabine) that are used in combination with other medicines to treat HIV.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE: INTRODUCTION

1.1 Background of the study

Human Immunodeficiency Virus (HIV) is responsible for destruction of Cluster of Differentiation 4 cells (CD4) which are part of the immune system, this leads to entry of opportunistic infections that cause Acquired Immune Deficiency Syndrome (AIDS) and this may cause death if not attended to early enough(Rachel Nall 2016).As of the year 2016 of 36.7 million people were living with HIV, 1.8 million newly infected, 76.1 million infections since inception and 35 million deaths. New HIV infections for adults-15 years and above-for the year 2016 was 1.7 million having decreased from a high of 2.5 million in the year 2000. Eastern and South Africa had the highest PLHIV at 19.4 million where Kenya is a member of Eastern Africa(UNAIDS 2017b).

HIV in Kenya was first detected in 1984 and by 1996 it was the leading cause of mortality with a prevalence of 10.5% overwhelming the healthcare system and the economy at large. By the year 2016, Kenya had 1.6 million PLHIV, 62,000 are new infections and 36,000 AIDS related deaths. Adult HIV prevalence stands at 5.4%(AVERT 2017). According to Kenya Aids Progress report 2016 (NACC 2016), Homabay county is leading in both the HIV prevalence of more than 15% which is endemic and total HIV new infections which is more than 10,000. Funding of HIV response comes from Kenya(17%), externa donors(70%) and private individuals(13%)(National AIDS Control Council 2014).

Key populations are people who inject drugs, men who have sex with men, transgender persons, sex workers and prisoners. Globally they face much higher rates of HIV and AIDS than the general population and they are most at risk to contracting HIV (USAID 2017). It is estimated that 45% of the new infections occur between people in the key population and their immediate partners (UNAIDS 2016). HIV in Kenya is generally concentrated in some population, it is deeply

rooted among the general population and concentration of very high prevalence among the key populations. (National AIDS Control Council 2014).

Fishing communities have been identified as being among the highest risk groups to HIV infections in countries whose rates of HIV and AIDS prevalence is high. The vulnerability stems from the way of life of the fishermen and women which include the time they spend away from home fishing, the ready cash and commercial sex and their sub-culture of risk taking making them vulnerable to infections and difficult to reach with anti-retroviral therapy and continuous prevention efforts (Seeley & Allison 2005).In the former Nyanza province, Kenya, HIV prevalence is highest at 26.2% in the beach communities along Lake Victoria. Female fish traders are often migrants to the beaches and highly mobile and the *sex for fish* exchange has made the risk even higher(Camlin et al. 2013).

Antiretroviral agents(ART) when adhered to well has the potential of reducing HIV transmission by reducing the HIV in blood and genital secretions hence leading to viral load suppression (Cohen & Gay 2010).The various strategies put up by WHO has been undergoing changes as results from studies come out. For instance, back in the years ARVs were given to those whose CD4 were below 350, later through various researches it was changed to 500 and currently we have test and treat. Test and treat means that if a person is tested HIV positive, that person needs to be put on ARVs immediately regardless of the CD4 count and this has great benefits especially for low resource contexts where opportunistic infections are rampant(Eholié et al. 2016). HIV patients on ARVs and adhering well and whose viral load is undetectable has low rates of transmission to HIV negative partners (Rodger et al. 2016).

Even with the *test and treat* model, we still have 62,000 new infections country wide and more than 10, 000 in Homabay County which is leading in Kenya(NACC 2016). With this numbers

at the back of our mind, then something needs to be done to achieve zero new infections and the strategy is PrEP. Pre-exposure prophylaxis (PrEP) is a way for people whom are HIV negative but who are at risk of HIV infection to prevent being HIV infected by takin a pill (truvada) daily. PrEP reduces the chances of getting HIV through sex by over 90% and 70% among people who inject drugs(CDC 2018b). WHO recommended use of PrEP on Men who sex with men (MSM) in the year 2014 but it later changed to include all population who are at substantial risk of HIV infection. It is recommended that PrEP should be a priority for populations with HIV incidence of 3 per 100 person-years or higher. It should be an additional service to a comprehensive service pack that includes voluntary male circumcision, male and female condoms, lubricants, test and treat for HIV infected people among others (World Health Organization 2015).

Along the beaches of Lake Victoria the fishermen and fish traders, motorcycle riders commonly known as boda boda riders, widows, female sex workers and serodiscordant couples are at risk of HIV infection and these people require PrEP to enable them remain negative (Mack et al. 2014).More and more people are becoming eligible for PrEP and the unfortunate thing is that there are a significant number of people on PrEP who are becoming HIV positive (Liz Highleyman 2016).

In Kenya, the healthcare industry generates huge amounts of data which is complex data regarding the patients, the resources the hospital has, disease management among others. These huge amounts of data then requires powerful mining tools to support health care professionals in diagnosis, treatment and managing HIV/AIDS (Iseu et al. 2017). What has been observed over the years is that the volume of the data has been increasing but the proportion that the people collecting it understands is decreasing rapidly. Lying in those volumes of data is information that is very useful but never made explicit or taken advantage of. To make use of this volumes we then need

what we call data mining which is *the extraction of implicit, previously unknown, and potentially useful information from data* (Witten et al. 2011).

In the healthcare sector, there is a lot of data that is generated day in day out and the data is too voluminous to be processed by traditional methods, data mining comes in to provide the methodology and technology to turn the huge data that exist to useful information that can be used to better manage patients. Healthcare insurers use it too to detect anomalies in the transactions among others. Over the recent past, researchers have embraced the data mining techniques to help them in decision making, for instance in HIV-AIDS management, (Oliveira et al. 2017) employed data mining technique in HIV-AIDS surveillance system to determine the factors influencing delays in HIV-AIDS reporting within the Portuguese surveillance system.

Most organizations that offer healthcare in Kenya have electronic Medical records' systems and the data they get from the systems is always analyzed traditionally not knowing that there are other hidden patterns that can be found in the data that are very novel and that can assist them to better manage the healthcare system.

1.2 Problem statement

Kenya and indeed the world at large has been struggling to reduce HIV menace and so many strategies have been put in place to help curb the epidemic. So many families have been left with no one to fend for them as most of bread winners have died or are so sick that they can't work to provide for their family. A lot of money has been pumped for research purposes, care and treatment for the HIV positive patients which run to millions of shillings per year and this puts burden on the country's economy and even to the donors. About 68% of funds for HIV response in Kenya come from external donors, the Kenyan government contributes about 17% and private individuals give around 13%. The Kenyan government allocation to HIV and AIDS response more than doubled since the year 2006/7 to 2012/13 from USD 57.49 million to USD 153 million respectively but this isn't sustainable as funds from international donors have been dwindling since Kenya's reclassification as a middle-income country making international donors pull out and support low-income countries(National AIDS Control Council 2014).

Over the years, the government has implemented various strategies to curb the epidemic which has reduced but not to zero prevalence. The introduction of Pre-Exposure Prophylaxis to help the negatives remain negative is facing challenges of non-adherence. After all the effort that has been put into this, there is still a number clients on Pre-exposure (PrEP) that are still becoming HIV positive. In their study on *Cost-Effectiveness of Pre-Exposure Prophylaxis (PrEP) in Preventing HIV-1 Infections in Rural Zambia: A Modeling Study*,(Nichols et al. 2013) found out that adherence to PrEP is the key to its efficacy. What this means is that those who become positive after having been enrolled to PrEP are not adhering well to PrEP.

There exists models for measuring adherence to PrEP, one of these is the use plasma and peripheral blood mononuclear cells (PBMC) from the clients or using hair/dried blood spots

(DBS)(Haberer 2016). These are sure ways of measuring how long the client has been exposed to the drug and after knowing this then the concerned parties can take up measures meant to help those not taking PrEP regularly to do so and the end result is good adherence. The approach is good but it doesn't tell us prior that the patients is likely not to adhere so that measures are taken in advance to help the participant.

Data mining has been applied to data that has been collected and stored using electronic medical records (EMRs) and this has been able to predict those to be enrolled into HIV Pre-Exposure Prophylaxis (PrEP) (Krakower et al. 2016). It has even been applied in HIV management by predicting the CD4 status of a client by use of stored data (Haile Mariam 2015). With this massive data that exist, we don't have a model that can be able to tell us who these people who after enrollment are not likely to adhere to PrEP are so that measures can be put in place to help them adhere and hence increase efficacy of PrEP leading to low HIV prevalence.

The research study seeks to employ data mining techniques to build a model to help in predicting non-adherence to Pre-exposure Prophylaxis among clients enrolled in the Pre-exposure Prophylaxis.

1.3 Research Objectives

1.3.1 Main Objective

The main objective of the study is to develop a model for predicting non-adherence among Pre-Exposure Prophylaxis (PrEP) clients along the lake region.

1.3.2 Specific Objectives

i. To identify the factors affecting adherence among Pre-exposure Prophylaxis (PrEP) clients during the treatment period.

ii. To review and identify appropriate data mining method for predicting non-adherence to Pre-Exposure Prophylaxis (PrEP) among the enrolled clients

iii. To develop a model for predicting non-adherence to Pre-Exposure Prophylaxis (PrEP) among the enrolled clients

iv. To evaluate the model for predicting non-adherence to Pre-Exposure Prophylaxis (PrEP) among the enrolled clients

1.4 Research Questions

i. What are the factors that encourage Pre-Exposure Prophylaxis (PrEP) clients not to adhere to PrEP?

ii. What are the appropriate data mining method(s) that can be used to develop a model for predicting PrEP non-adherence among the enrolled clients?

iii. What is the appropriate model for predicting PrEP non-adherence among the enrolled clients?

iv. How will the model for predicting PrEP non-adherence among the enrolled clients be evaluated?

1.5 Motivation of the study

HIV in Kenya is a big menace and a lot has been done to eradicate it. Poverty in HIV stricken areas is increasing day and night as all the resources are being directed to care and treatment and those who are sick cant fend for their families, children are dropping out of school to take care of their parents. The government together the international community has directed a lot of funds for HIV eradication, funds that could have been used to improve the livelihood of the people.

A lot of data has been collected in relation to HIV but less has been done to dig dip into the huge volumes of data to see if there are special patterns that can be used to better manage the

epidemic. The current strategy of using PrEP to ensure the negatives remain negative is a good strategy and HIV status change from negative to positive while on PrEP should not occur. The Kenyan government together with donor funded projects that are concerned with HIV care and treatment has done a lot to ensure that those tested HIV positive are put on care and treatment to ensure that they remain alive and healthy and to reduce transmission to some extent. Introduction of PrEP is aimed at ensuring the HIV negatives remain negative and together with care and treatment for the HIV positives, the country is able to achieve zero new HIV infections.

With the facts at hand, there are clients on PrEP that are turning HIV positive after some time on treatment and one contributing factor to this is non-adherence to PrEP which reduces its efficacy.

1.6 Significance of the study

The research findings from the study has the potential of providing useful insights to HIV prevention programs for those in charge of policy formulation, the domain experts and also those providing the services. For instance putting more emphasis on those likely to become HIV positive even after taking PrEP will be a great strategy

The findings has the potential of increasing enrollment of those at a high risk of HIV infection to PrEP and with this, the population is likely to benefit since most of those at a high risk of HIV will have low infection rates thus reducing new infections

Data mining technology has not been widely explored and the findings of the study has the potential of adding knowledge to the field of research and better utilization of the data that has already been collected and lying idle. The findings can also encourage more researchers to explore deeper in the same area and in the end add more knowledge.

1.7 Scope of the study

The study will be based in Suba region along the lake areas. The target population are the most at risk for HIV infection and they include fishermen and fish traders, motorcycle riders commonly known as boda boda riders, widows, female sex workers and serodiscordant couples who are enrolled in PrEP in 9 health facilities. The researcher will use data that has been collected on the target group.

CHAPTER 2

LITERRATURE REVIEW

2.1 Introduction

The current study is a model for predicting non-adherence among Pre-Exposure Prophylaxis clients and this chapter is primarily focused on going through the available literature to see what has been done thus far, how it relates to the study area and whether there are any gaps that can be picked in the course of going through the literature and how the gaps can be filled by the current study.

2.2 Data mining

Data mining is a process that involves various steps in order to end up with the knowledge generation. The figure below summarizes the steps.

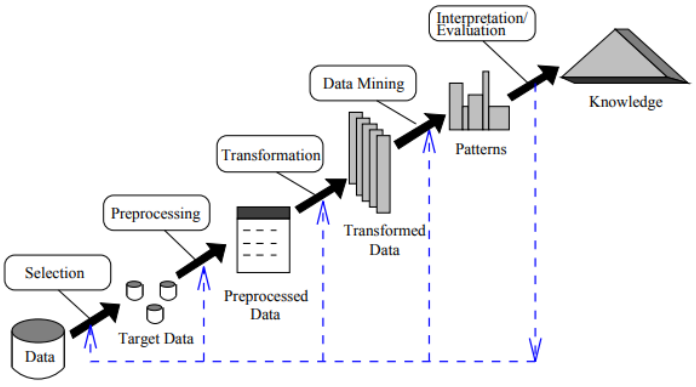


Figure 1: Data mining process (Yongjian Fu 2011)

Data mining has various tasks including classification, prediction, time-series analysis, association, clustering, and summarization. All the 6 tasks can be grouped to either predictive or descriptive data mining tasks. Predictive data mining tasks comes up with a model from the available dataset to predict unknown values of another dataset of interest like a doctor trying to

diagnose a disease based on testing results. Descriptive data mining tasks finds data that describes patterns and comes up with new interesting information like using data to see which products are bought together.

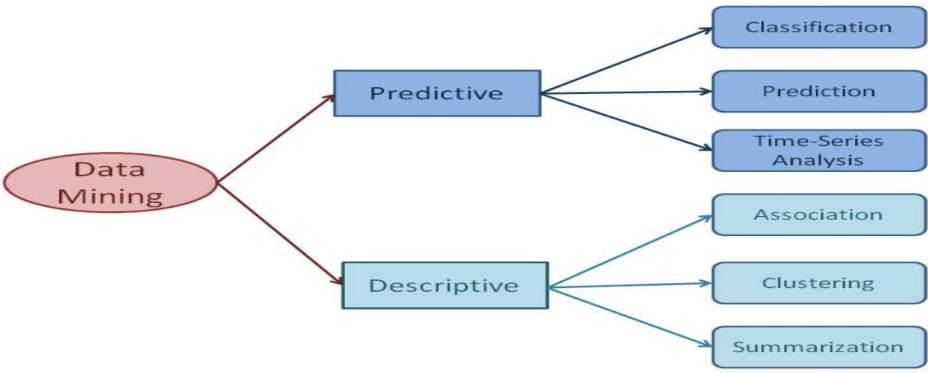


Figure 2: Data mining tasks

2.3 Factors affecting adherence to PrEP

Despite PrEP's efficacy, its implementation as a strategy to prevent new HIV infections has been slow and one of the challenges to its implementation is adherence. Various studies have shown that PrEP is more effective if taken daily and consistently and this is possible if adherence is at a high notch. This then makes adherence one of the most contributing factor that can lead to effectiveness of PrEP or its failure. There are various factors that can hinder adherence of clients that have been initiated on PrEP(Haire 2015)

2.2.1 PrEP related stigma

PrEP is associated to high sexual activity and this makes it to have multiple stigmas. One of the stigma is that PrEP is related to HIV (HIV also has its own stigma) and the fact PrEP is viewed as an alternative to condom use. In both contexts for example most clients who have been enrolled to PrEP care often feel shy to disclose that they are on PrEP and to this extent most of them are so much worried how their parents will react if they ever discovered that they are on

PrEP, their religion also plays an important part in their stigma. For those in relationships, they are worried how their partners will take if they found out that they are taking PrEP especially that they will definitely question their trust(Haire 2015).

Stigmatization and criminalization of the key population has made it difficult for them to access HIV related services. In the presence of supportive environment devoid of stigma and criminalization among the key populations, there is strong evidence of decrease in new HIV infections. For example, a study done in St. Petersburg, Russia between the years 2006 to 2012 found a 73% reduction in HIV seroprevalence among street youth 15-19 years of age who were involved in drug injection use. The reduction in HIV seroprevalence was possible due to implementation of extensive support programs and socio-economic improvements(Kornilova et al. 2017)

In a clinical trial in Kenya, (Van der Elst et al. 2013) found out that social impacts of PrEP included stigma, being implicated in rumors and experiencing difficulties in relationships due to misinterpretation of their HIV status. Being in such instances, one will definitely opt out of PrEP and the end result may be HIV infection.

2.2.2 Side effects of PrEP

According to (GILEAD 2017), PrEP has various known side effects that clients taking PrEP should be aware about and report in case they have similar symptoms, they include but not limited to kidney problems including kidney failure, too much lactic acid in blood severe liver problems and bone problems. Common side effects may include abdominal pain, headache and decreased weight. Clients are likely to discontinue with PrEP if they have any of the side effects. The anticipation of the side effects may also make a client on PrEP discontinue almost

immediately after starting the drug like worry about kidney and liver which may manifest themselves later in life (Arnold et al. 2017)

2.4 Related work

There are various studies that have been done that have tried to look at why there are HIV status changes due to adherence failures among people who have been enrolled in PrEP and the ways to avert that. The whole idea of PrEP is to help those that are negative to remain negative and if this is achieved then we will not have new infections since those that are on PrEP will not transmit to the other uninfected individuals and they will not be infected. The other national guideline of test and treat will take care of those infected to the extent that they will achieve viral suppression meaning that they can't transmit hence achieving *"U=U Undetectable = Untransmitable* (UNAIDS 2017a)

In their study on *Predictors of First-Year Medication Adherence in a National Cohort of Veterans Initiating Pre-Exposure Prophylaxis (PrEP) to Prevent HIV Infection,* (Ohl et al. 2017)linked older age, male sex, white race, diabetes, and lack of substance use to high adherence. Almost half of the cohort, 44%, stopped PrEP based on a 120 day gap in coverage. Poverty level, residence (urban, rural) and the US region did not affect high adherence. In their conclusion they noted that the differences in adherence related to age, gender and race suggest disparities in PrEP effectiveness in routine practice.

Data mining as a technology has also been applied in the field of health and specifically in HIV prevention and care. In their paper on *Application of Data Mining Techniques on Pre ART Data: The Case of FelegeHiwot Referral Hospital* (Tarekegn & Sreenivasarao 2016)recommended use of unpruned J48 classifier for predicting clients' eligibility for ART. This was after experimenting J48 decision tree and Naïve Bayesian classifier using both10-fold

cross validation and percentage split (66%) test modes. The classification rules obtained revealed that; Women who are pregnant are eligible for ART if their CD4 count is less than or equal to 350 and age is greater than or equal to 5years, all infants whose age is less than two years old are eligible for ART, those whose age is between two tears and five years old and CD4 percent is less or equal 25 are eligible for ART and the patients whose age is between two years and five years and whose CD4 count is less than or equal to 750 are eligible for ART.

Another study by (Haile Mariam 2015) on *Application of Data Mining Techniques for Predicting CD4 Status of Patients on ART in Jimma and Bonga Hospitals, Ethiopia* recommended use of J48 decision as best compared to PART rule induction algorithm. From the study's outcome, J48 algorithm was effective in the prediction of CD4 status of patients that are on ART. The model that was built shows that Eligible reason, ART status, ART start year, OA weight, OAWHO stage, Current regimen, Family planning, Functional status, Marital status, Past ARV were the most determining factors for predicting CD4 status. From this study, it is clear that mining has been used on available data to predict the variables that affect a dependent variable. With such outcomes, then CD4 status of patients can be determined from the data that has been collected hence saving the duration of making decisions that affect the health of the patient.

Social media is another area that has been targeted in the fight against HIV. A study by (Breen et al. 2016) on *Mining Pre-Exposure Prophylaxis Trends in Social Media* used Natural Language Processing (NLP) to monitor data from twitter which is a social media to identify PrEP discussion trends. From the results, it was clear that PrEP and HIV discussion dynamics are able to be identified over time and positive or negative sentiments can be assigned on the PrEP-

related tweets. They concluded that public health professional can monitor PrEP related discussions and use this to identify strategies to improve HIV prevention through PrEP.

Other studies have been done that targets ways of assessing adherence from individuals who are already on PrEP, one of such studies done by (Baxi et al. 2015) on *Comparing the Novel Method of Assessing PrEP Adherence/Exposure Using Hair Samples to Other Pharmacologic and Traditional Measures*compared the different measures of drug exposure that included self-report, medication event monitoring system (MEMS)-caps openings, and TFV/FTC levels in hair and other biomatrices. They recommended the use of hair to measure exposure to PrEPsince *hair levels correlated strongly with PBMC, plasma concentrations, and MEMS-caps openings*(Baxi et al. 2015).

Looking at most of the literature of the previous studies, there has been a lot that has been done to ensure PrEP works. For instance the assessing PrEP adherence using hair is a good way of knowing who never adhered to PrEP but the problem is for those who have not adhered to PrEP most probably could have contracted HIV and there is nothing much that can be done. On a positive note, such metric for assessing PrEP adherence can be useful if the tests are done regularly at close intervals so that we are able to identify those that are not adhering well and probably come up with ways to do thorough follow ups to ensure they adhere, with this then we are sure of PrEP efficacy.

We have huge volumes of data that can be utilized to ensure PrEP works and this can be possible if data mining technology is employed, but from the literature there has not been much that has been done in exploiting this data using data mining so as to make sense out of it.

The big question that comes up, as a gap, from the literature is how then we can know these people who are not going to adhere to PrEP beforehand so that most resources are directed towards them to ensure they have 100% adherence. There is data that has been collected from the various PrEP studies and other healthcare facilities, this data needs to be utilized to ensure PrEP works, a model is needed to identify the specific individuals that are likely not to adhere to PrEP immediately they are being enrolled so that we can ensure they take the pill. Waiting until a PrEP client seroconverts is not an option.

2.5 Conceptual framework

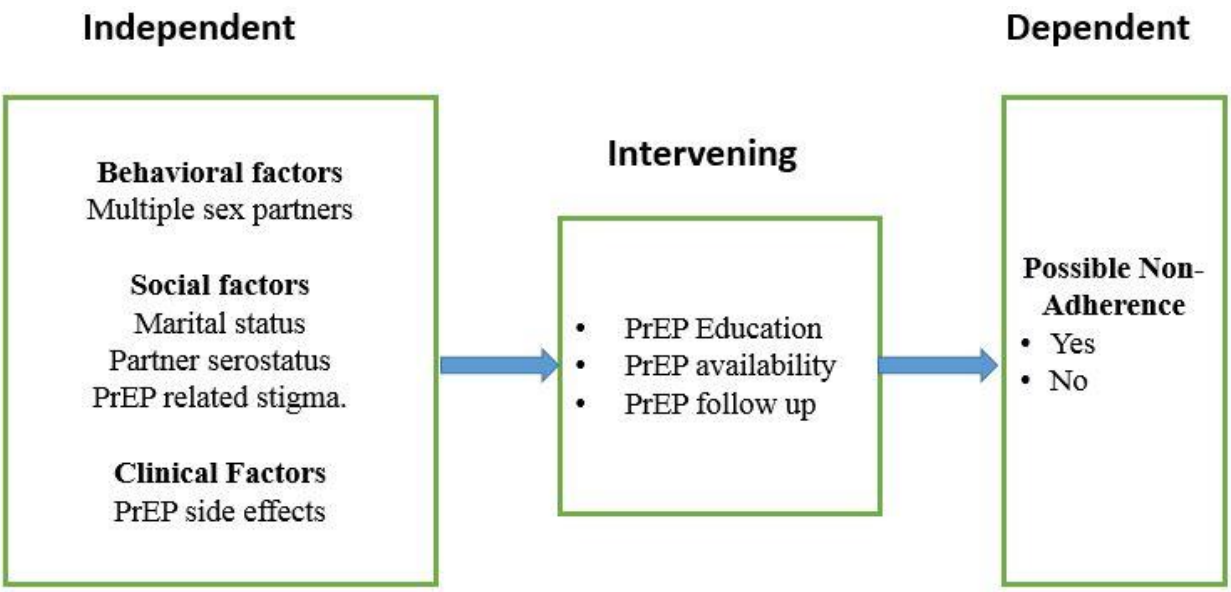


Figure 3: Conceptual Model for Non-Adherence to PrEP

2.6 Operationalization of the variables

Variables	Sub-variables	Indicators(Symptoms)	Values(data)
Factors affecting adherence to PrEP	Behavioral factors	Multiple sex partners	Yes No
	Social factors	Marital status	Single Married Monogamous Married polygamous
		Partner serostatus	Positive Negative
		PrEP related stigma	Yes No
	Clinical factors	PrEP side effects	Yes No
Non-Adherence prediction		Non-adherence	Possible Non-Adherence Yes No

Table 1: Operationalization of variables

CHAPTER THREE
RESEARCH METHODOLOGY

3.1 Introduction

The research methodology chapter is aimed at discussing the research design, the target population, sample size, sampling techniques and their justifications, describe the methods of data collection and how the data will be analyzed.

3.2 Research Design

The researcher used review of literature, review of the data and data mining methodology in order to accomplish the objectives of the study.

To achieve objective 1 review of the study data was used, objective 2 a review of literature was used and data mining methodology was used to achieve objective 3 and 4.

Data mining methodology can be defined as a system that consist of rules, procedures, methods and principles that guide the data mining process(Berry & Linoff 2009). Data mining methodology is an 11 step methodology designed to a successful data mining project.

Data mining is never a linear process but one that has nested loops in such a manner that things that are learnt later on in the process can lead to going back to the early ones to make changes. The 10 step data mining methodology is as shown below and as adopted from (Berry & Linoff 2009) and slightly modified by linking *get to know data* to *fix problems with the data* and thus removing *create a model set* process (Beatrice &Mwendia 2017)

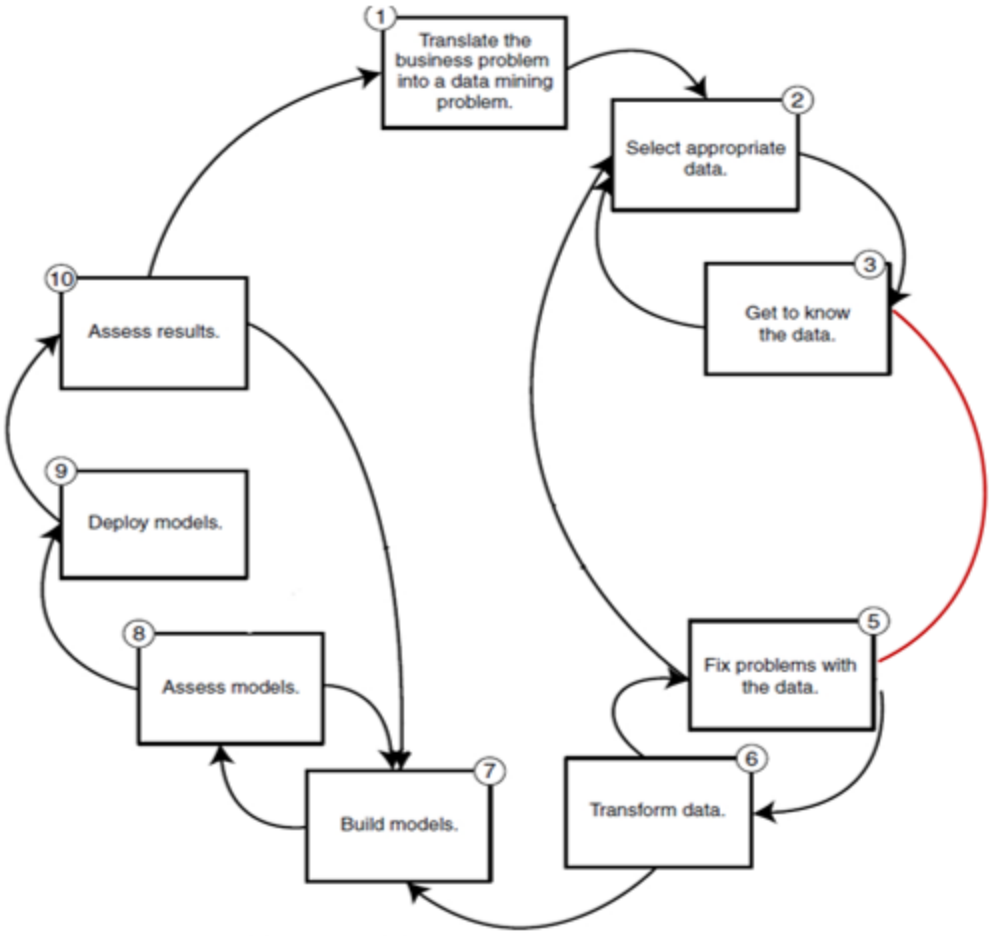


Figure 4: Data mining methodology(Berry & Linoff 2009)

The PrEP data to be used in this process has been de-identified to maintain the confidentiality of the participant data. The process will involve removing of names of the participants and their phones numbers.

Step 1: Define business problem

A well-defined business problem leads to a better design of a data mining model that will help solve the problem. Whoever wants to solve any business problem using data mining has to fully understand what the problem is and break it down to more specific issues so that more specific

goals to sort the issue can be tackled. The researcher was able to clearly define what the problem was using the problem statement and how data mining can be applied on the available data to try and solve the problem.

Step 2: Translate business problem into a data mining problem.

A successful translation of a business problem to a data mining problem requires the problem to be formulated in one of the data mining tasks. This was accomplished by looking at the problem and since the problem required a predictive model then the best technique to use was classification.

Step 3: Select Appropriate data

For one to use data mining, the data has to be available with variables that are needed. The PrEP data is available in the individual health facilities in Suba region. The data includes enrollment and follow up data. The variables were selected from the two i.e enrollment and follow up datasets using a unique identifier called searched as depicted in the images below. The data is always updated per visit, it is available and it is historically accurate.

	A	B	C	D	E	F	G	H	I	J
1	tdate	prepid	searchid	vweek	mvisit	lvisit	curriskhiv	hivrsk3months	hivantibody	rhivantibody
2	15/05/2018	P64033	38219548097-2	4	0	2	1	1	1	2
3	04/04/2017	P764122	38219548094-3	4	1	2	0	0	0	
4	21/02/2018	P64026	38219548083-1	4	0	1	1	1	1	2
5	28/03/2017	P964111	38219548078-2	4	0	2	1	0	1	2
6	23/03/2017	P964109	38219548069-1	4	0	2	1	0	1	2
7	23/03/2017	P964108	38219548063-2	4	0	2	1	1	1	2
8	24/04/2018	P64027	38219501065-1	4	0	2	1	1	1	2
9	07/05/2018	P64039	38219501063-1	4	0	2	1	1	1	2
10	23/03/2017	P764141	38219501056-1	4	0	2	0	0	0	
11	03/05/2018	P64040	38219501055-1	4	0	2	1	1	1	2
12	03/05/2018	P64037	38219501041-5	4	0	2	1	1	1	2
13	23/03/2017	P964106	38219501039-6	4	0	2	0	1	1	2
14	24/04/2018	P64034	38219501039-2	4	0	2	1	1	1	2
15	20/04/2017	P65016	38218597069-2	4	0	1	1	1	1	2
16	09/03/2017	P964043	38218597055-8	4	0	1	1	1	1	2
17	09/03/2017	P964042	38218597055-1	4	0	1	1	1	1	2
18	13/03/2017	P964060	38218597052-2	4	0	2	1	1	1	2
19	09/03/2017	P964039	38218597039-2	4	0	1	1	1	1	2

Figure 5: Follow up visits data

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	kptdate	Occupation	searchid	dob	sex	marital	polygamy	wives	require_testing	fpdrawn	rapid_test1	rapid_test2	rapid_tes	hivstat	malar	syphi
2	05/04/2018 13:05	fishing	38219548097-2	01/01/1992	1	5	-9	-9	1	1	0	-9	-9	0	-6	-6
3	05/04/2018 11:21	fishing	38219548094-3	02/10/1992	1	1	-9	-9	1	1	0	-9	-9	0	-6	-6
4	05/04/2018 10:28	fishing	38219548091-5	21/08/1995	1	1	-9	-9	1	1	0	-9	-9	0	-6	-6
5	05/04/2018 14:42	Street Vendor	38219548083-1	10/08/1973	2	2	1	3	1	1	0	-9	-9	0	-6	-6
6	05/04/2018 15:26	Housewife WIDOW	38219548068-1	16/04/1978	2	3	-9	-9	0	1	4	4	4	1	-6	-6
7	03/04/2018 14:13	Drug Abuse	38219548067-13	08/04/1967	1	2	1	2	1	1	0	-9	-9	0	0	-6
8	05/04/2018 11:40	fishing	38219548062-2	01/01/1993	2	2	0	-9	0	1	4	4	4	1	-6	-6
9	03/04/2018 10:41	Discordant	38219501065-1	01/01/1962	1	2	0	-9	1	1	0	-9	-9	0	-6	-6
10	05/04/2018 13:12	fishing	38219501063-1	01/01/1972	1	2	0	-9	1	1	0	-9	-9	0	-6	-6
11	05/04/2018 15:38	fishing	38219501056-2	01/01/1987	2	3	-9	-9	0	1	4	4	4	1	-6	-6
12	05/04/2018 11:18	Housewife WIDOW	38219501055-1	01/01/1963	2	3	-9	-9	1	1	0	-9	-9	0	-6	-6
13	05/04/2018 12:23	Housewife WIDOW	38219501054-12	15/06/1993	2	3	-9	-9	0	1	4	4	4	1	-6	-6
14	05/04/2018 14:25	fishing	38219501053-1	01/01/1983	1	2	1	2	0	1	4	4	4	1	-6	-6
15	05/04/2018 11:32	fishing	38219501050-2	01/01/1990	2	2	0	-9	1	1	0	-9	-9	0	-6	-6
16	05/04/2018 11:09	fishing	38219501050-1	01/01/1983	1	2	0	-9	1	1	0	-9	-9	0	-6	-6
17	05/04/2018 10:56	fishing	38219501046-1	01/01/1982	1	2	0	-9	0	1	4	4	4	1	-6	-6
18	05/04/2018 13:53	WIFE OF A FISHERMAN	38219501046-2	01/01/1983	2	2	0	-9	0	1	4	4	4	1	-6	-6
19	05/04/2018 14:38	fishing	38219501045-1	01/01/1986	1	2	0	-9	1	1	0	-9	-9	0	-6	-6
20	06/04/2018 16:52	fishing	38219501044-2	01/01/1970	2	2	0	-9	1	1	0	-9	-9	0	-6	-6
21	05/04/2018 10:25	Adoscents	38219501041-4	01/01/1994	1	1	-9	-9	1	1	0	-9	-9	0	-6	-6
22	05/04/2018 14:03	fishing	38219501041-5	01/01/1998	1	1	-9	-9	1	1	0	-9	-9	0	-6	-6
23	03/04/2018 15:17	Street Vendor	38219501039-12	11/06/1960	2	2	0	-9	1	1	0	-9	-9	0	-6	-6

Figure 6: Enrollment data

Step 4: Get to know the data

Before building models, it is important to go through the selected data just to familiarize yourself with it and through this one is able to discover a lot that may raise questions which could have

otherwise not come up, anomalies in the data may also be noticed and corrections made in the process. This will be accomplished through Computer Human Interaction (CHI).

Step 5: Fix problems with the data

Problems in the selected data has to be fixed. Normally all data that we intend to use to build models will most likely be dirty and hence have problems. The problems vary per the technique used. For instance, missing values and outliers may not cause trouble in decision trees but will definitely be a problem in neural networks. The required variables were merged from the two data datasets to a single file. Missing and incomplete values were fixed as shown below.

	B	C	D	E	F	G	H	I	J	K	L	M
1	prepid	DOB	searchid	Gender	Marital	Polygamy	lvisit	Occupation	curriskhi	hivrsk3month	hivantibod	rhivantibod
3	P52005	01/01/1995	32101522019-2	1	1	-9	2	farmer	1	1	1	2
4	P52006	01/01/1996	32101526001-2	1	1	-9	2	Transport	1	1	1	1
5	P52007	01/01/1993	32101512007-2	1	1	-9	2	Student	1	1	1	
6	P52008	05/10/1986	32106532057-2	2	2	0	1	FISHERFOLK	1	1	1	2
7	P52009	15/06/1978	32108524013-1		3	-9	2	fishing	1	1	1	2
8	P52010	15/06/1988	32102527018-2	2	2	1	2	fishing	1	1	1	2
9	P52011	01/01/1970	32111504003-1	2	2	1	2	fishing	0	0	1	2
10	P52012	01/01/1992	32108535008-1	2	2	0	2	Student	1	1	1	2
11	P52014	01/01/1994	32108514012-2		1	-9	2	fishing	1	1	1	2
12	P52015	01/01/1990	32107533009-1	2	2	1	2	Fishing	1	1	1	2
13	P52016	01/01/1994	32109508008-3	2	2	0	1	fishing	1	1	1	0
14	P52017	01/01/1978	32111509013-2	2	2	0	1	-Student	0	0	1	2
15	P52018	01/01/1985	32107534008-2	2	2	1	2	fishing	1	1	1	2
16	P52019	01/01/1995	32108535003-2	1	2	0	2	-Student	1	1	1	2
17	P52020	01/01/1996	32108535011	1	2	0	2	fishing	1	1	1	2
18	P52021	01/01/1983	32107547021-1	M	2	0	2	fishing	1	1	1	2
19	P52022	01/01/1986	32107547021-2	2	2	0	2	fishing	1	1	1	2
20	P52023	01/01/1993	32115547014-3	2	1	-9	2	-Student	0	0	1	2
21	P52024	01/01/1989	32104507014-2	2	2	0	1	fishing	1	1	1	2
22	P52025	01/01/1990	32106531030-3	2	2	0	2	Hotel worker	1	1	1	2

Figure 7: Merged dataset with incomplete values

B	C	D	E	F	G	H	I	J	K	L
prepid	DOB	searchid	Gender	Marital	Polygamy	lvisit	Occupation	curriskhi	hivrsk3month	hivantibod
P52005	01/01/1995	32101522019-2	1	1	-9	2	farmer	1	1	1
P52006	01/01/1996	32101526001-2	1	1	-9	2	Transport	1	1	1
P52007	01/01/1993	32101512007-2	1	1	-9	2	Student	1	1	1
P52008	05/10/1986	32106532057-2	2	2	0	1	fishing	1	1	1
P52009	15/06/1978	32108524013-1	2	3	-9	2	fishing	1	1	1
P52010	15/06/1988	32102527018-2	2	2	1	2	fishing	1	1	1
P52011	01/01/1970	32111504003-1	2	2	1	2	fishing	0	0	1
P52012	01/01/1992	32108535008-1	2	2	0	2	Student	1	1	1
P52014	01/01/1994	32108514012-2	1	1	-9	2	fishing	1	1	1
P52015	01/01/1990	32107533009-1	2	2	1	2	Fishing	1	1	1
P52016	01/01/1994	32109508008-3	2	2	0	1	fishing	1	1	1
P52017	01/01/1978	32111509013-2	2	2	0	1	Student	0	0	1
P52018	01/01/1985	32107534008-2	2	2	1	2	fishing	1	1	1
P52019	01/01/1995	32108535003-2	1	2	0	2	Student	1	1	1
P52020	01/01/1996	32108535011-3	1	2	0	2	fishing	1	1	1
P52021	01/01/1983	32107547021-1	1	2	0	2	fishing	1	1	1
P52022	01/01/1986	32107547021-2	2	2	0	2	fishing	1	1	1
P52023	01/01/1993	32115547014-3	2	1	-9	2	Student	0	0	1
P52024	01/01/1989	32104507014-2	2	2	0	1	fishing	1	1	1
P52025	01/01/1990	32106531030-3	2	2	0	2	Hotel worker	1	1	1

PREP fuv

Figure 8: Merged dataset with incomplete values fixed

Step 6. Transform the data

After fixing problems within the data, the data has to be converted to a format that is appropriate for mining. This may include things like adding some derived variables to make the data more understandable, removing outliers, binning numeric variables and grouping classes for categorical variables. After cleaning the merged data, variables like age was derived from date of birth by using a formula =DATEDIF(C2,NOW(),"Y") in excel, agegroup was also introduced and was derived from the age variable to come up with 3 age groups i.e 15-24 as youth, 25-49 as adult and 50 and above as elderly using the formula =IF(AND(C2>14,C2<25),"Youth",IF(AND(C2>24,C2<50),"adult",IF(C2>49,"elderly"))). Numeric data was also converted to text like in gender where 1 represents male and 2 female.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	prepid	searchid	community	age	age_category	agegroup	Gender	Marital_Status	visit_location	occupation	curriskhiv	hivrsk3months	rhivantibody
2	P52007	32101512007-2	Nyamrisra	25	25-49	adult	Male	single	offsite	student	Yes	Yes	Neg
3	P99209	32101512013-1	Nyamrisra	33	25-49	adult	Female	Married Polygamous	Clinic	housewife	Yes	Yes	Neg
4	P52035	32101512014-1	Nyamrisra	43	25-49	adult	Male	Married Monogamous	Clinic	Fishing	Yes	Yes	Neg
5	P98051	32101512017-3	Nyamrisra	35	25-49	adult	Male	single	offsite	Fishing	Yes	Yes	Neg
6	P72157	32101522001-2	Nyamrisra	26	25-49	adult	Female	Married Polygamous	offsite	student	No	No	Neg
7	P99001	32101522005-4	Nyamrisra	22	15-24	Youth	Male	single	offsite	Fishing	Yes	Yes	Neg
8	P99053	32101522006-1	Nyamrisra	59	50 +	elderly	Male	Married Monogamous	offsite	Farmer	No	No	Neg
9	P99052	32101522010-1	Nyamrisra	74	50 +	elderly	Male	Married Monogamous	offsite	Fishing	Yes	Yes	Neg
10	P99007	32101522014-3	Nyamrisra	24	15-24	Youth	Male	Married Monogamous	offsite	Fishing	No	No	Neg
11	P52067	32101522018-1	Nyamrisra	67	50 +	elderly	Male	Married Monogamous	offsite	Farmer	Yes	Yes	Neg
12	P52005	32101522019-2	Nyamrisra	23	15-24	Youth	Male	single	offsite	no job	Yes	Yes	Neg
13	P52006	32101526001-2	Nyamrisra	22	15-24	Youth	Male	single	offsite	transport	Yes	Yes	Pos
14	P52041	32101526021-2	Nyamrisra	38	25-49	adult	Female	Married Monogamous	offsite	business	Yes	Yes	Neg
15	P52160	32102521003-2	Nyamrisra	25	25-49	adult	Female	Married Monogamous	Clinic	Farmer	Yes	Yes	Neg
16	P99054	32102521005-1	Nyamrisra	28	25-49	adult	Male	Married Monogamous	offsite	Fishing	Yes	Yes	Neg
17	P52159	32102521005-2	Nyamrisra	24	15-24	Youth	Female	Married Monogamous	offsite	Fishing	Yes	Yes	Neg
18	P99157	32102525003-2	Nyamrisra	23	15-24	Youth	Female	Married Monogamous	Clinic	business	Yes	Yes	Neg
19	P52058	32102525039-1	Nyamrisra	40	25-49	adult	Male	Married Polygamous	offsite	Fishing	No	No	Neg
20	P99158	32102525046-2	Nyamrisra	26	25-49	adult	Female	Married Monogamous	offsite	business	No	No	Neg
21	P99003	32102527002-4	Nyamrisra	22	15-24	Youth	Male	single	offsite	student	Yes	Yes	Neg
22	P99208	32102527006-1	Nyamrisra	57	50 +	elderly	Male	Widowed	offsite	Farmer	Yes	Yes	Neg
23	P52010	32102527018-2	Nyamrisra	30	25-49	adult	Female	Married Polygamous	offsite	Fishing	Yes	Yes	Neg
24	P99050	32103528011-1	Nyamrisra	38	25-49	adult	Male	Married Polygamous	offsite	Fishing	Yes	Yes	Neg
25	P52158	32103528011-2	Nyamrisra	25	25-49	adult	Female	Married Monogamous	offsite	business	No	No	Neg
26	P52221	32103528012-2	Nyamrisra	36	25-49	adult	Female	Married Monogamous	offsite	Farmer	No	No	Neg
27	P99048	32103529001-1	Nyamrisra	35	25-49	adult	Male	Married Monogamous	Clinic	Fishing	No	No	Neg
28	P52042	32103529004-1	Nyamrisra	29	25-49	adult	Female	Married Monogamous	Clinic	business	Yes	Yes	Neg
29	P62055	37305531013-1	tommboya	47	25-49	adult	Male	Married Polygamous	offsite	business	Yes	Yes	Neg
30	P964103	38216578031-1	sena	38	25-49	adult	Female	Married Monogamous	Clinic	government worker	Yes	Yes	Neg
31	P52163	32103529018-5	Nyamrisra	18	15-24	Youth	Female	single	Clinic	student	Yes	Yes	Neg
32	P52031	32104507014-1	Nyamrisra	33	25-49	adult	Male	Married Monogamous	offsite	Fishing	Yes	Yes	Neg

Figure 9: Transformed data

	A	B	C	D	E	F	G	H	I	J	K
1	prepid	gender	age_c	marital	pc	location_vi	hivrsk3mont	occupation	highbloodpressu	monthsoutsidecommuni	non_adheren
2	P52001	male	Youth	single	no	clinic	yes	no_job/hwife	no	seven-nine	yes
3	P52003	male	Youth	single	no	offsite	no	fishing	no	never	yes
4	P52004	male	adult	married	yes	clinic	no	fishing	no	never	yes
5	P52005	male	Youth	single	no	offsite	no	no_job/hwife	no	never	yes
6	P52006	male	Youth	single	no	offsite	yes	transport	no	never	yes
7	P52007	male	Youth	single	no	clinic	yes	no_job/hwife	no	never	yes
8	P52008	female	adult	married	no	offsite	no	fishing	no	never	yes
9	P52009	female	adult	widowed	no	offsite	no	fishing	no	never	yes
10	P52010	female	adult	married	yes	clinic	yes	fishing	no	never	no
11	P52011	female	adult	married	yes	offsite	yes	fishing	no	never	no
12	P52013	male	adult	married	no	offsite	no	fishing	no	four-six	yes
13	P52014	male	Youth	single	no	offsite	yes	fishing	no	never	no
14	P52015	female	Youth	married	yes	offsite	yes	fishing	no	never	no
15	P52016	male	adult	married	no	offsite	yes	fishing	no	never	no
16	P52017	female	adult	married	no	offsite	yes	fishing	no	one-three	no
17	P52018	male	adult	married	yes	offsite	yes	fishing	no	never	no
18	P52019	male	Youth	married	no	offsite	yes	fishing	no	never	no
19	P52020	male	Youth	married	no	offsite	yes	fishing	no	never	no
20	P52021	male	adult	married	no	offsite	yes	fishing	no	never	no
21	P52022	female	adult	married	no	offsite	yes	fishing	no	never	no

prep

Figure 10: Data ready for mining

Step 7: Build models

Building models as a data mining process takes up little time in a data mining project since modern data mining software has automated the process. In directed data mining, like what the study is using, the training set is used to provide an explanation of the independent or target variable using the independent or input variables. There are various tools that are employed in data mining process and those that intersect multiple reviews are Tanagra, Orange, KNIME, Weka, and Rapid Miner(Wimmer & Powell 2015).(Chandan Goopta 2014) highlights six open source data mining toolkits and they include Orange, Weka, Rapid Miner, JHepWork, and KNIME. Another website that talks of tools is (AUZA 2010) which discusses 5 of the Best Free and Open Source Data Mining Software that include Rapid Miner, Weka, Orange, R, KNIME, and NTLK. The current study used WEKA to build models

	Orange	Tanagra	Rapid Miner	KNIME	R	Weka
K-means Clustering	Yes	Yes	Yes	Yes	Yes	Yes
Association Rule Mining	Yes	Yes	Yes	Yes	Yes	Yes
Linear Regression	Yes	Yes	Yes	Yes	Yes	Yes
Logistic Regression	Yes	Yes	Yes	Yes	Yes	Yes
Naïve Bayesian Classifiers	Yes	Yes	Yes	Yes	Yes	Yes
Decision Tree	Yes	Yes	Yes	Yes	Yes	Yes
Time Series Analysis	No	No	Some	Yes	Yes	Yes
Text Analytics	Yes	No	Yes	Yes	Yes	Yes

Big Data Processing	No	No	No	No	Yes	Yes
Visual Work Flows	Yes	Yes	Yes	Yes	No	Yes

Table 2: Data mining tools comparison*(Wimmer & Powell 2015)*

Step 8: Assess the models

This step determines if the model is working right. At this point the questions like the accuracy of the model and confidence of the predictions from the model comes into place. The researcher used k-fold cross validation and in this case k was 10.

Step 9: Deploy the models

This is the step where the model is moved from the mining environment to the scoring environment. Depending on the development environment, the process may be hard or easy. The hard part comes from the fact that some development may have been done in a special environment where the software can't run anywhere else and the miner has to move it and recode it in another programming language to make it possible for it to run.

Step 10: Assess the results

In this step, actual results are compared against what you expect as results. The actual measure of data mining is the value of actions taken as a result of data mining. Having a measure of lift can help you choose a model and use of these models can help you choose how to apply the results from the models. In addition to using lift, it is also important to do a measure at the field.

Step 11: Begin again

The completion of a data mining project raises more questions without answers. This is true since there comes new relationships that were not thought of before and this creates a new question to answer hence starting of data mining process again

3.3 Target population

The target population are clients who have been enrolled in PrEP Suba region of Homabay County that are at a higher risk of HIV infection. Homabay County has a population of approximately 963794 people, Adult HIV prevalence of 25.7%, the study targets clients who are 15 years and above and that have been enrolled for PrEP within the health facilities in the region.

3.4 Sampling and sampling procedure

The researcher used all the 971 records for the research.

3.5 Data collection methods

The researcher intends to use data from Sustainable East Africa Research in Community Health research study that is conducting HIV testing among the most at risk of HIV infection. Those found to be HIV negative and accept to be enrolled to take PrEP are linked to facilities nearby to continue with PrEP. Data about the participants is collected from the time they are enrolled and continues for every visit they make to collect drugs. The researcher intends to use this data that has been collected for the period from June 2016 to June 2018.

CHAPTER FOUR
RESEARCH FINDINGS AND DISCUSSIONS

4.1 Introduction

This chapter discusses the findings of the study. The researcher sought to develop a model to predict non-adherence among clients enrolled in Pre-Exposure Prophylaxis in Suba region. To achieve this, the researcher sought to understand the factors that leads PrEP clients not to adhere to the pill during the treatment period, establish appropriate data mining method for predicting non-adherence among PrEP clients, develop a model for predicting possible non-adherence among PrEP clients and evaluate the model for its effectiveness.

4.2 Research findings

4.2.1 Factors affecting adherence among Pre-exposure Prophylaxis clients during the treatment period

The researcher was interested in finding out the factors that leads PrEP clients to forfeit taking PrEP drug. Looking at the data, a number of factors were identified and discussed below.

- ***Self-risk assessment***

People who feel at risk of HIV infection are more likely to adhere to PrEP and those who don't feel at risk for a number of months to come are likely not to adhere to PrEP and this is clear when looking at the data, those that didn't feel at risk for the following 3 months were found not to adhere at 81% and those who felt at risk adhered well to PrEP at 73%

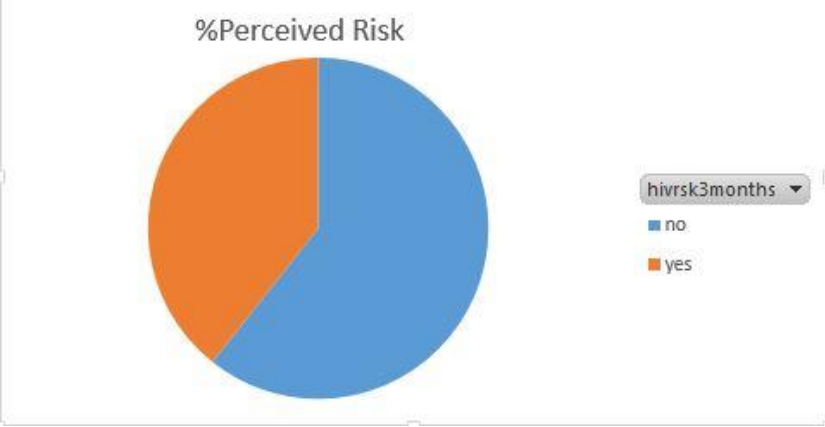


Figure 11: Self-risk assessment

- ***Demographic characteristics***

Age of a participant enrolled in PrEP could be a reason why he/she will take the daily pill or ignore for some days. The age can be categorized into youth, adult and elderly. The data shows adults (70%) having a higher non-adherence percentage followed by the youth (23%) and finally the elderly(7%)

Gender of a participant, male or female, can be another factor that influences adherence or non-adherence to PrEP. From the data males (59%) have a higher percentage of non-adherence followed by female (41%).

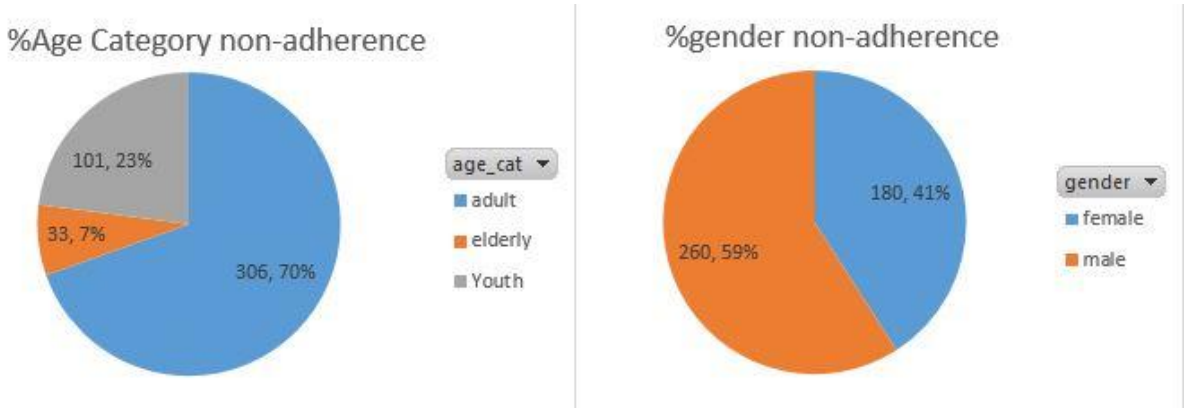


Figure 12: Demographic factors

- ***Social Factors***

These factors include the relationship status of the client which can be divided into those married in monogamous marriages and polygamous families, the single, divorced/separated and the widowed.

The data shows that those in monogamous marriages (49%) leads in non-adherence followed by the single (30%), polygamous marriages (15%), widowed (4%) and finally (2%)

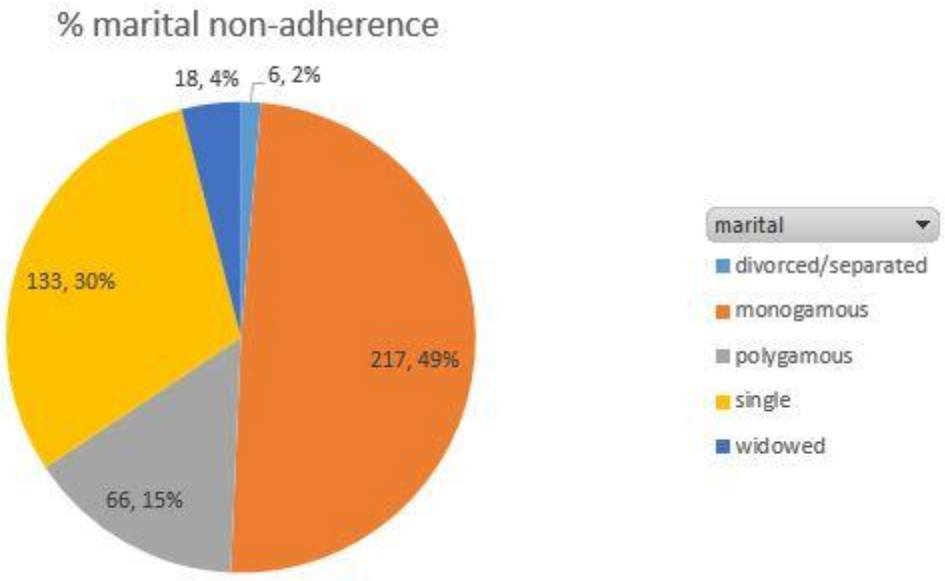


Figure 13: Social factors

4.2.2 Appropriate data mining method for predicting non-adherence to Pre-Exposure Prophylaxis (PrEP)

The researcher sought to find out the appropriate data mining method for predicting non-adherence to PrEP. Through the review of literature, classification was the chosen method for building a predictive model for non-adherence among PrEP clients.

From the literature review, predictive models are built using classification. In classification, the independent variables are used to sort the data into distinct classes. It can be

implemented using various algorithms which include decision trees, rule induction, k-nearest neighbors, naïve Bayesian, artificial neural networks, and support vector machines.

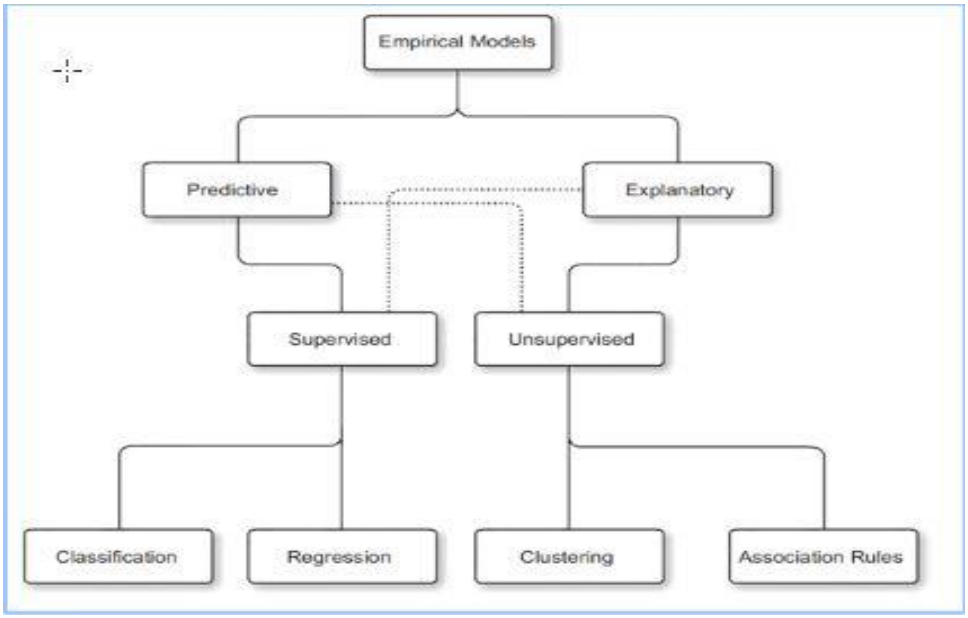


Figure 14: Models(Elkan 2010)

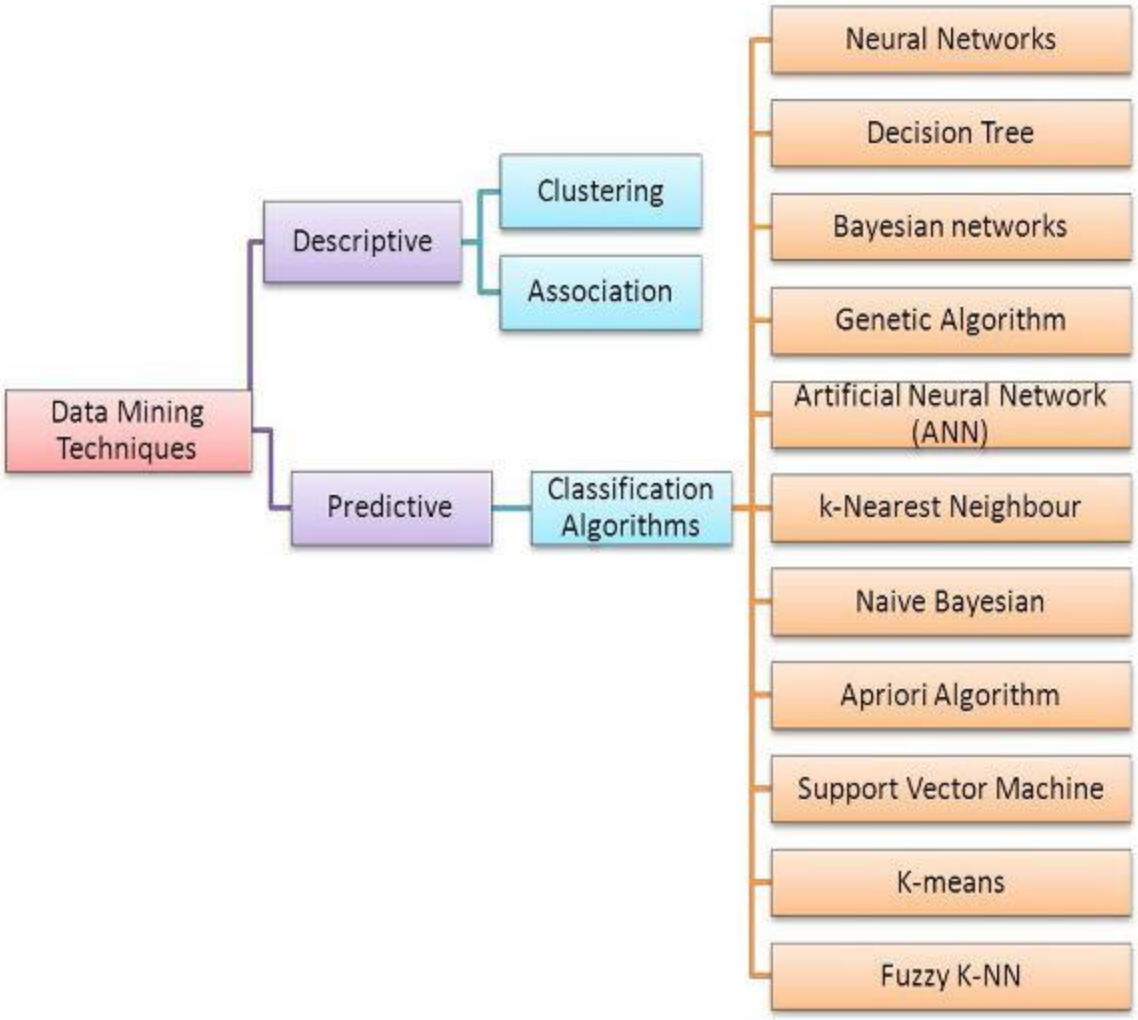


Figure 15: Different techniques in Healthcare domain ((Kharya 2012)

Various studies have used classification for prediction purposes and most of these are in health and specifically in HIV. Below is a table showing the various studies and the algorithms employed.

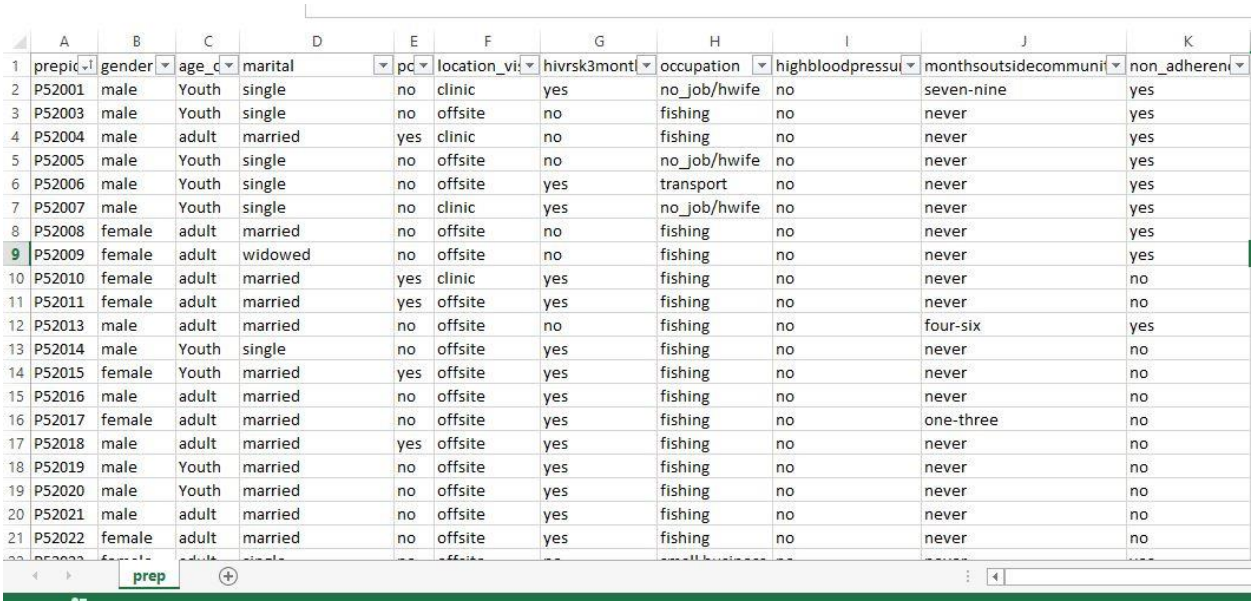
Algorithm	**Author, year**	**Outcome**
Random tree, J48, neural networks and logistic regression	(Hailu 2011)Comparing Data Mining Techniques in HIV Testing Prediction	Predicted whether an individual was being tested or not for HIV given that wealth index, education level, residence, HIV related stigma, knowledge on family planning, knowledge

		related to HIV, region, age group, and risky sexual behavior as inputs with an accuracy of 96%, 79%, 78% and 74% respectively
C4.5 Decision Tree, IBK, Naïve Bayes, Support Vector Machines and Random Forest	(Tucker et al. 2015)Machine learning classification of medication adherence in patients with movement disorders using non-wearable sensors.	C4.5 algorithm had the most consistent predictive performance, achieving the highest performance followed by IBK and SVM as second and third respectively
random forests, random tree, logistic model trees, J48, rotation forest , SVM, radial basis function network, Bayesian network, Naive Bayes , multiple layer perceptron and simple classification and regression tree (simple CART).	(Karanasiou et al. 2016)Predicting adherence of patients with HF through machine learning techniques	combination of different classifiers with different FS measures revealed that features expressing medical condition of the patient, as well as, the medication treatment can act as predictors for the adherence

Table 3: Classification Studies

4.2.3 Developing a model for predicting non-adherence among pre-exposure prophylaxis

The figure below shows the data in csv form that has been cleaned and transformed ready for mining to build the predictive model.



	A	B	C	D	E	F	G	H	I	J	K
1	prepid	gender	age_c	marital	pc	location_vi	hivrsk3mont	occupation	highbloodpressu	monthsoutsidecommuni	non_adheren
2	P52001	male	Youth	single	no	clinic	yes	no_job/hwife	no	seven-nine	yes
3	P52003	male	Youth	single	no	offsite	no	fishing	no	never	yes
4	P52004	male	adult	married	yes	clinic	no	fishing	no	never	yes
5	P52005	male	Youth	single	no	offsite	no	no_job/hwife	no	never	yes
6	P52006	male	Youth	single	no	offsite	yes	transport	no	never	yes
7	P52007	male	Youth	single	no	clinic	yes	no_job/hwife	no	never	yes
8	P52008	female	adult	married	no	offsite	no	fishing	no	never	yes
9	P52009	female	adult	widowed	no	offsite	no	fishing	no	never	yes
10	P52010	female	adult	married	yes	clinic	yes	fishing	no	never	no
11	P52011	female	adult	married	yes	offsite	yes	fishing	no	never	no
12	P52013	male	adult	married	no	offsite	no	fishing	no	four-six	yes
13	P52014	male	Youth	single	no	offsite	yes	fishing	no	never	no
14	P52015	female	Youth	married	yes	offsite	yes	fishing	no	never	no
15	P52016	male	adult	married	no	offsite	yes	fishing	no	never	no
16	P52017	female	adult	married	no	offsite	yes	fishing	no	one-three	no
17	P52018	male	adult	married	yes	offsite	yes	fishing	no	never	no
18	P52019	male	Youth	married	no	offsite	yes	fishing	no	never	no
19	P52020	male	Youth	married	no	offsite	yes	fishing	no	never	no
20	P52021	male	adult	married	no	offsite	yes	fishing	no	never	no
21	P52022	female	adult	married	no	offsite	yes	fishing	no	never	no

prep

Figure 16: Data ready for data mining in WEKA

After the data was subjected to WEKA and using J48 algorithm, the tree that was produced is as shown below.

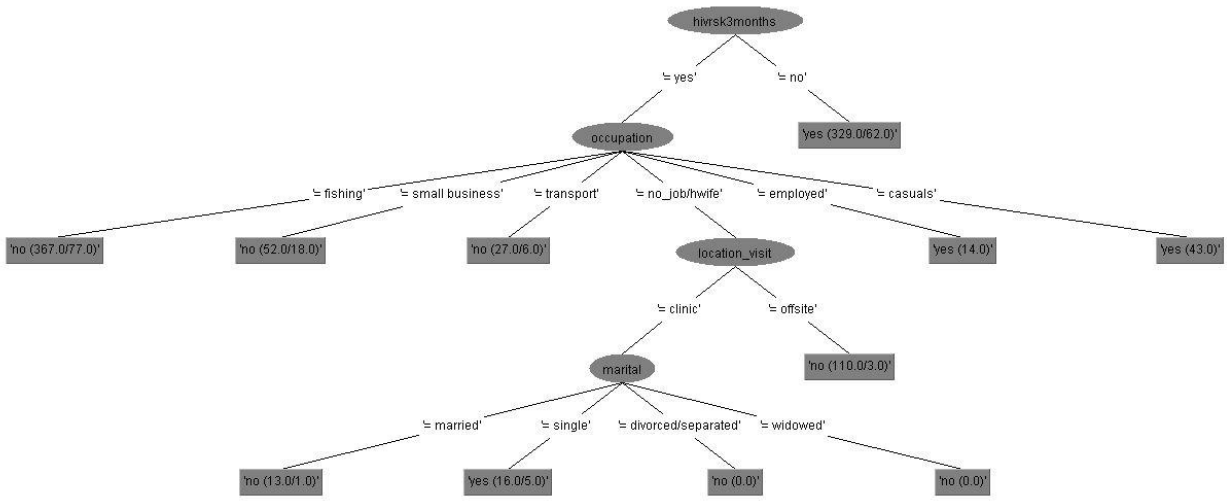


Figure 17: Decision tree for predicting non-adherence to PrEP

From the tree above, it is clear that those people who perceive themselves to be at risk for the following three months and are either in the fishing, transport or small business occupation are likely to adhere to PrEP or those that perceive themselves to be at risk in the following three months but have no job and their drug delivery points are at their points of choice (offsite) or they prefer the clinic as their pick up points and are either married, widowed, divorced or separated.

In the contrary, those people who don't perceive themselves as being at risk of HIV infection for the following three months are most likely not to adhere to PrEP. The same applies to those that

perceive themselves as being at risk of HIV infection in following three months and they are either casuals or employed or they have no job and are single.

The researcher extracted the following from the decision tree:

If *HIV risk for 3 months=No* THEN *Possible non-adherence*

If *HIV risk for 3 months=Yes, Occupation=no_job/housewife, Location of Visit=Clinic, marital=single* THEN *Possible Non-adherence*

If *HIV risk for 3 months=Yes, Occupation= {employed, Casuals}* THEN *Possible Non-adherence*

If *HIV risk for 3 months=Yes, Occupation={fishing, small business, transportation}*THEN *Possible adherence*

If *HIV risk for 3 months=Yes, Occupation=no_job/housewife, Location of Visit=clinic, marital={married, divorced/separated, widowed}* THEN *Possible adherence*

If *HIV risk for 3 months=Yes, Occupation=no_job/housewife, Location of Visit=offsite* THEN *Possible adherence*

4.2.4 Evaluating the model for predicting non-adherence to Pre-Exposure Prophylaxis (PrEP)

The researcher used 10 fold cross validation during the model creation and below measures were used to test the model accuracy.

1. **Kappa statistic.** This is the measure of how the class labels and the attributes are related. It is the correlation of attributes and the class labels just like the correlation coefficient. A value of zero (0) means there is no relationship between the class labels and the attributes

whereas a value approaching one (1) shows a strong relationship. The Kappa statistic from the decision tree was 0.6199 and 6125 for Naïve Bayes which falls in the bracket of 0.60-0.79 and this can be interpreted as *substantial agreement* as summarized in the Kappa statistic measurements and their interpretation table below.

Measurement (value or range)	**Interpretation**
0.0	Complete disagreement (do not resemble each other).
0.40 to 0.59	Moderate agreement
0.60 to 0.79	Substantial agreement
above 0.80	Outstanding agreement
1.0	Complete agreement (items strongly resemble each other)

Table 4: Kappa statistic measurements and their interpretation

2. **Mean Absolute Error (MAE).**This is the sum of absolute errors divided by the number of predictions. It measures the set of predicted values to the actual value and what this means is that MAE measures how close the predicted model is to the actual model. A small value of MAE means that a better prediction of the model. From the evaluation of the model the MAE was 0.2805 for J48 and 0.2921 for Naïve Bayes which is a small value.
3. **Root Mean Square Error (RMSE).** This is the square root of sum of squares error divided number of predictions. It measures the difference between values predicted by the model and the actual values observed. The smaller the value of RMSE the better

accurate the model. From the evaluation of the model, the value of RMSE was 0.3793 for J48 and 0.3763 for Naïve Bayes which is below the half mark of 0.5

4. **F-Measure.** This is a combined measure of precision and recall, they are both used together for F-Measure. From the model evaluation, the F-Measure value was 0.812 for J48 and 0.808 for Naïve Bayes
5. **Receiver operating Characteristic (ROC) curve.** This is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The ROC area value was 0.837 for J48 and 0.861 for Naïve Bayes
6. **True Positive:** The number of items correctly retrieved as belonging to the positive class. TP value in J48 is 0.813 and 0.809. These values are high and all tend to one (1) which means that most of the items were correctly classified.
7. **False positive:** Items that are incorrectly labeled as belonging to positive class (Type1 error). FP value in J48 is 0.196 and Naive Bayes is 0.202, this shows that very few cases were misclassified compared to those that were correctly classified.

Below are the evaluation results of the decision tree.

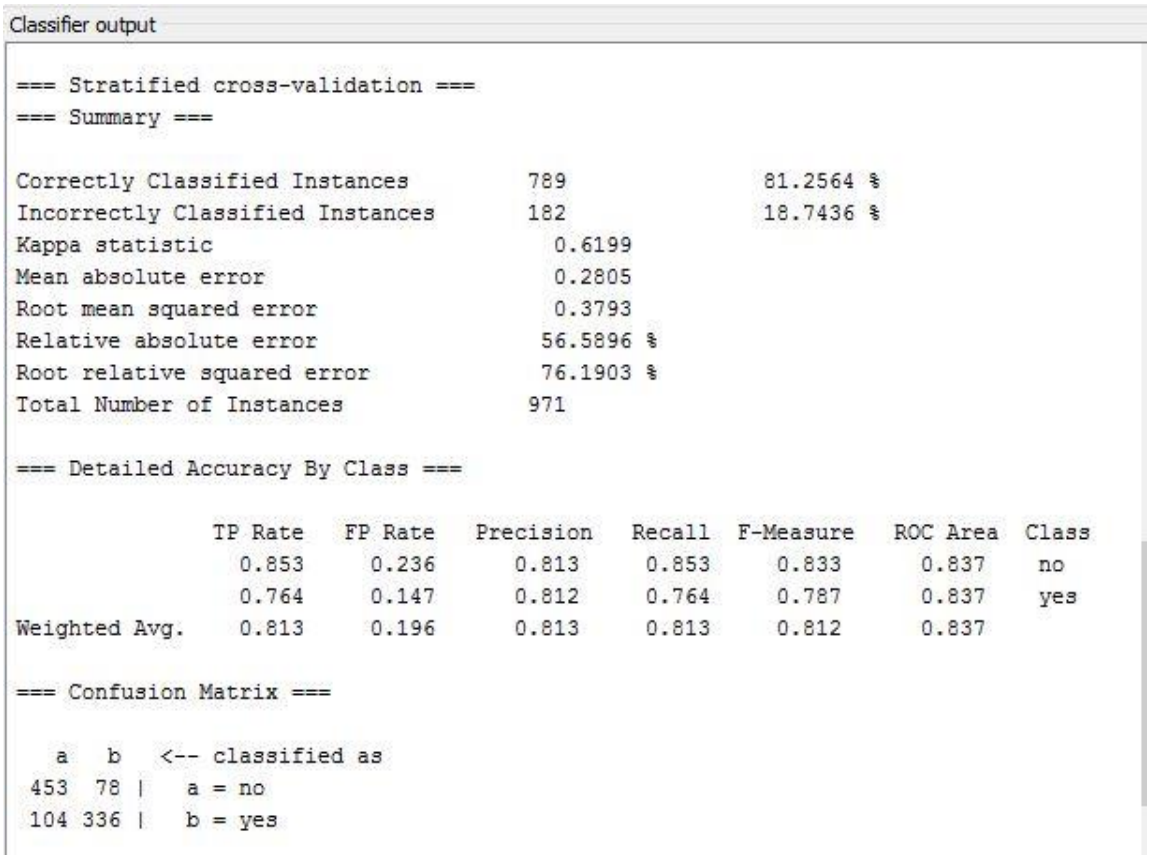
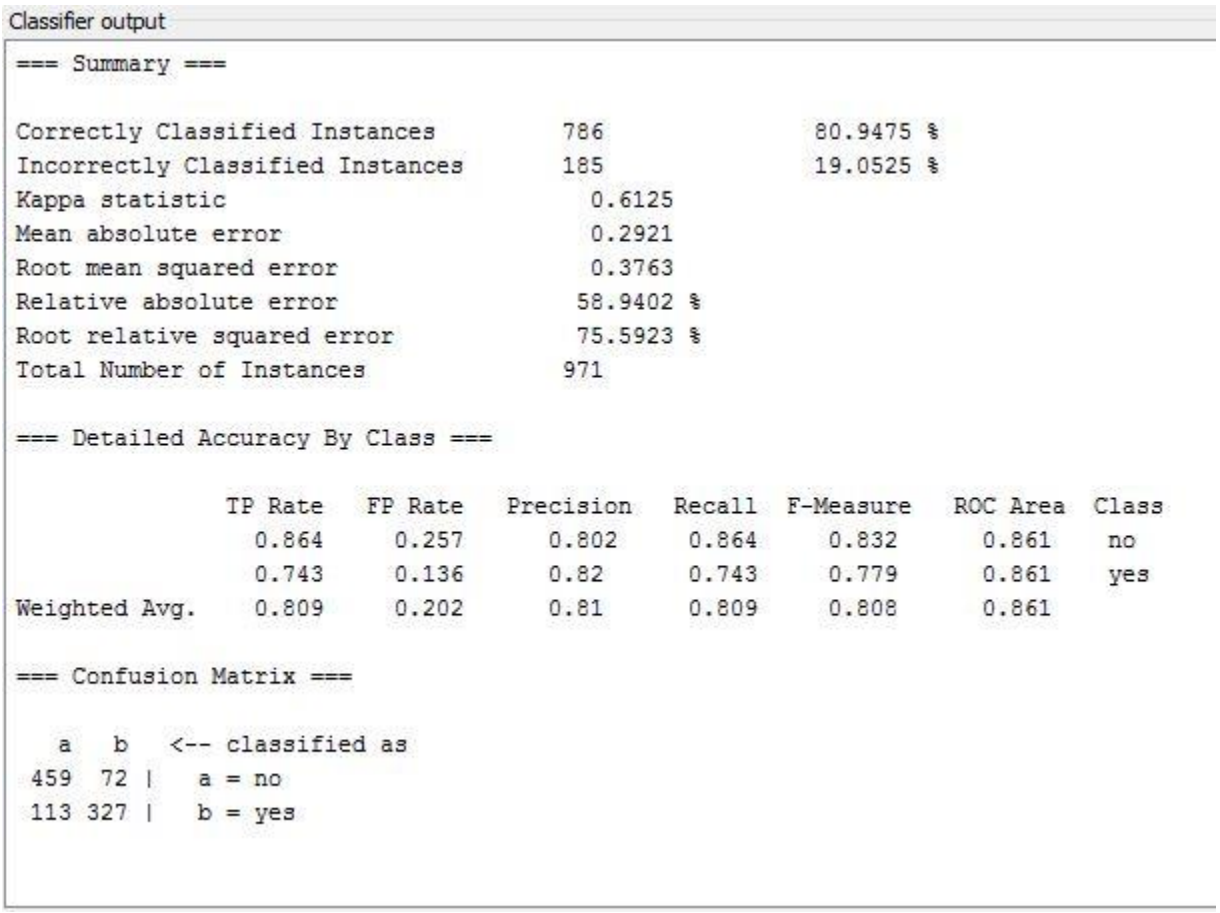


Figure 18: J48 Decision tree evaluation results

From the results from J48 algorithm, the Kappa statistic is substantial and the values of MAE and RMSE are both lower and ROC area and F-measure values tends to one (1) which is a perfect test and overall this indicates a dependable model



Classifier output

```
=== Summary ===

Correctly Classified Instances         786               80.9475 %
Incorrectly Classified Instances       185               19.0525 %
Kappa statistic                          0.6125
Mean absolute error                      0.2921
Root mean squared error                  0.3763
Relative absolute error                 58.9402 %
Root relative squared error             75.5923 %
Total Number of Instances              971

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.864     0.257      0.802     0.864     0.832      0.861    no
                 0.743     0.136      0.82      0.743     0.779      0.861    yes
Weighted Avg.    0.809     0.202      0.81      0.809     0.808      0.861

=== Confusion Matrix ===

   a   b   <-- classified as
 459  72 |   a = no
 113 327 |   b = yes
```

Figure 19: Naive Bayes evaluation results

From the Naïve Bayes evaluation results, the value for Kappa statistic indicates a substantial agreement, the values of MAE and RMSE are both smaller and ROC Area and F-Measure both tends to one (1) which is a perfect test.

Below is a summary of the various evaluation results as shown above

Evaluation Result	Algorithm I (Decision Tree)	Algorithm II (Naïve Bayes)
Kappa Statistics	**0.6199**	0.6125
Mean Absolute Error	**0.2805**	0.2921
Root Mean Square Error	0.3793	**0.3763**
Relative absolute Error	**56.59%**	58.94%
F-Measure	**0.812**	0.808
ROC Area	0.837	**0.861**

Table 5: Evaluation Results from various algorithms

Looking at the evaluation results from the 2 algorithms, it is clear that the model has a satisfactory inter rater reliability as depicted in the Kappa statistic of 0.6199 and 0.6125 in J48 decision tree and Naïve Bayes respectively and the values falls in the category of substantial agreement of 0.60-0.79. The findings from the study shows an accurate model as evidenced from the F-Measure of 0.812 and 0.808 in J48 decision tree and Naïve Bayes respectively. The area under the curve is used as a measure of effectiveness and from the evaluation results a value of 0.837and 0.861 in decision tree and Naïve Bayes respectively shows effectiveness. Most of the items are correctly labeled as belonging to the positive class at >0.8 for both J48 and Naïve Bayes and those not correctly labeled at <0.21

4.3 Discussion of Results

The results from this study shows four main factors that affect adherence to PrEP and these include self-identified risk to HIV infection in the following three months, the occupation of the PrEP client, the location where the medication is delivered during every visit and the marital status of the client.

From the results, people who take drugs from the clinic could be adhering well to medication since they are not stigmatized and those that choose to have PrEP taken to them may not really appreciate what PrEP does or they may be stigmatized and could be fearing to take PrEP.

The results from this study clearly shows that the way an individual feels at risk of HIV infection has a greater determination of whether he/she will adhere to PrEP medication. Part of the enrollment criteria was self-identified risk of HIV infection and looking at the results, then it is high time that self-risk assessment be given more weight. A study by (Koss et al. 2018)on *Early adopters of HIV pre-exposure prophylaxis in a population-based combination prevention study in rural Kenya and Uganda* identified various adopters to PrEP uptake and that included self-identified risk to HIV infection. What this means then is that one will intentionally not adhere to PrEP just because the individual feels that he or she is not at risk of HIV infection even after starting the medication.

In their study of *Factors influencing uptake of pre-exposure prophylaxis: some qualitative insights from an intervention study of men who have sex with men in China*, (Liu et al. 2017) identified perception of low HIV risk, mistrust of the national PrEP program, and concerns of side effects as reasons for not wanting to use PrEP, from the current study, identified perception of low risk also contributes to one not to adhere to PrEP.

A study done by (Ohl et al. 2017) that linked older age, male sex, white race, diabetes, and lack of substance use to high adherence, this is contrary to what the results from the study shows. Sex and age appear in both studies but they seem to have had a greater impact on adherence to PrEP which is contrary to what the study results show. The difference could be due to the difference between the two countries where the studies were done, which is Kenya and the United States of America. The two countries are different, the way of life of the countries are different Kenya being a developing nation and United States of America being a developed country. The researcher carried out the study on blacks only but looking at the (Ohl et al. 2017) study, it was a study of mixed races.

The study by (Tarekegn & Sreenivasarao 2016) on *Application of Data Mining Techniques on Pre ART Data: The Case of FelegeHiwot Referral Hospital* recommended use of unpruned J48 classifier after experimenting with Naïve Bayes at 10-fold and 66% split which is almost similar with the current research that has recommended use of J48 classifier after comparing it with Naïve Bayes at 10-fold cross validation though not much difference between the two.

(Joshi et al. 2015) compared J48 and Naïve Bayes in a diabetes dataset and after validation Naïve Bayes perfomred better than J48 in the time taken to build the model, the number of correctly classified instances and prediction accuracy and therefore concluded that Naïve Bayes perform better on diabetes dataset. This is different from the resulst from the study which shows J48 perfoming better but with no much differences between J48 and Naïve Bayes.

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDATIONS

7.1 Introduction

This chapter presents the conclusions from the study findings and the recommendations for policy and future research.

7.2 Conclusions

This study was intended to develop a model for predicting non-adherence among Pre-Exposure Prophylaxis clients in Suba region. Consequently various factors that affect adherence were noted that include HIV self-risk assessment, demographic factors such as gender and age group and social factors which is marital status.

A client's perception of low risk of HIV, their occupation employed/casual/no job, location of visit clinic and marital status single predicted one's non-adherence.

The fact that PrEP has the potential to reduce the HIV epidemic is one breakthrough in HIV prevention and for the potential to be realized then adherence to PrEP has to be put into consideration in order to make PrEP successful in fighting HIV, once this is achieved then we expect reduction in HIV infections. The use of models to predict those likely to take up PrEP is one thing that has been done and will likely increase enrolments and the use of model for predicting non-adherence to PrEP will likely ensure good adherence and this will increase PrEP efficacy.

Use of data mining on client data has been successful in coming up with the predictive model and use of J48 algorithm was more effective compared with Naïve bayes using 10-fold cross validation.

Using the model and being able to put more emphasis on this subgroup of clients who are likely not to adhere will help in PrEP adherence and thus improve PrEP efficacy. The emphasis include PrEP education, making it easy for PrEP access and continuous follow up.

7.3 Contributions of the study

1. The study has been able to show that data mining can be used on existing data to develop models that can aid in management of clients on Pre-Exposure Prophylaxis (PrEP)
2. The study has identified the kind of clients who are likely not to adhere to Pre-Exposure Prophylaxis, this is likely to help in the development of interventions to specifically target this kind of people.
3. The results from this study combined with what other researchers have been able to come up with can help policy implementers to come up with strategies to increase Pre-Exposure Prophylaxis (PrEP) enrollments and at the same time ensure adherence is at a high notch. With this the success of PrEP will surely be seen.

7.4 Recommendations for future research

1. Further research should be conducted to include data mining tools in Electronic Medical Records (EMR) which can get the data, mine it, predict and visualize the results for interpretation and action.
2. Further research on the subject area can be done in different areas with different economic activities, culture and tribe so that the results can be compared and possibly generalized.
3. The data for the PrEP clients should be shared across implementing partners so that any client on PrEP can visit any location that is offering PrEP and be able to take drugs from there. The data shared should also enable the care givers to understand the incoming client and be able to offer services that are tailored to suit him or her.
4. A similar study should be done among the HIV positive clients and who are on care but this time it should be about adherence to ARVs but applying data mining methodology

5. Since the study results has shown the profiles for non-adherent clients, the policy makers should therefore come up with specific guidelines that are taking into account this subgroup who are likely not to adhere to PrEP to better manage them and achieve high adherence and better efficacy of PrEP.
6. The study recommends use of other tools like Tenagra and KNIME and compare results.

APPENDICES

4.1 Research Schedule

Table 6: Research schedule

Activity	Begin	End
Proposal drafting and writing	January 2018	May 2018
Proposal submission and presentation	June 2018	June 2018
Work in progress 1 writing, submission and presentation	June 2018	July 2018
Data preparation and analysis	July 2018	August 2018
Thesis defense	September 2018	September 2018
Final report submission	October 2018	October 2018

4.2 Budget

Table 7: Budget

Activity	**Unit Cost(KSh.)**	**Total cost(KSh.)**
Transport to various health facilities	4000	4,000
Transport to and from KCA, 5 trips	2000	10,000
Internet	2500	2,500
Printing and binding of thesis, 5 copies	1000	5,000
	Totals	21,500

REFERENCES

Arnold, T. et al., 2017. Social, structural, behavioral and clinical factors influencing retention in Pre-Exposure Prophylaxis (PrEP) care in Mississippi J. A. Caylà, ed. *PLOS ONE*, 12(2), p.e0172354. Available at: http://dx.plos.org/10.1371/journal.pone.0172354 [Accessed April 29, 2018].

AUZA, 2010. 5 of the Best Free and Open Source Data Mining Software | TechSource. Available at: http://www.junauza.com/2010/11/free-data-mining-software.html [Accessed August 16, 2018].

AVERT, 2017. HIV and AIDS in Kenya. , pp.1–7. Available at: http://www.avert.org/aids-uganda.htm [Accessed April 18, 2018].

Baxi, S.M. et al., 2015. Comparing the Novel Method of Assessing PrEP Adherence/Exposure Using Hair Samples to Other Pharmacologic and Traditional Measures. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 68(1), pp.13–20. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25296098 [Accessed May 7, 2018].

Berry, M. & Linoff, G.S., 2009. *Data Mining Techniques : Theory and Practice Course Notes*,

Breen, P. et al., 2016. Mining Pre-Exposure Prophylaxis Trends in Social Media. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, pp. 214–221. Available at: http://ieeexplore.ieee.org/document/7796907/ [Accessed May 7, 2018].

Camlin, C.S., Kwena, Z.A. & Dworkin, S.L., 2013. Jaboya vs. jakambi: Status, negotiation, and HIV risks among female migrants in the "sex for fish" economy in Nyanza Province, Kenya. *AIDS education and prevention : official publication of the International Society for AIDS Education*, 25(3), pp.216–31. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23631716 [Accessed April 20, 2018].

CDC, 2018a. PrEP | HIV Basics | HIV/AIDS | CDC. Available at: https://www.cdc.gov/hiv/basics/prep.html [Accessed May 21, 2018].

CDC, 2018b. PrEP 101/PrEP 101. Available at: https://www.cdc.gov/hiv/pdf/library/factsheets/prep101-consumer-info.pdf [Accessed April 22, 2018].

Chandan Goopta, 2014. Six of the Best Open Source Data Mining Tools - The New Stack. Available at: https://thenewstack.io/six-of-the-best-open-source-data-mining-tools/ [Accessed August 16, 2018].

Cohen, M.S. & Gay, C.L., 2010. Treatment to prevent transmission of HIV-1. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 50 Suppl 3(0 3), pp.S85-95. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20397961 [Accessed April 20, 2018].

Eholié, S.P. et al., 2016. Antiretroviral treatment regardless of CD4 count: the universal answer to a contextual question. *AIDS research and therapy*, 13, p.27. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27462361 [Accessed April 22, 2018].

Elkan, C., 2010. *Predictive analytics and data mining*, Available at:

http://www.mendeley.com/research/data-mining-and-predictive-analysis/.

Van der Elst, E.M. et al., 2013. High Acceptability of HIV Pre-exposure Prophylaxis but Challenges in Adherence and Use: Qualitative Insights from a Phase I Trial of Intermittent and Daily PrEP in At-Risk Populations in Kenya. *AIDS and Behavior*, 17(6), pp.2162–2172. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23080358 [Accessed April 29, 2018].

GILEAD, 2017. TRUVADA for PrEP™ Side Effects. Available at: https://www.truvada.com/what-is-truvada/side-effects [Accessed April 29, 2018].

Haberer, J.E., 2016. Current concepts for PrEP adherence in the PrEP revolution: from clinical trials to routine practice. *Current opinion in HIV and AIDS*, 11(1), pp.10–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26633638 [Accessed April 30, 2018].

Haile Mariam, T., 2015. Application of Data Mining Techniques for Predicting CD4 Status of Patients on ART in Jimma and Bonga Hospitals, Ethiopia. *Journal of Health & Medical Informatics*, 06(06). Available at: https://www.omicsonline.org/open-access/application-of-data-mining-techniques-for-predicting-cd4-status-ofpatients-on-art-in-jimma-and-bonga-hospitals-ethiopia-2157-7420-1000208.php?aid=64757 [Accessed April 26, 2018].

Hailu, T.G., 2011. Comparing Data Mining Techniques in HIV Testing Prediction. *Intelligent Information Management Intelligent Informa-tion Management*, 7(7), pp.153–180. Available at: http://www.scirp.org/journal/iim [Accessed August 12, 2016].

Haire, B.G., 2015. Preexposure prophylaxis-related stigma: strategies to improve uptake and adherence - a narrative review. *HIV/AIDS (Auckland, N.Z.)*, 7, pp.241–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26508889 [Accessed April 29, 2018].

Iseu, G., Mwangi, W. & Kimwele, M., 2017. A Framework To Support Management Of HIV / AIDS Using K-Means And Random Forest Algorithm. , 6(06).

Karanasiou, G.S. et al., 2016. Predicting adherence of patients with HF through machine learning techniques. *Healthcare technology letters*, 3(3), pp.165–170. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27733922 [Accessed September 4, 2018].

Kornilova, M.S. et al., 2017. Decline in HIV seroprevalence in street youth 2006-2012, St. Petersburg, Russia: moving toward an AIDS-free generation. *International journal of STD & AIDS*, 28(4), pp.345–356. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27147268 [Accessed April 19, 2018].

Koss, C.A. et al., 2018. Early Adopters of Human Immunodeficiency Virus Preexposure Prophylaxis in a Population-based Combination Prevention Study in Rural Kenya and Uganda. *Clinical Infectious Diseases*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/29741594 [Accessed August 29, 2018].

Liu, C. et al., 2017. Factors influencing uptake of pre-exposure prophylaxis: some qualitative insights from an intervention study of men who have sex with men in China. *Sexual Health*, 15(1), pp.39–45. Available at: http://www.ncbi.nlm.nih.gov/pubmed/28859729 [Accessed September 23, 2018].

Liz Highleyman, 2016. HIV & AIDS Information :: HIV seroconversion is rare if Truvada

PrEP is used consistently. Available at: http://www.aidsmap.com/HIV-seroconversion-is-rare-if-iTruvadai-PrEP-is-used-consistently/page/3068919/ [Accessed April 22, 2018].

Mack, N. et al., 2014. Barriers and facilitators to pre-exposure prophylaxis (PrEP) eligibility screening and ongoing HIV testing among target populations in Bondo and Rarieda, Kenya: results of a consultation with community stakeholders. *BMC health services research*, 14, p.231. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24886646 [Accessed April 22, 2018].

NACC, 2016. Kenya AIDS Progress Report 2016. , p.92. Available at: www.nacc.or.ke.

National AIDS Control Council, 2014. Kenya AIDS Strategic Framework 2014/2015 - 2018/2019. *National AIDS Control Council*, pp.1–84. Available at: http://www.nacc.or.ke/attachments/article/181/KNASP III.pdf [Accessed April 19, 2018].

Nichols, B.E. et al., 2013. Cost-Effectiveness of Pre-Exposure Prophylaxis (PrEP) in Preventing HIV-1 Infections in Rural Zambia: A Modeling Study E. S. McBryde, ed. *PLoS ONE*, 8(3), p.e59549. Available at: http://dx.plos.org/10.1371/journal.pone.0059549 [Accessed April 22, 2018].

Ohl, M. et al., 2017. Predictors of First-Year Medication Adherence in a National Cohort of Veterans Initiating Pre-Exposure Prophylaxis (PrEP) to Prevent HIV Infection. *Open Forum Infectious Diseases*, 4(suppl_1), pp.S438–S438. Available at: http://academic.oup.com/ofid/article/4/suppl_1/S438/4294516/Predictors-of-FirstYear-Medication-Adherence-in-a [Accessed April 30, 2018].

Oliveira, A. et al., 2017. Data Mining in HIV-AIDS Surveillance System. *Journal of Medical Systems*, 41(4), p.51. Available at: http://www.ncbi.nlm.nih.gov/pubmed/28214992 [Accessed May 23, 2018].

Rachel Nall, 2016. How HIV Affects the Body. Available at: https://www.healthline.com/health/hiv-aids/how-hiv-affects-the-body [Accessed April 13, 2018].

Rodger, A.J. et al., 2016. Sexual Activity Without Condoms and Risk of HIV Transmission in Serodifferent Couples When the HIV-Positive Partner Is Using Suppressive Antiretroviral Therapy. *JAMA*, 316(2), p.171. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27404185 [Accessed April 22, 2018].

Seeley, J.A. & Allison, E.H., 2005. HIV/AIDS in fishing communities: Challenges to delivering antiretroviral therapy to vulnerable groups. *AIDS Care*, 17(6), pp.688–697. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16036255 [Accessed April 20, 2018].

Tarekegn, G.B. & Sreenivasarao, V., 2016. Application of Data Mining Techniques on Pre ART Data: The Case of Felege Hiwot Referral Hospital. *International Journal of Research Studies in Computer Science and Engineering*, 3(1), pp.2349–4840. Available at: www.arcjournals.org [Accessed May 4, 2018].

Tucker, C.S. et al., 2015. Machine learning classification of medication adherence in patients with movement disorders using non-wearable sensors. *Computers in biology and medicine*, 66, pp.120–34. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26406881 [Accessed

September 4, 2018].

UNAIDS, 2017a. Ending Aids Progress Towards the 90-90-90 Targets. *Global Aids Update*, p.198. Available at: http://www.unaids.org/sites/default/files/media_asset/Global_AIDS_update_2017_en.pdf.

UNAIDS, 2017b. Fact sheet - Latest global and regional statistics on the status of the AIDS epidemic. *2016*, (June), p.8. Available at: http://www.unaids.org/sites/default/files/media_asset/UNAIDS_FactSheet_en.pdf [Accessed April 17, 2018].

UNAIDS, 2016. HIV prevention among key populations | UNAIDS. Available at: http://www.unaids.org/en/resources/presscentre/featurestories/2016/november/20161121_keypops [Accessed April 19, 2018].

USAID, 2017. Key Populations: Targeted Approaches Toward an AIDS-Free Generation | U.S. Agency for International Development. Available at: https://www.usaid.gov/what-we-do/global-health/hiv-and-aids/technical-areas/key-populations [Accessed April 19, 2018].

Wimmer, H. & Powell, L.M., 2015. A Comparison of Open Source Tools for Data Science. *2015 Proceedings of the Conference on Information Systems Applied Research*, pp.1–9. Available at: http://iscap.info.

Witten, I.H., Frank, E. & Hall, M. a., 2011. *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition*, Available at: http://www.cs.waikato.ac.nz/~ml/weka/book.html%5Cnhttp://www.amazon.com/Data-Mining-Practical-Techniques-Management/dp/0123748569.

World Health Organization, 2015. Pre-exposure prophylaxis (PrEP). WHO expands recommendation on oral PrEP of HIV infection. *World Health Organization*, (November), p.2. Available at: apps.who.int/iris/bitstream/10665/197906/1/WHO_HIV_2015.48_eng.pdf?ua=1 [Accessed April 22, 2018].

Yongjian Fu, 2011. Data mining models and tasks. Available at: http://academic.csuohio.edu/fuy/Pub/pot97.pdf [Accessed April 23, 2018].