

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCES

DEPARTMENT OF STATISTICS

Modeling CD4 $^{+}$ Cell Counts of HIV-Positive Patients Following Antiretroviral Therapy (ART): A Case of Yekatit 12 Hospital, Addis Ababa.

By

Kebadu Tadesse

A thesis submitted to the School of Graduate Studies of Addis Ababa University in partial fulfillment of the requirements for the degree of Master of Science in Statistics (Biostatistics)

June, 2016

Addis Ababa, Ethiopia

Modeling CD4 $^{+}$ Cell Counts of HIV-Positive Patients Following Antiretroviral Therapy (ART): A Case of Yekatit 12 Hospital, Addis Ababa.

A thesis submitted to the school of Graduate Studies, Department of Statistics, College of Natural and Computational Sciences, Addis Ababa University in partial fulfillment of the requirements for the Master of Science Degree in Statistics (Biostatistics)

By:

Kebadu Tadesse

Advisor: Dr. Birhanu Teshome (Ph.D)

June, 2016

Addis Ababa, Ethiopia

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCES

DEPARTMENT OF STATISTICS

This is to certify that, the thesis prepared by Kebadu Tadesse, entitled: **Modeling CD4 $^{+}$ Cell Counts of HIV-Positive Patients Following Antiretroviral Therapy (ART): A Case of Yekatit 12 Hospital, Addis Ababa** and submitted in partial fulfillment of the requirements for the Degree Master of science in Statistics (Biostatistics) compiles with the regulations of the University and meets the accepted standards with respect to originality and quality.

Approved by the board of examiners:

Dr. Birhanu Teshome (Ph.D) Signature_____

Advisor

Dr. Shibru Temesgen (Ph.D) Signature_____

Examiner

Dr. Emmanuel G/Yohannes (Ph.D) Signature_____

Examiner

_____

Chair of Department or Graduate Program Coordinator

DECLARATION

I hereby declare that, the thesis is my genuine work and all sources of materials used for writing it have been duly acknowledged. I declare that I have not so far submitted this thesis to any other institution anywhere for that award of any academic degree, diploma or certificate.

Name: Kebadu Tadesse

Signature: _____

Place of submission: Department of Statistics, College of Natural and computational Sciences, Addis Ababa University

Date of submission: June, 2016

This thesis has been submitted for examination with my approval as a University advisor.

Dr. Birhanu Teshome (Ph.D)----------

Advisor's Name signature

ACKNOWLEDGEMNT

First of all I would like to thank Almighty God for being there with me in every step which I took. Next I am sincerely grateful to my advisor Dr. Birhanu Teshome for his help, guidance, consultation and encouragement in the completion of this work.

I sincerely thank my entire family for their love, encouragement and support throughout this phase of my academic journey.

I would like to thank Addis Ababa University and Wolkite University for providing me to attend the study and Yekatit 12 Referral Hospital for providing me with all the data in the Hospital.

Abstract

Modeling CD4 $^{+}$ Cell Counts of HIV-Positive Patients Following Antiretroviral Therapy (ART): A Case of Yekatit 12 Hospital, Addis Ababa.

Kebadu Tadesse

Addis Ababa University, 2016

HIV infection leads to severe depletion of $CD4^{+}$ counts with subsequent reduced levels of circulating $CD4^{+}$ lymphocytes in the peripheral blood. $CD4^{+}$ counts are the primary laboratory markers used to track the progression of HIV to AIDS. It is the strongest predictor for risk of death and AIDS (DHHS, 2011) at the time of initiating therapy and used to decide when to start ART treatment. The main objective of this study was to model the progression of $CD4^{+}$ cell counts of HIV- positive patients following ART. A longitudinal data was obtained from ART case unit of HIV positive patients in Yekatit 12 Hospital who were following ART from September 10, 2009 to September 9, 2012 G.C and followed until January 9, 2016 G.C, Addis Ababa. The data was extracted from SMART CARE of patients chart and patients with at least three times of measurements and those in first line ART regimens were included. The linear mixed model analysis was applied taking the correlation between $CD4^{+}$ counts of patients into account. We included 217 patients of whom 141 (65%) were females and only 51.61% of patients were followed until the last follow up time inducing 48.39% missingness. The results from the linear mixed model with AR(1) covariance structure showed that baseline CD4+ count, WHO clinical stage, time and the interaction effects of sex, Age, baseline $CD4^{+}$ count and regimen class with time were statistically significant factors for the progression of $CD4^{+}$ count at a square root level over time. Results from the full likelihood and multiple imputation approaches were found to be almost identical for covariates that were significant. The mean $CD4^{+}$ count showed an increasing progress over time after patients were initiated on ART. We recommend that patients should be initiated to ART regimens at an early WHO clinical stages, at a higher $CD4^{+}$ count and ART regimens should be taken continuously to increase the progress of $CD4^{+}$ count.

***KEY WORDS**: ART, $CD4^{+}$, Linear Mixed Model, Longitudinal Data, Missing Data, AR(1), ML, MI, MAR.*

LIST OF ACRONYMS

3TC	Lamivudine
AIC	Akaike Information Criteria
AIDS	Acquired Immunodeficiency Syndrome
ART	Anti-Retroviral Therapy
ARV	Anti-Retroviral
AR	Autoregressive
AZT	Azidothymidine/ Zidovudine
CD4	Cluster Of Differentiation
CDC	Center for Disease Control
USDHHS	U.S Department of Health and Human Services
d4t	Stavudine
DHS	Demographic and Health Survey
EFV	Efavirenz
EPHI	Ethiopian Public Health Institute
GLMM	Generalized Linear Mixed Model
HIV	Human Immune Deficiency Virus
MAR	Missing at random
MCAR	Missing completely at random
MDG	Millennium Development Goal
MI	Multiple Imputation
MNAR	Missing Not at random
MTCT	Mother to child transmission
NVP	Neverapine
OARAC	Office of AIDS Research Advisory Council
PMTCT	Prevention of Mother to child transmission
TDF	Tenofovir
UNAIDS	Joint United Nations Program on HIV/AIDS
WHO	World Health Organization

Table of Contents

Acknowledgment ……...…… i

Abstract ……..…… ii

List of Acronyms …….. iii

List of Tables ……...…… vii

List of Figures ……...……. viii

CHAPTER ONE

INTRODUCTION ……..…… 1

1.1.Background of the Study ……. 1

1.2. Overview of HIV epidemic in Ethiopia ……... 3

1.3.Statement of the Problem ……..5

1.4.Objectives of the Study ……6

1.4.1. Main objective ……..……. 6

1.4.2. Specific Objectives ……...……6

1.5.Significance of the study …….. 7

CHAPTER TWO

LETRATURE REVIEW ……..... 8

2. Introduction …… 8

2.1. Measurement of CD4 Count to Determine ART eligibility ……...…… 10

CHAPTER THREE

DATA AND METHODOLOGY ……. 13

3.1.Data ……... 13

3.2.Variables of the study …… 13

3.2.1. The Response Variable ……. 13

3.2.2. Covariates ……..13

3.3. Methods of Data Analysis …… 13

3.3.1. Longitudinal Data Analysis…….. 14

3.3.1.1.Longitudinal Data Structures …… 14

3.3.2. Exploratory Data Analysis 15
3.3.2.1.Variance Covariance Structures 16
3.4. Theory of linear mixed model 17
3.4.1. Linear mixed model for longitudinal data 18
3.4.1.1.Marginal Models 22
3.4.1.2. Assumptions of Linear mixed Models 23
3.5. Estimation of fixed effects 23
3.5.1. Maximum likelihood estimation (ML) 24
3.5.1.1.Restricted maximum likelihood estimation (REML) 24
3.6. Inference for the marginal model 25
3.6.1. Inference for the fixed effects 27
3.6.2. Inference for the variance components 30
3.7. Random Coefficient Model 30
3.8. Information criteria 30
3.9. Missing Data 31
3.9.1. Handling Missing Data 32
3.9.1.1.Maximum Likelihood (ML) 33
3.9.1.2.Multiple Imputations 33
3.10. Model comparison techniques 34
3.11. Model Checking Technique 34
CHAPTER FOUR
RESULTS AND DISCUSION 35
4.1.Descriptive Statistics 35
4.1.1. Descriptive Statistics of Continuous Variables 35
4.1.2. Descriptive Statistics of Categorical Covariates by Time..... 36
4.1.3. Distribution of $CD4^+$ Count Data 37
4.2.Exploratory Data analysis 39
4.2.1. Exploring Individual profiles 39
4.2.1.1.Scatter plots for $CD4^+$ count 41
4.2.2. Exploring the Mean Profile 43
4.2.3. Exploring Variance profiles 44

4.2.4. Exploring the Random Effect ……..…… 46
4.2.5. Exploring correlation structure and Scatter plot of Correlation matrix ……47
4.3. Linear Mixed Models ……...… 48
4.3.1. Univariable Marginal Models …….. 48
4.3.2. The full Marginal Model Including Variable selection …… 48
4.3.2.1.Selection of Covariance structure ……..… 48
4.3.2.2.Variable selection ……..… 48
4.3.2.3. Correlation Structure Checking …….. 49
4.3.3. The random effects model ……... 51
4.3.3.1.Univariable Random effects Model ……. 51
4.3.3.2.The random intercepts and slope model ……..... 51
4.3.4. Mixture Chi-square Distribution ……. 55
4.4. Handling Missing Data ……..56
4.3.4. Maximum Likelihood …….. 56
4.3.5. Multiple Imputations ……... 56
4.5. Model comparison ……..... 57
4.6. Model Diagnostics …… 58
4.7. Discussion of the results …… 61
CHAPTER FIVE
CONCLUSION AND RECOMMENDATION …… 63
5.1. Conclusion ……..……... 63
5.2. Recommendation ……...63
References …..…… 64
Appendix …….. 69

List of Tables

Table 4.1: Descriptive Statistics of Continuous variables ……..…… .35
Table 4.2: Descriptive Statistics of Continuous Covariates at baseline ……. .36
Table 4.3: Descriptive Statistics of categorical Covariates by Time ……...…….....37
Table 4.4: Shapiro-Wilk tests of Normality ……..…….. 38
Table 4.5: Correlation Matrix of $CD4^+$ count over Time …….……. 47
Table 4.6: Fit Statistics for checking correlation Structure …….…….……. 49
Table 4.7: Parameter Estimates for full linear Mixed Model ……...…….…… 50
Table 4.8: Type 3 Tests for full linear Mixed Model ……..…….……. 51
Table 4.9: Fit Statistics for random intercepts and slope model ……...…..…… 52
Table 4.10: Covariance Parameter Estimates …….……... 52
Table 4.11: Parameter Estimates for Random Slope and Intercept Model …….…….. 53
Table 4.12: Type 3 Tests of Random Slope and Intercept Model ……..……... 54
Table 4.13: Covariance Parameter Estimates for random Intercept only Model ……..…….. 54
Table 4.14: Fit Statistics for Random Intercept only Model ……..…….…….. 54
Table 4.15: Parameter Estimates of RandomIntercept Model ……. 54
Table 4.16: Mixture Chi-square Distribution …….…… 55
Table 4.16: Parameter Estimates for Multiple Imputation ……...…… 57

List of Figures

Figure 4.1: Mean $CD4^+$ Count Plot by Time ……36

Figure 4.2. Normal Q-Q plots of $CD4^+$ counts ……..……..…… 39

Figure 4.3 Box Plots of $CD4^+$ count by Time …….. 39

Figure 4.4. Individual Profile plot of $CD4^+$ count by Time ……..... 40

Figure 4.5: Individual Profile Plot of Square Root $CD4^+$ Count by Time ……...…… 41

Figure 4.6. Individual Profile plots of the response by Time and Sex ……...….. 41

Figure 4.7. Scatter plots of $CD4^+$ count against Covariates by time 43

Figure 4.8: Mean Profile of Patients $CD4^+$ count of Categorical Covariates ……..… 44

Figure 4.9. Variance Profile of $CD4^+$ count by Categorical Covariates over Time ……..….. 45

Figure 4.10 Confidence Interval Plot of random intercept and Linear Time Effect ..…….. 46

Figure 4.11: Scatter plot of correlation matrix ……...…….. 47

Figure 4.12 Residuals plot for square root $CD4^+$ count ……..……. 59

Figure 4.13 Studentized residual plots of square root $CD4^+$ count……...… 60

Figure 4.14 Pearson Residual Plots of Square Root $CD4^+$ Count ……...…….…….. 60

CHAPTER ONE

INTRODUCTION

1.2.Background of the Study

The Human Immunodeficiency Virus (HIV) causes Acquired Immune-Deficiency Syndrome (AIDS), a condition in which the immune system begins to decline, exposing infected individuals to life-threatening opportunistic infections (Duncan *et al*, 1999). It is an epidemic that affects every part of the globe.

Since the first case was identified in 1981, around 78 million (71 million-87 million) have been infected with HIV and 39 million (35 million-43 million) people have died of AIDS related illnesses (UNAIDS, 2014).

According to the latest figures, an estimated 36.9 million people globally are living with HIV of which 25.8 million are in Sub-Saharan Africa (UNAIDS, 2015). There were two million new infections in 2014 with 1.4 million (65%) of these occurring in Sub-Saharan Africa. In 2014, 1.2 million people died of AIDS-related illnesses globally of which 790,000were in Sub-Saharan Africa. As of June 2015, 15.8 million people living with HIV were accessing antiretroviral therapy, up from 13.6 million in June 2014. In sub-Saharan Africa, 10.7 million people were accessing antiretroviral treatment, 41% of all people living with HIV in the region, up from fewer than 100,000 people in 2002 (UNAIDS, 2015).

HIV is associated with immune suppression which can be measured by the levels of T cell (Roit*et al.,* 2006). Specifically, HIV has been shown to reduce levels of $CD4^+$ and $CD3^+$ counts in addition to other hematological values (Borough *et al.,* 2005).

HIV infection leads to severe depletion of $CD4^+$counts in the gut-associated lymphoid tissue with subsequent reduced levels of circulating $CD4^+$ lymphocytes in the peripheral blood. $CD4^+$counts are reduced precipitously in acute HIV infection, but usually rebound over several weeks as HIV-specific CD8 T cells help to lower plasma viremia. In the untreated patient, $CD4^+$counts subsequently decline over several years. Population-based studies of the natural history of HIV infection among men who have sex with men show that the mean $CD4^+$ count prior to seroconversion is about 1000 cells/mm3; $CD4^+$counts decline to a mean of 780cells/mm3 at six months post-seroconversion and to 670cells/mm3 at one year of follow-up. Subsequently, the

$CD4^+$ cell count declines at an average yearly rate of approximately 50cells/mm3, but there is substantial variation among patients'. Significant depletion of $CD4^+$counts can lead to opportunistic infections and mortality in the untreated patient.

$CD4^+$ T lymphocyte ($CD4^+$) counts are the primary laboratory markers used to track the progression of HIV to AIDS. It is the strongest predictor for risk of death and AIDS (DHHS, 2011) at the time of initiating therapy. Initiating HAART at higher $CD4^+$ counts has been demonstrated to "the risk of death, opportunistic infections and non-HIV related co morbidities" (Lichtenstein *et al*, 2008.). $CD4^+$ count is also used to decide when to start antiretroviral therapy (ART).The use of the $CD4^+$ count as an independent and reliable marker for treatment outcome is attractive from various aspects. First, $CD4^+$ counts are already the most important factor in deciding whether to initiate antiretroviral therapy and opportunistic prophylaxis in all HIV-positive patients in high-income countries, and an increasing number of patients in low-income countries have a baseline $CD4^+$ count at entry into care. Secondly, the $CD4^+$ count is a relatively objective and simple marker to follow. Finally, the cost of $CD4^+$ counts has become more affordable, including in developing countries.

Normal levels of $CD4^+$ counts in healthy individuals ranges from 1000 to 1200 cells/µl of blood (Borough *et al.,* 2005). HIV patients with or less than 200 CD4 cells/µl of blood have been shown to develop AIDS which is associated with opportunistic infections and malignancies (Gilks, 1998).

The 2010 WHO revised guideline defined immunological failure as a fall of $CD4^+$ count to baseline level or below, or 50% fall from on-treatment peak value or persistent $CD4^+$ count below 100 cells/mm^3. There must, however, be absence of concomitant infection to cause transient $CD4^+$count decrease. A patient presenting with immunological or clinical failure (new or recurrent stage four disease) with viral load copies >5000 copies/mL is deemed to have treatment failure and switched to second-line regimens (WHO, 2010). However, clinicians still debate on the appropriate CD4 threshold at which to initiate HIV therapy.

ART involves taking a combination of HIV medicines (called an HIV regimen) every day. It prevents HIV from multiplying and destroying our infection-fighting from $CD4^+$counts. ART cannot cure HIV, but it can help the patient live a longer, healthier life and reduce their risk of HIV transmission.

HAART is the mainstay of treatment for those infected with HIV **(**Jayasuriya, *et al*., 2007). Since its introduction in 1996, mortality and morbidity rates in HIV-infected individuals in countries with widespread access to HAART have plummeted **(**Mocroft*et al*., 2003). The main effect of HAART is to suppress viral replication, allowing the individual's immune system to recover and protecting him/her from the progress of AIDS and death.

The optimal time to start treatment for HIV/AIDS has been a contentious issue since the introduction of HAART. ART is recommended for everyone with HIV, but the urgency to start ART is greater in people with low or rapidly falling $CD4^+$ counts. A falling $CD4^+$ count indicates that HIV is advancing and damaging the patient's immune system. Most treatment guidelines recommend that all individuals with a $CD4^+$ count less than 350 cells/µl should be treated (Hammer *et al*., 2008).

The 2013 ARV guidelines recommend initiating ART earlier at $CD4^+$ count ≤500 cells/mm3 – and immediately initiating ART for serodiscordant couples, pregnant women living with HIV, people with TB and HIV, people with HIV and Hepatitis B, and children living with HIV who are younger than five years, irrespective of $CD4^+$ count. For patients with $CD4^+$counts in the range 350-500 cells/µl, opinion remains divided as to whether treatment should or should not be. Those with severe and advanced clinical disease (WHO clinical stage three and four) should start ART irrespective of their $CD4^+$ count. Most treatment guidelines recommend that among patients with $CD4^+$ counts in this range, treatment be considered in those with high viral loads, those experiencing a rapid drop in $CD4^+$ count, or those with low CD4 percentages (Gazzard, 2008).

1.2. Overview of HIV epidemic in Ethiopia

In 2016, there will be an estimated 740,251 people living with HIV including 143,201 children according to the latest EPHI estimate. Of the estimates living with HIV in Ethiopia, 495,194 of them are living in urban areas and the rest 245,075 are in rural areas. As per the same estimate, the estimated HIV population in Ethiopia is mostly adults who were vertically infected in earlier years which cover 597,050 of the total and the rest 143,201 estimates living with HIV are children's aged under 14 years. There will be an estimated 11,479 new infections in 2016 with 9,374 (82%) of these are adults living in the country. In 2016, an estimated 20,354 people annually will die of AIDS-related illnesses in the country of which **18,174** are above 15 years old and 13,616 are in urban areas of the country. According to the 2014 estimate by EPHI of HIV related estimates and

projections for Ethiopia, in 2016 697,449 people living with HIV will needs accessing antiretroviral therapy, up from 580,067 in 2013 (EPHI, 2014).

There were approximately 45,200 (36,500-55,200) AIDS related deaths in 2013 and about 898,400 (770,700–1,048,500) AIDS orphans in the same year. HIV adult prevalence was estimated at 1.5% in 2011, the year in which the last Ethiopian Demographic Health Survey (DHS) was conducted. However prevalence varies according to age, gender and geographical location. According to the 2011 DHS, adult prevalence was almost twice as high among females compared to males at 1.9% versus 1.0% respectively. The distribution of HIV prevalence also varies by age, peaking earlier in females in the 30-34 years age group compared to 35-39 years in males. Looking at the younger age groups it can be seen that young women have a two to six fold higher HIV prevalence than young men (ranging from 15-17 years: no males and there was 0.2% females to 20-22 years: 0.1% males vs. 0.6% females). Marked variation in urban rural prevalence was also reported in the 2011 EDHS with urban areas showing a seven fold higher HIV prevalence compared to rural areas (4.2% versus 0.6%) (FDRE, 2014).

The federal HIV/AIDS prevention and control office of Ethiopia launched a five year strategic plan with objectives of implementing high impact and target prevention program, intensifying targeted HIV testing and counseling service, attaining virtual elimination of mother to child transmission of HIV and optimizing and sustaining quality care treatment by the year 2015-2020 (FDRE, 2014).

1.6.Statement of the Problem

HIV/AIDS epidemic remains to be one of the major challenges of development in both developed and developing countries of the world. It affects peoples' welfare, undermining social cohesion and security, incapacitating global workforces, and diminishing social and economic productivity (World Bank, 1999).Ethiopia is among the most affected countries with HIV /AIDS in the world and the disease is one of the top ten causes of death in the country. The epidemic affects the population as the number of persons living with the disease, the number of new infections each year, the number of deaths related to HIV and the number of persons in need of ART is high in Ethiopia especially in urban areas of the country.

It is recommended that HIV infected patients have to start ART in order to reduce AIDS related morbidity and mortality, or to improve their quality of life. But in most cases, factors influencing the progression of $CD4^+$ counts/ HIV were not well identified. Classical statistical models are not appropriate for longitudinal $CD4^+$ count data as the $CD4^+$ counts are not expected to be independent within a patient over time. Linear mixed models are appropriate for modelling the progression of $CD4^+$ count and to identify associated risk factors over time taking the correlation of $CD4^+$ counts within a patient into account.

As is the case in most longitudinal studies, missing data were also observed in our CD4+ count studies. Hence, a study that deals with progression of $CD4^+$ count and risk factors influencing this progression should apply statistical methods that could handle the missing values in the data. Direct likelihood and multiple imputations are appropriate methods of handling missing data in case of $CD4^+$ count data by linear mixed models.

Hence, our study problems include: assessing and modelling the progression of $CD4^+$ counts over time; identifying factors influencing the progression of $CD4^+$ count; comparing the progression of $CD4^+$ count between groups and handling missing observations by direct likelihood and multiple imputation.

1.7.Objectives of the Study

1.7.1. **Main objective**

The main objective of the study was to model the progression of $CD4^+$ counts of HIV-positive patients following ART for a period of four years in Yekatit 12 Referral Hospital.

1.7.2. Specific Objectives

- ✓ To assess progression of $CD4^+$ cell counts using Linear Mixed models (LMM)
- ✓ To compare the progression of $CD4^+$ T -cells among ART regimens, gender, marital status, functional status and WHO stage
- ✓ To identify determinant factors for the change in $CD4^+$ cell count in HIV positive patients in ART case unit of Yekatit 12 Hospital.

1.8.Significance of the study

Longitudinal modeling of $CD4^+$ counts of HIV patients initiated on ART is a longitudinal study as the number of $CD4^+$ counts is recorded over time and they are expected to be correlated within a patient. Models that assume independence among observations are not efficient for longitudinal data because measurements taken on the same subject tend to be more similar than measurements taken on different subjects and measurements taken close in time on the same subject tend to be more similar than measurements taken far apart in time and the variances of longitudinal data often change with time. Hence, special statistical methods like linear mixed models are needed for longitudinal data analysis. These potential patterns of correlation and variation may combine to produce a complicated covariance structure that must be taken into account to draw valid statistical inferences. Hence, mixed effect models allow studying the progression of $CD4^+$ counts over time and can model it with the predictors and correlation by incorporating random effects.

Due to the presence of missing observations in our study on progression of $CD4^+$ counts over time, full likelihood and Multiple Imputation approaches were used as missing data handling techniques.

The results of this study might be useful in the development of an effective HIV care and ART patient monitoring system. Even though perfect eradication of HIV from an individual is not currently possible, we can delay the progression of disease by using ART which provide protection against the development of HIV-related complications by a long-term increase in $CD4^+$ T-cells. This study will assess the importance of attending ART program by showing how much $CD4^+$ T-cells change over time so that patients' functional status and disease protective ability also increase.

CHAPTER TWO
LETRATURE REVIEW

2. Introduction

There is a need for close supervision and improvement on the existing techniques that would ensure monitoring $CD4^+$ trends and variations in trends on HIV/AIDS patients on HAART (Stohr*et al.,* 2007). A prospective observational study of more than 8500 HIV/AIDS patients (called Eurosida) was conducted in 63 hospitals of 20 European countries using absolute $CD4^+$ counts coupled with viral load monitoring (Florence *et al*., 2003). Patients for this study were all who started HAART with moderate immunologic suppression and with base line $CD4^+$ count below 350 cells/ µl measured within previous six months at most. Association between risk factors potentially associated with a low $CD4^+$ count response in the Eurosida study was assessed by using a logistic regression model. The risk factors taken into account for low $CD4^+$ count response were demographic factors (age, sex ratio, ethnic origin, HIV transmission group, region), antiretroviral treatment factors (number of nucleotide reverse transcriptase inhibitors (NRTI) use before HAART, start date of HAART, type of HAART, HAART duration), immunological factors (previous episodes of opportunistic infections).

According to Allison (2008), as untreated HIV progresses, $CD4^+$ count decreases by about approximately 4% every year. With successful ART follow up, $CD4^+$ count might increase by greater than 50 cells per micro liter within weeks after viral suppression and it tends to vary. Additionally, it may increase by 50-100 cells per micro liter per year thereafter, until a threshold is reached (Allison, 2006). However, it is important to note that in some patients, $CD4^+$ count may not increase this quickly or steadily even with durable viral suppression.

A study by Akinsegun*et al*, 2012 on $CD4^+$ count pattern and demographic distribution of treatment-naive HIV patients in Nigeria concluded that females account for more than half of registered patients in HIV clinic and have a relatively higher $CD4^+$ count than males. About three-quarter of HIV positives require antiretroviral therapy at registration.

A study using generalized linear mixed models and generalized estimating equations by Obeagu *et al* (2014) on CD4+ count of ART and non-ART HIV positive patients attending Living Word Mission Hospital in Abayi Aba, Abia State, Nigeria, showed a significant increase in the CD4+

count of the ART HIV patients compared to the non-ART HIV patients. This shows that it is important to know your HIV status on time and to monitor the $CD4^+$level to start ART early enough to ensure improved well-being of the infected patients. There was a significant improvements in females placed on ART more than their male counterparts. According to that study, more enlightment campaign should be carried out to the populace for adequate health education to start the treatment early to prolong the life span and happiness of those infected with HIV.

A study on $CD4^+$ count trends after commencement of antiretroviral therapy among HIV infected patients in Tigray, Northern Ethiopia was conducted by Addisu*et al*., (2015). Logistic regression was used to examine factors associated with the risks of immunological non-response and a multiple least-squares linear regression analysis was employed to assess changes in CD4 cell count. The results showed that higher baseline $CD4^+$ counts and being male were significant predictors of subsequent $CD4^+$ recovery.

A study by Marlon (2010) on factors affecting $CD4^+$ count response to HAART in HIV/AIDS patients within 24 months of treatment at Chreso ministries art centers' in Zambia was conducted. The Chi-Square test was used to determine the association between the identified variables and CD4-Lymphocyte count response to HAART and logistic regression analysis was used to predict the probability of CD4 count response to HAART using the variables of the study. The study finds that gender, alcohol consumption, base line $CD4^+$ count and regimen type affect $CD4^+$ count response to HAART. It was found that men, non-alcohol consumers and those that start HAART with baseline $CD4^+$ count above 350 cells/ mm^3 experienced a good $CD4^+$ response to HAART. On the other hand, age, smoking and employment status were not found to affect $CD4^+$ count response to HAART.

In a study by Wellington (2013) on comparison of $CD4^+$ T-Cell changes in response to HAART in adolescents and children enrolled at Parirenyatwa Hospital family care centre (2005-2010) in Zimbabwe, differences between groups in CD4 cell response at different time points was assessed using Wilcoxon rank-sum test. Mixed effects model was also used to compare the pattern of changes in CD4 count over time between adolescents and children and to identify the factors which were associated with changes in CD4 count after ART initiation. Baseline $CD4^+$ counts and age group were found to predict the changes in the square root of $CD4^+$ count over time in the multivariate analysis. The increase in the square root of $CD4^+$ count over time for those who initiate HAART at adolescence stage were 0.0853 times less when compared to those who initiated

whilst they were still children, adjusting for other variables. Adjusting for other baseline variables, patients with $CD4^+$count less than 100 cells/ mm3 experienced a significant increase in the square root of $CD4^+$ count when compared to those with baseline $CD4^+$ count of more than 300 cells/ mm3.

A study on time to and predictors of $CD4^+$ T-Lymphocytes recovery in HIV-Infected Children initiating HAART in Ghana was conducted using Cox proportional hazards (CPHs) models (Lorna *et al*, 2011). The study found that severe suppression of baseline $CD4^+$ T-lymphocyte count adjusted for age, age at HAART initiation, gender and having parents alive was statistically significant in predicting time to $CD4^+$ T-lymphocyte recovery.

2.1. Measurement of CD4 Count to Determine ART eligibility

$CD4^+$ count is a marker of immune function in HIV infection and is recommended by WHO as the most appropriate means of determining immunological stage of HIV disease and in ascertaining when ART should be initiated in settings with appropriate clinical laboratory infrastructure. Measurement of $CD4^+$ count can be particularly challenging for HIV–infected patients in Sub Saharan Africa and often requires multiple visits to tertiary hospitals to complete measurement with substantial out of pocket expenditure. A cohort study from South Africa showed that only 31% of newly diagnosed HIV positive individuals had completed $CD4^+$count measurement by 12-months and in Mozambique 23% had not completed $CD4^+$ count by six months (WHO, 2012).

Smith *et al., (*2004) conducted a study on factors influencing increases in $CD4^+$ counts of HIV-positive persons receiving long-term HAART. By linear regression and mixed models, it was found that white ethnicity, higher pre-HAART virus load, and lower pre-HAART CD4 and CD8 cell counts were associated with greater increases in $CD4^+$ counts during the first three months of HAART. From three months on ward, a greater cumulative proportion of time spent with virus load <400 copies/ mL was associated with a more favorable change in CD4+ count. According to the same study, sex, risk group, age, and HAART regimen were not associated with increases in $CD4^+$ counts.

Yarchoan*et al., (1991)* assessed $CD4^+$ count and the risk of death in HIV-infected patients on HAART. Nearly all deaths occurred in patients with fewer than 50 CD4 cells/mm3.A common denominator amongst all the guidelines on initiating HAART in HIV-positive patients is the use of $CD4^+$ count in deciding when to initiate ART in HIV positive patients. When some HIV+

patients present at registration with low $CD4^+$ count, i.e., less than 200cells/mm3, probably due to late presentation or diagnosis, they have to be commenced on HAART almost immediately irrespective of clinical symptoms. Some asymptomatic HIV-positive patients do not require antiretroviral (ART) drugs on enrollment because of their high $CD4^+$ count at registration hence their inability to meet criteria for initiation of therapy laid down by various organizations like World Health Organization (WHO), Centre for Disease Control (CDC) Atlanta, Presidential Emergency Program for AIDS relief (PEPFER), Office of AIDS Research Advisory Council (OARAC), and European AIDS Clinical Society Guidelines, among others.

A study on longitudinal analysis of change in $CD4^+$ counts of HIV-1 patients on ART in the Builsa District Hospital in Ghana using linear Mixed effects modelling approach showed that the initial $CD4^+$ cell count of a patient, the duration of treatment and the drug type used in the treatment were factors that significantly determine a patient's current $CD4^+$ cell count (Adams *et al. (2013))*.

A study by Steve., *et al* (2014) on $CD4^+$ count recovery after initiation of ART in a resource limited setting of a retrospective cohort analysis found that age and base line $CD4^+$ count played important role in $CD4^+$ count recovery. A study on factors affecting the survival of HIV-infected children after ART initiation in Bahir-Dar, Ethiopia, was conducted using Cox proportional hazards regression model (Habtamu*et al.,* 2012). The results showed that baseline hemoglobin level, WHO clinical stage and age at baseline had significant impacts on the survival of children during 30 months of follow up.

A retrospective longitudinal study on predictors of change in CD4 Lymphocyte count and weight among HIV infected patients on ART treatment in Ethiopia was conducted using multiple regression analyses (Ayalu, *et al.* (2013)). According to this study, the duration of ART, advanced WHO clinical stage, lower baseline CD4 cell count and baseline hemoglobin levels were factors associated with decline in weight. Actively working patients had higher CD4 lymphocyte count and weight compared to those that were ambulatory.

A study on effectiveness of first-line antiretroviral therapy and correlates of longitudinal changes in CD4 and viral load among HIV-infected children in Ghana employed a mixed linear modeling approach. The result of the study found that age at initiation of therapy, child's gender, having a parent as a primary care giver, severity of illness, and type of regimen were associated with

treatment outcomes and concludes that first-line ART regimens were effective and well tolerated (Oliver *et al.* (2013)).

A study on predicting AIDS disease progression using longitudinal $CD4^+$ count among adult HIV/AIDS patients was conducted in South West Ethiopia using a semi-Marcov process (Dinberu*et al* (2016)). The result of the study revealed that having lower $CD4^+$ count at the start of the ART treatment resulted in higher probability of death.

A statistical analysis of CD4+ cell counts progression of HIV-positive patients enrolled in ART atHossana District Queen ElleniMohamad Memorial Hospital, South Ethiopiaused generalized linear mixed, Poisson, Poisson-gamma, Poisson-normal, and Poisson-normal-gamma models. The covariates significant for the progression of CD4+ cell counts were age of the patient, time since seroconversion and sex at 5% level of significance (Getachew (2016)).

CHAPTER THREE

DATA AND METHODOLOGY

3.3.Data

The data was obtained from the ART case unit of Yekatit 12 Hospital HIV positive patients who initiated on ART from September 10, 2009 to September 9, 2012 and followed until January 9, 2016. Longitudinal data are extracted from the patients' chart which contains epidemiological, laboratory and clinical information of all HIV patients under ART follow-up including a detailed antiretroviral therapy history. In this study, patients with at least three times $CD4^+$ count measurements and those who started first line regimens were included.

3.4.Variables of the study

The response and predictor variables considered in this study are defined as follows.

3.4.1. The Response Variable

The response variable was the number of $CD4^+$ counts of HIV Positive patients during the follow up time from the date of ART initiation.

3.3.2. Covariates

Explanatory variables that were assumed to influence the outcome of $CD4^+$ count after initiation of ART are sex, age in years (at ART initiation), marital status (married and single), baseline $CD4^+$ count, weight in kgs, regimen type (ART first line regimens), WHO staging at ART initiation (Stage I, Stage II, Stage III and Stage IV), functional status (working, ambulatory and bed redden) and time in months. Detail of the explanatory variables and their codings are presented in Table A_1 of the appendix.

3.4. Methods of Data Analysis

Repeated responses may arise in two common situations; one is when the measurements are taken on the same subject at different times and the other is when the measurements are taken on related subjects (clusters). In both of these cases, the responses are likely to be correlated. The term repeated measures is used to describe both the longitudinal and clustered data.

Longitudinal data is defined as the data in which repeated measurements are taken on a subject through time. Models can distinguish changes over time within subjects from differences between subjects at their baseline levels. The objectives of longitudinal data analysis are to examine and compare responses over time. The defining feature of a longitudinal data model is its ability to

study changes over time within subjects and changes over time between groups. Longitudinal studies are in contrast to cross-sectional studies in which a single outcome is measured for each individual. For example, longitudinal models can estimate individual-level (subject-specific) regression parameters and population-level regression parameters.

Longitudinal data sets differ from time series data sets because longitudinal data usually consists of a large number of a short series of time points. On the other hand, time series data sets usually consist of a single, long series of time points (Diggle, Liang, and Zeger 1994).

Traditional statistical methods such as univariate analysis of variance (ANOVA) cannot be applied to longitudinal data analysis due to their restrictive model assumptions that the residual or error terms are independently and identically distributed. Linear mixed effects models (LMM) have less restrictive assumptions. The variance covariance structure of the random variables does not have to be independent. Data can be unbalanced or each subject does not need to have the same number of observations per subject and repeated measurements do not have to have equal time intervals.

3.11.1. Longitudinal Data Analysis

The main objective of statistical analysis is to address variations in the data. For longitudinal data, there are two sources of variations: within-subject variation' the variation in the measurements within each subject, and between-subject variation' the variation in the data between different subjects. Modeling within-subject variation allows studying changes over time, while modeling between-subject variation allows understanding differences between subjects (Bijleveld, van derKamp *et al*, 1999*).*

3.3.1.1. Longitudinal Data Structures

Before formulating the models, let's illustrate the longitudinal data structures briefly. First we consider the structures for univariate longitudinal data, and then we continue with their multivariate counterparts.

Suppose that there are N subjects for whom data is collected repeatedly over time. Then, $Y_i$corresponds to a $n_{i\mathrm{x}}1$response vector of the i^{th}subject where $Y_i$can be illustrated by

$$Y_i = \begin{pmatrix} Yi,1 \\ Yi,2 \\ \vdots \\ Yi,ni \end{pmatrix}, i = 1, 2, \dots, N$$

The numbers of repeated measures for subject i is defined as $\mathbf{n_i}$, since it is assumed that they can vary from subject to subject. Also, it is assumed that the time at which measurements are taken might be uncommon.

Additionally, X_i is an associated $\mathbf{n_i x p}$ covariate matrix where p is the number of covariates and X_i can be illustrated by

$$X_i = \begin{pmatrix} \mathrm{X_{i11}} & \mathrm{X_{i12}} & & \mathrm{X_{i1p}} \\ \mathrm{X_{i21}} & \mathrm{X_{i22}} & \dots & \mathrm{X_{i2p}} \\ \vdots & \vdots & & \vdots \\ \mathrm{X_{ini1}} & \mathrm{X_{ini2}} & & \mathrm{X_{inip}} \end{pmatrix}, i = 1, 2, \dots, N$$

Where X_{itp} corresponds to the pth covariate for subject i at time t. Note that the elements of X_i might be changing with time or not changing with time.

Multivariate counterparts are extension of these data structures. We continue assuming the existence of N subjects. This time Y_i corresponds to an $\mathbf{n_i x k}$ response matrix of the ith subject where $\mathrm{Y_i}$ can be illustrated by

$$Y_i = \begin{pmatrix} \mathrm{Y_{i,1,1}} & & \mathrm{Y_{i,1,k}} \\ \mathrm{Y_{i,2,1}} & \dots & \mathrm{Y_{i,2,k}} \\ \vdots & & \vdots \\ \mathrm{Y_{i,ni,k}} & & \mathrm{Y_{i,ni,k}} \end{pmatrix}, i = 1, \dots, N$$

The associated covariate matrix for the multiple responses is similar to the one for the univariate case.

3.11.2. Exploratory Data Analysis

The first step in any model building process is exploratory data analysis. Data exploration is a very helpful tool in the selection of appropriate models and understanding of the data nature. Analyses of longitudinal data compare profiles over time. Examining the data for clues about the likely nature of the mean structure and to see how the mean profile changes over time is essential for specifying the functional form of the mean response of the model. So as to understand the possible relationships among means over time for balanced data, graphical inspection can be used by

connecting the average values computed at each time point separately. If the data is not balanced smoothing could be used instead. This will give idea as to how the mean profile evolves over time. The results of mean profile exploration will be useful to choose the appropriate fixed-effects structure of the linear mixed model.

Graphical methods can also be used to explore the magnitude of person-to person variability in outcomes over time. One approach is to create a panel of individual line plots for each study participant. These plots can be inspected for both the amount of variation from subject-to-subject in the overall “level" of the response, and the magnitude of variation in the “trend" over time in the response.

The correlation structures can be studied through the correlation matrix, or a scatter plot matrix. Graphically, pair-wise scatter plots are used for exploring the correlation between any two repeated measurements.

3.11.2.1. Variance Covariance Structures

Measurements made on the same subject are likely to be more similar than measurements made on different individuals. For an analysis to be valid, the co-variances among repeated measures must be modeled properly.

The four most commonly used covariance structures are compound symmetry (CS), Toeplitz (TOEP), unstructured (UN), and autoregressive (1) (AR (1)).

1. **Compound symmetry (CS)**: The correlations between all pairs of measures are the same. That is, the observations of the same subject have homogeneous variance and homogeneous covariance. With four repeated measures, the CS will have the following shape:

$$CS = \sigma^2 \begin{bmatrix} 1 & \rho 2 & \rho 2 & \rho 2 \\ \rho 2 & 1 & \rho 2 & \rho 2 \\ \rho 2 & \rho 2 & 1 & \rho 2 \\ \rho 2 & \rho 2 & \rho 2 & 1 \end{bmatrix}$$

Constant variance (σ^2) overtime and equal correlation $\sigma_1^2/(\sigma^2 + \sigma_1^2)$ between any two measurements from the same subject. This variance covariance matrix might not be appropriate in situations where measurements close together have higher correlation than measurements that are far apart in time.

2. **Autoregressive (AR (1))**: This structure resolves some of the objections to the use of compound symmetry. The correlation that AR (1) uses between two responses that are *m* measurements apart

is ρ^m. Since ρ is less than 1 in absolute value, the greater the power, the smaller the magnitude. Thus, the farther apart measurements have lower correlation.

$$\text{AR (1)} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho 2 & \rho 3 \\ \rho & 1 & \rho & \rho 2 \\ \rho 2 & \rho & 1 & \rho \\ \rho 3 & \rho 2 & \rho 1 & 1 \end{bmatrix}$$

3. **Toeplitz(TOEP)**: Similar to AR (1) in that all correlations at the same distance have the same correlation, but no assumption of exponential decay. The AR (1) model can be estimated with a single parameter; the Toeplitz model has as many parameters as there are distances. If there were 3 measures, then there would be two distances:-one unit distant, two units distant and hence two parameters would be estimated.

$$\text{TEOP} = \sigma^2 \begin{bmatrix} 1 & \rho_1^2 & \rho_2^2 & \rho_3^2 \\ \rho_1^2 & 1 & \rho_1^2 & \rho_2^2 \\ \rho_2^2 & \rho_1^2 & 1 & \rho_1^2 \\ \rho_3^2 & \rho_2^2 & \rho_1^2 & 1 \end{bmatrix}$$

4. **Unstructured**: Here the assumption is that all the variance and covariance are different. This let the data dictate what they should be and requires the estimation of many parameters. But, the more data that are used to assess the covariance structure, the more it is used to estimate the parameters of linear models.

$$\text{UN} = \sigma^2 \begin{bmatrix} \rho_{11} & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{21} & \rho_{22} & \rho_{23} & \rho_{24} \\ \rho_{31} & \rho_{32} & \rho_{33} & \rho_{34} \\ \rho_{41} & \rho_{42} & \rho_{34} & \rho_{44} \end{bmatrix} \text{ where, } \rho_{ij} = \rho_{ji}$$

3.12. Theory of linear mixed model

Mixed models provide a flexible and powerful tool for the analysis of data with complex covariance structure, such as longitudinal correlated data. Linear mixed effects models for repeated measures data formalize the idea that an individual's pattern of responses is likely to depend on many characteristics of that individual, including some that are unobserved (Der and Everitt, 2006).These unobserved effects are then included in the model as random variables, or equivalently called, random effects (Der and Everitt, 2006). A random effects model means that the levels of the factor variable in the data being modeled comprise a random sample of levels in the target population.

A fixed effect is considered to be a constant which we wish to estimate, but the random effect is considered as just an effect coming from a population of effects (McCulloch *et al*., 2008). To

emphasize on this distinction, we use the term prediction of random effects rather than estimation (McCulloch *et al*., 2008). Deciding whether a factor is random or fixed is not always easy and can be controversial. The collection of such class of linear models is referred to as linear mixed models (Littell*et al*., 2006). A repeated measures model is a special case of the general linear mixed model. The distinguishing feature of the repeated measures model lies in the specification of the covariance structure of the repeated measures. The choice of fixed and random effects is not always determined by the structure of the experiment, but may depend on the information required. The model for repeated measurements from the same individual implies an underlying correlation structure between measurements on the same subject which constitute a cluster of observations. Before stating the linear mixed model, the general multiple regression model is briefly discussed. The model is given by Y_i

$$Y_i = X_i\beta + \varepsilon_i \quad \text{(3.1)}$$

where$Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in})$' is a vector of n repeated measurements from the ith subject, X_i is the matrix of covariates, β is a vector of parameters and $\varepsilon_i \sim N(0; \sum_i)$. The distribution of Y_i is $Y_i \sim N(X_i\beta, \sum_i)$. Assuming independence across individuals, β and the parameters in $\sum_i$ can be estimated by optimizing the likelihood given by:

$$L_{ML} = \prod_{i=1}^{N}\{(2\pi)^{-n/2}|\sum_i|^{-1/2}\exp\left(-\frac{1}{2}(Yi - Xi\beta)'\sum_i^{-1}(Yi - Xi\beta)\right)\} \quad \text{(3.2)}$$

The inference is based on classical maximum likelihood theory namely likelihood ratio (LR) tests and asymptotic WALD tests.

3.12.1. Linear mixed model for longitudinal data

A commonly used mixed effects linear model for continuous response variables was proposed by Laird and Ware (1982). The name mixed model indicates that the model contains both the fixed or the mean model component and the random component and variable effects are either fixed or random depending on how the levels of the variables that appear in the study are selected (Laird and Ware, 1982).

Through random effects models, we can make inferences over a wider population in LMM than possible with GLM. The first step in the model building process for a linear mixed-effects model, after the functional form of the model has been decided, is choosing which parameters in the

model, if any, should have a random-effect component included to account for between-group variation.

Unlike general linear model family, linear mixed models (LMM) are models that handle data where observations are not independent. Many longitudinal studies are designed to investigate change over time in a characteristic which is measured repeatedly for each patient (Laird and Ware, 1982). Analyses of multiple observations measured on the same individual over time are different from observations measured on different people. The advantage of a longitudinal study is its effectiveness for studying change. There are natural changes like peoples' change in weight and there are also changes caused by interventions or treatment which we are often interested in. Often, we cannot fully control for the circumstances under which measurements are taken, and there may be considerable variation among individuals in the number and timing of observations (Laird and Ware, 1982). Investigators gather repeated measures or longitudinal data in order to study change in a response variable over time as well as to relate these changes in explanatory variables over time (McCulloch *et al*., 2008). To estimate treatment effect, it is important to adequately model the covariance structure of the repeated measurements. Therefore, it is important to try and model the correct correlation structure and this will yield more precise estimators of interest.

In addition, modeling the true correlation structure becomes important in the presence of missing values when the number of observations per subject is not large. There are two types of covariates in longitudinal studies in general. These are time invariant or baseline covariates (e.g. gender) and time varying covariates (e.g. weight). The Linear Mixed Model (LMM) has become the most commonly used tool for analyzing continuous repeated measures data from a sample of individuals and the term `individual' will have different interpretation or meaning for different areas of application. Next we will explain how measurements may be taken repeatedly on the same unit.

The units are patients in a longitudinal study where measurements of biological laboratory markers such as $CD4^+$ count are taken at every six monthly visits. Thus the patient is measured repeatedly giving rise to a cluster of observations from each patient. Also the within subject $CD4^+$ count are subject to biologic variation and measurement error, both of which can be considerable (Malone *et al*., 1990).

Thus, it is imperative that between subject variability and within subject correlation are adequately accounted for in order to get reliable inference about parameters of interest. In this study the unmeasured characteristics can be that different patients have variable immune response hence respond differently to HAART. All these factors might affect their $CD4^+$ count improvement post HAART initiation. In many cases the correlation between two repeated measurements decreases as the time span between those measurements increases (Hedeker, 2004). Hence a correlation structure that accounts for the distance between pairs of observations may be more admissible.

The general linear mixed model (equation 3.3) extends the general linear model by the addition of random effect parameters and by allowing a more flexible specification of the covariance matrix of the random errors. The matrix Z can contain continuous or dummy predictor variables, just like the matrix$X$. The general linear mixed model can be fitted to longitudinal data by assuming that the vector of repeated measurements on each subject follows a linear regression model where some of the regression parameters are population-specific (fixed-effects), but other parameters are subject-specific (random-effects). The subject-specific regression coefficients reflect how the response evolves over time for each subject. These subject-specific models can be very flexible, but in practice polynomials involving time will often suffice. However, extensions of this flexibility, such as fractional polynomial models or extended spline functions, can be considered as well (Verbeke and Molenberghs, 2000).

How the subject-specific regression coefficients relate to known covariates is reflected in the fixed-effect parameter estimates. These parameter estimates represent the population average. A model with the assumption that all variability in subject specific intercepts and slopes is attributed to the fixed-effect variables can be obtained by omitting the random-effect variables. Omitting only the random variables reflecting subject-specific slopes generates a random intercepts model.

The Linear Mixed Model is defined as:

$$Y_i = X_i \beta + Z_i b_i + \varepsilon_i \quad (3.3)$$

Where:

$Y_i$is the response vector for the ith subject: $Y_i = = (Y_{i1}, Y_{i2}, \ldots, Y_{ini})'$

$X_i$is a n_i x p design matrix of fixed effects

βstands for a p-dimensional vector of subject specific regression coefficients

$Z_i$is a n_i x $q$design matrix of random variables

$b_i$is a q-dimensional vector of random-effect parameters $b_i \sim N(0,D)$, b1, …, b_N are independent,

$\varepsilon_i$isn_i x 1 vector of random errors and $\varepsilon_i \sim N(0,\Sigma_i)$

D andΣ_i are variance components.

The model in equation 3.3 assumes both fixed and random effects. If $\Sigma_i = \delta^2 I_{ni}$ where I_{ni} denotes an identity matrix, then we call the model the `conditional-independence model' since it implies the n_i responses on individual i are independent, conditional on bi and β (Laird and Ware, 1982). Furthermore, $b_1, \ldots, b_N$ and $\varepsilon_1, \ldots, \varepsilon_N$ are assumed to be independent. The elements for the variance components are in the matrices D andΣ_i. In the so-called error component models, ε_1, can be decomposed into two components representing both subject specific variation and variation over time that is serial correlation (Hallahan, 2003). Therefore$\varepsilon_i = \varepsilon_{1(i)} + \varepsilon_{2(i)}$ where $\varepsilon_{1(i)}$ is the measurement error associated with i^{th} subject and $\varepsilon_{2(i)}$ is associated with serial correlation for i^{th}subject. $\varepsilon_{1(i)} \sim N(0,\delta^2 I_{ni})$ and $\varepsilon_{2(i)} \sim N(0,\tau^2 H_i)$.

The correlation matrix H is assumed to have its (j; k)th element given by $h_{ijk} = g(|tij - tik|)$ for some decreasing function g(.) with g(0) = 1. This means that the correlation between $\varepsilon_{2(ij)}$ and $\varepsilon_{2(ik)}$ only depends on the time interval between the measurements y_{ij} and y_{ik} and decreases if the length of this interval increases. A stochastic process with mean zero, constant variance and a correlation function which only depends on the time lag between the measurements, is often called (second order) stationary (Verbeke and Molenberghs, 2000). $\varepsilon_{2(i)}$ represents the belief that part of an individual's observed profile is a response to time varying stochastic processes operating within that individual (Verbeke and Molenberghs, 2000). The model given by equation (1) is restated more explicitly in matrix notation as

$$\begin{bmatrix} Y1 \\ . \\ . \\ . \\ YN \end{bmatrix} = \begin{bmatrix} X1 \\ . \\ . \\ . \\ XN \end{bmatrix} \beta + \begin{bmatrix} Z1 & . & . & . & 0 \\ . & . & & . & \\ . & & . & . & \\ . & & . & . & \\ 0 & . & . & . & ZN \end{bmatrix} \begin{bmatrix} b1 \\ . \\ . \\ . \\ bN \end{bmatrix} + \begin{bmatrix} \varepsilon 1 \\ . \\ . \\ . \\ \varepsilon N \end{bmatrix} \quad \ldots\ldots\ldots\

This gives us the between and within subject variation contained in the first and second components of V_i. Both D and $\Sigma_i$can be estimated by either the method of Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML) estimation. But Diggle *et al*. (1994, 2002) report that ML estimation produces biased estimators. Hallahan (2003) said that REML estimators are less sensitive to outliers than ML estimators.

3.12.1.1. **Marginal Models**

Assuming we are sampling from a Gaussian distribution then marginally

$$Y_i \sim N(X_i\beta, Z_i D Z_i' + \Sigma_i) \quad \text{(3.7)}$$

If one is interested in the fixed effects (population averages) then one should focus more on the marginal model. In this case random effects are considered nuisance parameters since the method does not use the Z matrix to predict the response Y, rather the role of random effects enter into the var(Y) under the components of variance in D.

The conditional mean and variance of Yi is given by

$$E(Y_i / b_i) = X_i \beta + Z_i b_i \quad \text{(3.8)}$$

and

$$V(Y_i / b_i) = \sum_i \quad \text{(3.9)}$$

The parameter β is the same for all patients and b_i is specific to patient$i$. If the random effects in b are zero implying that D is also zero then, the marginal and conditional distributions are the same and $\sum_i$is called the marginal variance. Thus, the conditional model for longitudinal Gaussian data can also be written as

$$Y_i / b_i \sim N(X_i \beta + Z_i b_i, \sum_i) \quad \text{(3.10)}$$

If for example, in a clinical trial one wants to measure individual specific drug efficacy then a conditional model will do the job. However, if the interest is on drug efficacy in the population, then the marginal model is the way to go.

3.12.1.2. Assumptions of Linear mixed Models

Before making inferences about a fitted mixed-effects model, we should check whether the underlying distributional assumptions appear valid for the data or not. There are two basic distributional assumptions for the general linear mixed effects model.

I. $\varepsilon_i \sim N(0, \delta^2 I)$

The within-group errors are independent and identically normally distributed; with mean zero and variance $\delta^2 I$ and they are independent of the random effects. This assumption can be relaxed by allowing to model non constant variances or special within group correlation structures.

II. $b \sim N(0, D)$

The random effects are normally distributed with mean zero and covariance matrix D. But, D is not depending on the group as it is the covariance matrix of the random effects and are independent for different groups.

3.13. Estimation of fixed effects

Recall that the general linear mixed model is given by $Y_i = X_i \beta + Z_i b_i + \varepsilon_i$ where $b_i \sim N(0, D)$and $\varepsilon_i \sim N(0, \Sigma_i)$, b_i and $\varepsilon_i$are independent and thus the marginal model is given by $Y_i \sim N(X_i \beta, Z_i D Z_i' + \Sigma_i)$. The inferences based on the marginal model do not explicitly assume the presence of random effects representing the natural heterogeneity between subjects (Verbeke and Molenberghs, 2000). The marginal likelihood function is given by

$$L_{ML} = \prod_{i=1}^{N} \{(2\pi)^{-n/2} |Vi(\propto)|^{-1/2} \exp\left(-\frac{1}{2}(Yi - Xi\beta)' V_i^{-1} (\propto) (Yi - Xi\beta)\right)\} \quad \text{.....(3.11)}$$

where $\propto$ is the vector of all variance components in D and $\sum_i$ and $\Theta = (\propto'; \beta')'$ is the vector of all parameters in marginal model. The log likelihood function for subject i is

$$l_i = -\frac{ni}{2}\log(2\pi) - \frac{1}{2} log|Vi| - \frac{1}{2}(Yi - Xi\beta)' V_i^{-1} (Yi - Xi\beta) \quad \text{..... (3.12)}$$

and

$$\frac{dli}{d\beta} = -X_i v_i^{-1} X_i \beta + X_i' V_i^{-1} \quad \text{..... (3.13)}$$

If $\propto$ were known, then the MLE of β on combining all the information from all the N subjects equals

$$\hat{\beta}(\propto) = (\sum_{i=1}^{N} X_i'V^{-1}X_i)^{-1} \sum_{i=1}^{N} X_i'V^{-1}y_i \quad \text{..... (3.14)}$$

In most cases α is not known and needs to be estimated as say $\hat{\propto}$, then V_i^{-1}should subsequently be replaced by $V_i(\hat{\propto})^{-1}$. The two frequently used methods to estimate α are maximum likelihood and restricted maximum likelihood. The method of restricted maximum likelihood was introduced by Patterson and Thompson in 1971. It was developed in order to avoid biased variance component estimates that are produced by ordinary maximum likelihood estimation. This is because maximum likelihood estimation of variance components takes no account of the degrees of freedom used in estimating fixed effects. This means that ML estimates of variance component have a downwards bias which increases with the number of fixed effects in the model. This in turn leads to underestimates of standard errors for fixed effects, hence leading to unnecessarily more conservative confidence intervals than expected.

3.13.1. Maximum likelihood estimation (ML)

The maximum likelihood estimation method finds the parameter estimates that are most likely to occur given the data. The parameter estimates are derived by maximizing the likelihood function, which is a mathematical expression that describes the joint probability of obtaining the data expressed as a function of the parameter estimates.

In maximum likelihood estimation $\hat{\propto}$ is obtained by maximizing the profile likelihood $L_{ML}(\propto,\hat{\beta}(\propto$)) with respect to $\propto$. The resulting estimate of $\hat{\beta}(\ \hat{\propto}_{ML})$ for β will be denoted by $\hat{\beta}_{\mathrm{ML}}$. The estimates $\hat{\propto}_{ML}$ and$\hat{\beta}_{\mathrm{ML}}$ can also be obtained from maximizing L_{ML} (Θ) with respect to Θ that is, with respect to both $\propto$ and β simultaneously (Verbeke and Molenberghs, 2000).

3.13.1.1. Restricted maximum likelihood estimation (REML)

The REML estimation method applies ML estimation techniques to the likelihood function. The difference is that, the REML estimation method is associated with a set of "error contrasts" rather than associated with the original observations. Therefore, it will lose degrees of freedom and give less biased estimates of the variance components. The bias issue cannot be neglected, especially when the number of parameters is not small relative to the total number of observations (Verbeke and Molenberghs, 2000).

Both ML and REML are based on the likelihood principle, which has the properties of consistency, asymptotic normality, and efficiency but REML corrects for the downward bias in the ML

parameters in $\boldsymbol{D}$ and $\sum$. REML handles strong correlations among the responses more effectively than ML and the differences in estimating between them increases as the number of fixed effects in the model increases. The difference between ML and REML is the construction of the likelihood function. However, the two methods are asymptotically equivalent and often give very similar results.

The distinction between ML and REML becomes important only when the number of fixed effects is relatively large. In that case, the comparisons unequivocally favor REML. First, REML copes much more effectively with strong correlations among the responses for the subjects than does ML. Second, REML estimates do not have the downward bias that ML estimates have because REML estimators take into account the degrees of freedom from the fixed effects in the model. Finally, REML estimators are less sensitive to outliers in the data than ML estimators. In fact, when the estimates do vary substantially, Diggle, Liang, and Zeger (1994) favor REML.

To explain the concept of REML estimation mathematically, consider a sample of N observations $Y_1, \ldots, Y_N$ from $N(\mu,\delta^2)$. For known μ, the MLE of δ^2 is given by

$$\hat{\delta}^2{}_u = \sum_{i=1}(Yi - \mu)^2 / N \quad \ldots\ldots (3.15)$$

and clearly $\hat{\delta}^2$ is unbiased for δ^2. When μ is not known, the MLE of δ^2 is now given by

$$\hat{\delta}^2{}_b = \sum_{i=1}(Yi - \bar{Y})^2 / N \quad \ldots\ldots (3.16)$$

One should note that for unknown μ, $\hat{\delta}^2{}_b$ is a biased estimator for δ^2, that is

$$E(\hat{\delta}^2{}_b) = \frac{N-1}{N}\delta^2 \quad \ldots\ldots (3.17)$$

The biased expectation of $\hat{\delta}^2{}_b$ implies that an unbiased estimate of δ^2, should be

$$S_u^2 = \sum_i \frac{(Y_i-\bar{Y})^2}{N-1} = \frac{N-1}{N}\hat{\delta}^2{}_b \quad \ldots\ldots (3.18)$$

The estimator $S_u^2$is unbiased because $E(S_u^2) = \frac{N-1}{N-1}\delta^2 = \delta^2$. Apparently having to estimate μ introduces bias in the MLE for δ^2. To estimate δ^2without having to estimate μ consider $Y \sim N(\mu; \delta^2 I_N)$, where $Y = (Y_1, Y_2, \ldots, Y_N)'$. Ycan be transformed such that μ disappears from the likelihood as follows.

Let U = $\begin{pmatrix} Y1 - Y2 \\ Y2 - Y3 \\ . \\ . \\ . \\ Y_{N-2} - Y_{N-1} \\ Y_{N-1} - Y_N \end{pmatrix}$ = A'Y ~ (0, δ^2A'A) ……..…. (3.19)

The distribution of U does not depend on μ (Diggle et al., 1994, 2002). The MLE of δ^2 based on U is

$$S_u^2 = \sum_i \frac{(Y_i - \bar{Y})^2}{N-1} \quad \text{……..} \quad (3.20)$$

where A defines the set of N-1 linearly independent error contrasts of the vector $Y = (Y_1, Y_2, \ldots, Y_N)$'.

Therefore S_u^2 is called the REML estimate of δ^2, which is independent of A. This estimation can be extended to linear regression models. To do this consider a sample of N observations (Y_1, Y_2, …, Y_N from a linear regression model where Y ~ N(Xβ,δ^2I). The MLE of δ^2 is

$$\hat{\delta}^2{}_b = (\text{Y-X}\hat{\beta})'(\text{Y-X}\hat{\beta})/\text{N} \quad \text{……} \quad (3.21)$$

which is biased for δ^2 since

$$\text{E}(\hat{\delta}^2{}_b) = \frac{N-p}{N}\delta^2 \quad \text{……} \quad (3.22)$$

From the biased expression one can derive an unbiased estimate mean square error (MSE) given by

$$\hat{\delta}^2{}_u = \frac{(\text{Y}-\text{X}\hat{\beta})'(\text{Y}-\text{X}\hat{\beta})}{N-p} \quad \text{……..} \quad (3.23)$$

The MSE can also be obtained from transforming the data orthogonal to X like in the simple example above. The MSE is called the REML estimate of δ^2. So far we have done the estimation of variance for the normal population and also in the linear regression model. Now we consider a similar estimate in the case of linear mixed model. Consider models where Yi ~N(Xβ, V_i) where V_i was defined in equation (3.6) under the formulation of the linear mixed models. If we combine the subject-specific sub-models into one model we get Y~N(Xβ; V (∝)).

where V(∝) = $\begin{pmatrix} V1 & . & . & . & 0 \\ . & . & & & . \\ . & & . & & . \\ . & & & . & . \\ 0 & . & . & . & Vn \end{pmatrix}$

The REML estimator for the variance component $\propto$ is obtained from maximizing the likelihood function of a set of error contrasts U = A'Y where A is (n $_X$ (n -p)) matrix with columns orthogonal to X matrix such that U = A'Y ~ N(0, A'V ($\propto$)A), which is not dependent on β anymore. Verbeke and Molenberghs (2000) reported that the REML estimators for $\propto$ and for β can be found by maximizing the so-called REML likelihood function

$$L_{\text{REML}}=\left|\sum_{i=1}^{N} X_i' V_i^{-1} X_i\right|^{-\frac{1}{2}} L_{\text{ML}}(\Theta) \quad \ldots\ldots\ldots\

interest is only in inferences for the mean structure, little effort should be spent in modeling the covariance structure if the data set is sufficiently large (Verbeke and Molenberghs, 2000). A common problem in statistical analysis is that of testing hypothesis about group comparison or contrasts. For any known matrix L, consider testing the hypothesis

$$\left\{\begin{array}{c} Ho: L\beta = 0 \ldots\ldots\ldots\

reduces to a t-test. Having dealt with Wald, t- and F-tests one needs to consider the likelihood-ratio test (LR) as it is an alternative test that can be used to test hypotheses about model parameters. A likelihood-ratio test is a statistical test for goodness of fit between two nested models. A complex model is compared to a relatively simpler model to see if it fits the data significantly better. It is a statistical test for making a decision between two hypotheses based on the value of the ratio of the likelihood under a restricted parameter space compared to an unrestricted one.

The null hypothesis of interest is H0: β is an element of $\Theta_{\beta,0}$ for some subspace $\Theta_{\beta,0}$ of the parameter spaceΘ_{β} of the fixed effects β. Suppose $\hat{\beta}$ is the value of β that maximizes the likelihood function L(β)under Θ_{β} and $\hat{\beta}_0$ is the value of β which maximizes the likelihood under $\Theta_{\beta,0}$ pertaining to someof the elements of β. Then, the likelihood-ratio test statistic is given by: $\lambda = \frac{L(\hat{\beta}0)}{L(\hat{\beta})}$. We reject H_0declaring it unsupported by data if H_0 is too small is an indication that there are other hypotheses which are much better supported by data. It is often easier to use the negative of twice its logarithm:

$$-2\log\lambda = -2\log L(\hat{\beta}0) + 2\log L(\hat{\beta}) \quad \text{(3.29)}$$

which has an approximate null χ^2 distribution with degrees of freedom equal to the difference in dimension of $\Theta_{\beta}$and$\Theta_{\beta,0}$. This distributional approximation is exact in the normal case. In otherwords, -2logλ has a null distribution with degrees of freedom equal to the difference in number of parameters in complex model and simple model. By stating the likelihood as L($\hat{\beta}$), we have deliberately suppressed the presence of other parameters which we here assume known. Otherwise, the estimation and the inference procedure will have to take into account of other parameters both nuisance and those of importance in the model. For hypothesis testing regarding parameters of a regression model we assume that the distributional assumptions also hold under the following conditions:

- In the case of a nested model within the complex model we can resort to simple model from the complex model simply by fixing some parameters of complex model to zero.
- The same response variable set is employed to fit both complex and simple models

In case it is not clear that the hypothesis in question does not involve nested models, then other decision making criteria such as the Akaike information criteria (AIC) can be used.

3.14.2. Inference for the variance components

The inference for the mean structure is usually of primary interest. However, inferences for the covariance structure are of interest as well especially for the interpretation of the random variation in the data. One should also take note that over parameterized covariance structures may lead to inefficient inferences for the mean and likewise too restrictive covariance model may invalidate inferences for the mean structure. The ML and REML estimates of $\propto$ are asymptotically normally distributed with a mean $\propto$ and inverse Fisher information matrix $I^{-1}(\propto)$as covariance. Thus, the Wald and the F tests can also be used to test hypotheses about the variance components too.

3.15. Random Coefficient Model

Random coefficient models are useful for highly unbalanced data with many repeated measurements per subject. In random coefficient models, the fixed effect parameter estimates represent the expected values of the population of intercepts and slopes. The random effects for intercept represent the difference between the intercept for the ith subject and the overall intercept. The random effects for slope represent the difference between the slope for the i^{th}subject and the overall slope. Random coefficient models also have a random effect for the within-subject variation. Because there is too few data on a single subject to estimate its regression parameters, and to avoid theoretical obstacles, it is assumed that the random effects are independently and identically distributed random variables (Verbeke and Molenberghs 2000).

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + b_{0i} + b_{1i} Z_{ij} + \varepsilon_{ij} \quad \ldots\ldots\ldots\

Competing models may be ranked according to their AIC, with the one having the lowest AIC being the best. It should be strongly emphasized that information criteria only provide rules of thumb to discriminate between several statistical models and they should never be used or interpreted as formal statistical tests of significance (Verbeke and Molenberghs, 2000). For the comparison of models with different mean structures, information criteria should be based on ML rather than REML, as otherwise the likelihood values would be based on different sets of error contrasts, and therefore would no longer be comparable.

3.17. **Missing Data**

In this section, we discuss about missing data, assumptions of missing data and methods of handling missing data.

Missing data commonly occur in longitudinal studies and are defined as no data values are stored for a variable or variables. Missing data may arise due to several different reasons including refusal, attrition, measurement errors or simply ignorance of interviews. In this study's data, the missing data was due to dropout, lost to follow up, transfer out and death. No matter what the reason is, missing of observations is a problem that has to be dealt with in all statistical areas (Allison, 2001). Besides making sure that the missing observations really are missing observations (Schafer and Graham, 2002), assumptions, explicitly or implicitly, about the missing data mechanism are always made.

The problem of missing observations can frequently occur in all types of clinical trials, especially when observations are measured repeatedly at each scheduled visit for the same subject in a longitudinal study. In longitudinal studies, there are many possible causes leading to missing data including the duration of the study, the nature of the disease, the efficacy and adverse effects of the drug under study, accidents, patients' refusal to continue, moving, or other administrative reasons. Frequently, missing data can potentially lead to two serious problems in statistical practice: reducing the overall statistical power and having biases in the estimates.

Missing data is a problem since nearly all standard statistical methods presume complete information for all the variables is included in the analysis. A relatively few absent observations on some variables can dramatically shrink the sample size. As a result, the precision of confidence intervals is harmed, statistical power weakens and the parameter estimates may be biased. Appropriately dealing with missing can be challenging as it requires a careful examination of the

data to identify the type and pattern of missing data, and also a clear understanding of how the different imputation methods work.

When exploring missing data, it is important to come to a conclusion about the mechanism of missing data, that is, the hypothesized reason for why data are missing. This can range from arbitrary or random influences to purposeful patterns of non-response. There are various methods of dealing with missing data that range from simple classical methods to model based methods. Furthermore, each method is based upon a specific missing data mechanism but one needs to realize that at the heart of the missing value problem it is impossible in practice to identify the missing data mechanisms. In a longitudinal setting, each unit is measured on several occasions and hence it is not unusual in practice for some sequences of measurements to terminate early for reasons outside of the control of the experimenter or investigator and any unit so affected is called a dropout.

The three missing data mechanisms according to Rubin, 1977 are:

- ✓ Missing completely at random (MCAR) if the missingness is independent of both unobserved and observed data
- ✓ Missing at random (MAR) if conditional on the observed data the missingness is independent of the unobserved measurements,
- ✓ Missing not at random or non-random (MNAR) if the process is neither MCAR nor MAR.

In the context of likelihood inference, and when the parameters describing the measurement process are functionally independent of the parameters describing the missingness process, then MAR and MCAR are ignorable while a MNAR process is non-ignorable.

3.17.1. Handling Missing Data

Several different methods and strategies are available to handle missing data. Graham (2009) defines three conditions that should be satisfied in a good method. First, the method should yield unbiased estimates of a variety of different parameters. Second, the method should include a way to assess the uncertainty about the parameter estimates, and third, the method should have good statistical power. It is further good to remember that the goal of any missing data procedure is not to recreate the missing observations but rather to retain the characteristics of the data and the associations between variables such that valid and efficient inferences can be made (Schafer and Graham, 2002).

3.17.1.1. **Maximum Likelihood (ML)**

We can use the ML method to get the variance-covariance matrix for the variables in the model based on all the available data points, and then use the obtained variance- covariance matrix to estimate our regression model (Schafer, 1997).

There are two main ML methods:

a. Direct Maximum Likelihood: It implies the direct maximization of the multivariate normal likelihood function for the assumed linear model. Advantage: It gives efficient estimates with correct standard errors. Limitations: It requires specialized software (it may be challenging and time consuming).
b. The Expectation-maximization (EM) algorithm: It provides estimates of the means and covariance matrix, which can be used to get consistent estimates of the parameters of interest. It is based on an expectation step and a maximization step, which are repeated several times until maximum likelihood estimates are obtained. It requires a large sample size and that the data are missing at random (MAR) (Allison, 2001).

3.17.1.2. **Multiple Imputations**

Multiple imputation is the popular method for handling missing values in a data. It replaces each missing item with five or more acceptable values, representing a distribution of possibilities (Rubin, 1987). The advantage of the method is that once the imputed data set have been generated, the analysis can be carried out using procedures in virtually any statistical package, which makes the analysis simple. Also, standard error or p-value obtained from MI are generally valid because they incorporate uncertainty due to missing values. MI can be highly efficient even if the number of imputation is relatively small, especially when between-imputation variance is not too large.

The multiple imputation inference involves three distinct phases (Rubin, 1987):

a. The missing data are filled *m* times to generate *m* complete data sets
b. The *m* complete data sets are analyzed by using appropriate procedures
c. The results from the *m* complete data sets are combined for the inference.

The method of imputations mechanism choice depends on the type of missing data pattern. For monotone missing data patterns, either a regression method or propensity score method can be used. For an arbitrary missing data pattern, a Markov chain Monte Carlo (MCMC) method can be

used. Without the detail of theatrical methods, regression method is fitted for each variable with missing values with previous variables as covariates. Propensity Score method is that observations are grouped based on propensity scores, and an approximate Bayesian bootstrap imputation is applied to each group. At last, MCMC constructs a Markov chain long enough for distribution of the elements to stabilize to a common distribution (Yuan, 2000).

3.18. Model Comparison Techniques

The primary objective of model comparison is to choose parsimonious model that best fit the data. In this study, we used information criteria and Likelihood ratio test (LRT) to select the best model.

Akaike's information criterion (AIC) is a measure of goodness of fit of an estimated statistical model (Agresti, 1996). It is not a test on the model in the sense of hypothesis testing; rather it is a tool for model selection. The AIC penalizes the likelihood by the number of covariance parameters in the model, therefore AIC= -2Log (L) +2p where, L is the maximized value likelihood function for the estimated model and p represent the number of parameters in the model. The model with the lowest AIC value is preferred.

Likelihood ratio Test: It is constructed by comparing the maximized log likelihoods for the full & reduced models respectively and its test statistics is defined as:

$T^2_{LR} = -2\ln\lambda N = -2\ln\left(\frac{Lml(\hat{\alpha}Ml,o)}{Lml(\hat{\alpha}Ml)}\right)$ where,$\hat{\alpha}Ml, o$, and $\hat{\alpha}Ml$ are respective maximum likelihood estimates which maximize the likelihood functions of the reduced and full model. The asymptotic null distribution of the LR test statistic is a chi-square distribution with degrees of freedom equal to the difference between the numbers of parameters in the two models.

3.19. Model Checking Technique

In linear mixed effects model, it is assumed that the random effects are normally distributed and uncorrelated with the error term. Residual plots can be used visually to check normality of these effects and to identify any outlying effect categories. Examining the plot of the standardized residuals versus fitted values by any covariates of interest can give a better feeling (Molenberghs, 2008**)**. The assumption of normality for the within-group error was assessed with formal tests and the normal probability plot of the residuals by covariates. Similarly, normality of the random effects is assessed using Normal Plot of each random effect. Normal plot of estimated random effects helps for checking marginal normality and to identify outliers.

CHAPTER FOUR

RESULTS AND DISCUSION

4.1.Descriptive Statistics

In this section, exploratory data analysis of the $CD4^+$ count of HIV positive patients enrolled in the ART case unit of **Yekatit 12** Hospital from September 10, 2009 to September 9 2012 G.C and followed until January9, 2016G.C, in Addis Ababa was carried out. The total number of patients included in this study was 217 of whom 141(65%) were females and the rest 76(35%) were males. The aim of exploratory data analysis is to understand the data structure and determine the relevant modeling approaches suitable for the data.

4.1.1. Descriptive Statistics of Continuous Variables

The mean $CD4^+$ count of patient's increases with an increasing rate until time three, then increases slowly to time six and increases at an increasing rate to time seven to the maximum average 466.722 with a standard deviation of 233.172 at the last time point. The number of measurements for $CD4^+$ count showed increasing and decreasing observations between follow up times for the response indicating that the data have both intermittent and dropout missing observations.

Table 4.1: Descriptive Statistics of Continuous variables

Time	Baseline	1	2	3	4	5	6	7
N	217	179	191	162	168	141	116	97
Mean (CD4)	264.8848	351.3519	390.1518	410.2716	423.4643	426.6525	446.9138	466.72165
StdDev	210.9259	210.5959	231.6899	233.8812	209.3979	197.2816	209.0785	233.1719

From Table 4.2, we can observe that the mean baseline age of patients was 39 with a standard deviation of 10.888; the mean baseline $CD4^+$ count was 264.3641 with a standard deviation of 210.93 and the mean weight at baseline for HIV positive patients was 56.945 with a standard deviation of 12.376.

Table 4.2: Descriptive Statistics of Continuous Covariates at baseline

No	Variable	Age	BaseCD4	Weight
1	Mean	39.1106	264.3641	56.9447
2	StdDev	10.8878	210.9311	12.3759

Figure 4.1 shows the plot of mean CD4+ count over time. The plot indicates that the mean CD4+ count shows an increasing progress from baseline to the last time point and follows almost a linear trend.

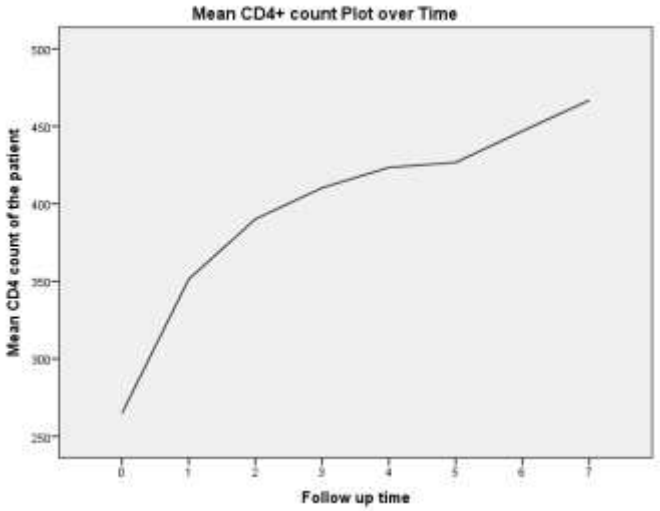


Figure 4.1: Mean $CD4^{+}$ Count Plot by Time

4.1.2. **Descriptive Statistics of Categorical Covariates by Time**

Descriptive statistics of categorical covariates by follow up time are presented in Table 4.3.The number of patients available for measurement decreases from 217 to 112 as the time of follow up increases from baseline to time 7 and the number of female patients were greater than male patients at all levels of follow up time, where time of follow up indicates (0 = Baseline, 1 = Month Six, 2 = Month 12, 3 = Month18, 4 = Month24, 5 = Month30, 6 = Month36, 7 = Month 42). Only 51.61% of patients are followed until follow up time seven (Month42) inducing 48.39% missing values.

Table 4.3: Descriptive Statistics of categorical Covariates by Time

		Follow up time in every six months							
		0	1	2	3	4	5	6	7
Sex of the patient	M	76	76	76	76	71	65	52	39
	F	141	141	141	140	139	120	96	73
Total		217	217	217	216	210	185	148	112
Marital status of the patient	Married	148	148	148	147	144	134	110	83
	Single	69	69	69	69	66	51	38	29
ART Regimens Taken	AZT-3TC-EFV	76	76	76	76	74	58	44	28
	AZT-3TC-NVP	56	57	58	58	56	50	39	31
	TDF-3TC-NVP	25	25	26	25	26	25	22	17
	d4t-3TC-NVP	11	10	8	8	6	6	4	3
	TDF-3TC-EFV	49	49	49	49	48	46	39	33
Total		217	217	217	216	210	185	148	112
Functional Status of the Patient	work	189	189	189	188	184	161	128	99
	Bed redden	17	17	17	17	16	14	11	7
	Ambulatory	11	11	11	11	10	10	9	6
Total		217	217	217	216	210	185	148	112
WHO-clinical Stage of the patient	Stage1	66	66	66	66	63	57	43	30
	Stage2	58	58	58	58	57	51	45	38
	Stage3	66	66	66	65	65	56	48	35
	Stage4	27	27	27	27	25	21	12	9
Total		217	217	217	216	210	185	148	112

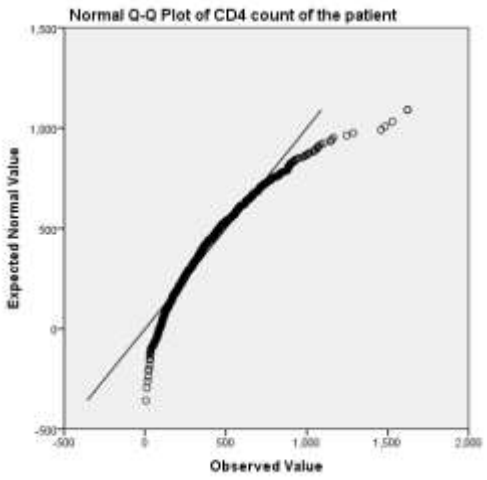
4.1.3. Distribution of $CD4^+$ Count Data

In any data analysis, before directly going to the analysis part firstly the assumptions of the data must be checked. In our $CD4^+$ count data Shapiro wilk test and q-q plots depicted in Table 4.2.1. and Figure 4.2 respectively were used for checking the normality of the data. Count data can be well approximated by a normal distribution when the number of the count becomes large. To normalize the data logarithmic and square root transformation were carried out. The logarithmic transformation doesn't make a visible change of normality, thus we used square root transformation. From the Shapiro wilk tests of normality in Table 4.4., we found that the actual

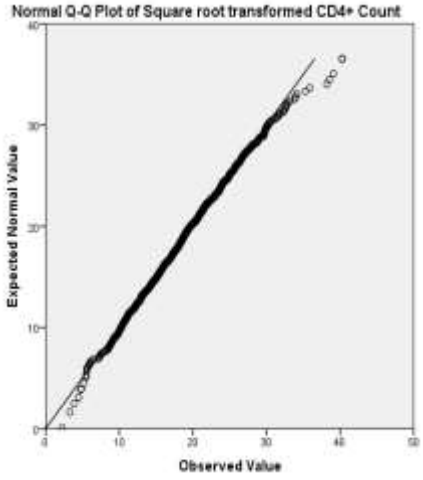
CD4+ counts were not normal at all-time points as the test showed significant deviation from normality. Likewise, the logarithmic transformed $CD4^+$ count were not normal by the Shapiro wilk test except at the last two time points. But, the test approves normality of square root transformed $CD4^+$ count at all-time points except at baseline, time three and time six. This coincides with Figure 4.2 which takes the q-q plot of the overall data suggesting square root transformation of the data to normality and detail of the q-q plots by time are depicted in Figure A_1 on the appendix.

Table 4.4: Shapiro-Wilk tests of Normality

Follow up time	Actual CD4 Count			Square Root CD4Count			Logarithm of CD4 Count		
	Tests of Normality			**Tests of Normality**			**Tests of Normality**		
	Shapiro-Wilk			Shapiro-Wilk			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.	Statistic	df	Sig.
0	.837	217	.000	.967	217	.000	.957	217	.000
1	.916	179	.000	.992	179	.452	.948	179	.000
2	.909	191	.000	.988	191	.093	.969	191	.000
3	.897	162	.000	.982	162	.034	.965	162	.000
4	.961	168	.000	.989	168	.215	.959	168	.000
5	.964	141	.001	.993	141	.688	.974	141	.008
6	.888	116	.000	.973	116	.018	.983	116	.136
7	.890	97	.000	.976	97	.068	.985	97	.340



A.



B.

C.

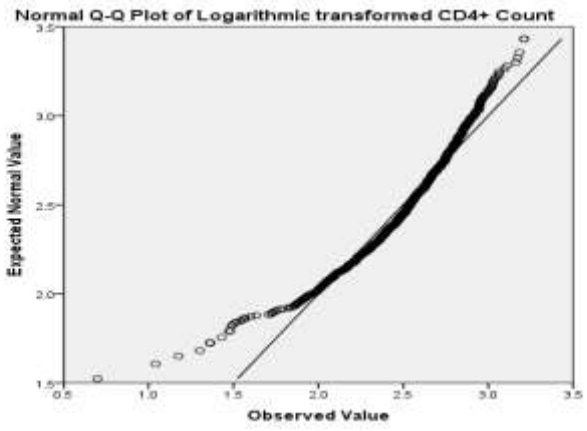


Figure 4.2. Normal Q-Q Plot of $CD4^+$ counts

The box plots in Figure 4.3 showed that, the outlier observations observed were minimized in the square root transformed CD4+ count than logarithmic transformed and actual CD4+ counts of patients. Hence, the Figure suggests square root transformation of the data.

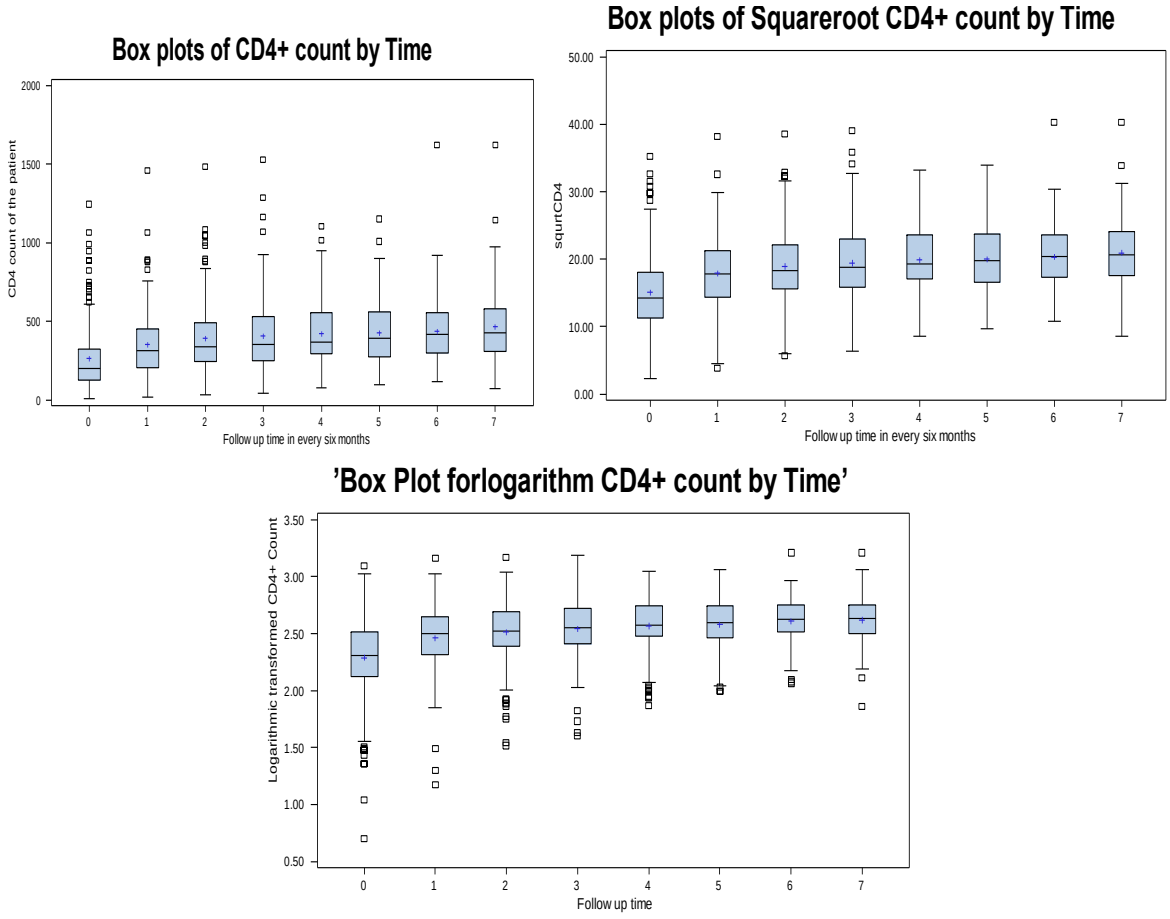


Figure 4.3 Box Plots of $CD4^+$ count by Time

4.2.Exploratory Data analysis

4.3.6. Exploring Individual profiles

Figure4.3 and Figure4.4 show individual profiles of all patients. Plotting observed profiles over time helps to identify general trends within patients and may detect change over time that provides information about the variability at given times. Connecting the repeated measurements for each patient's $CD4^+$ count over time shows that there is a non-constant pattern common to most patients. These individual profiles can also provide some information on between patient's $CD4^+$ count variability and illustrate that there is change among patient's $CD4^+$ count over time. Similarly, it appears that, there is a fluctuation in $CD4^+$ count over time after they were initiated on ART and the variability of the patient's $CD4^+$ count seemed larger at the beginning compared to the end in both $CD4^+$ count and square root transformed $CD4^+$ count of patients. The variation at the beginning was higher than at the end of the follow up time periods.

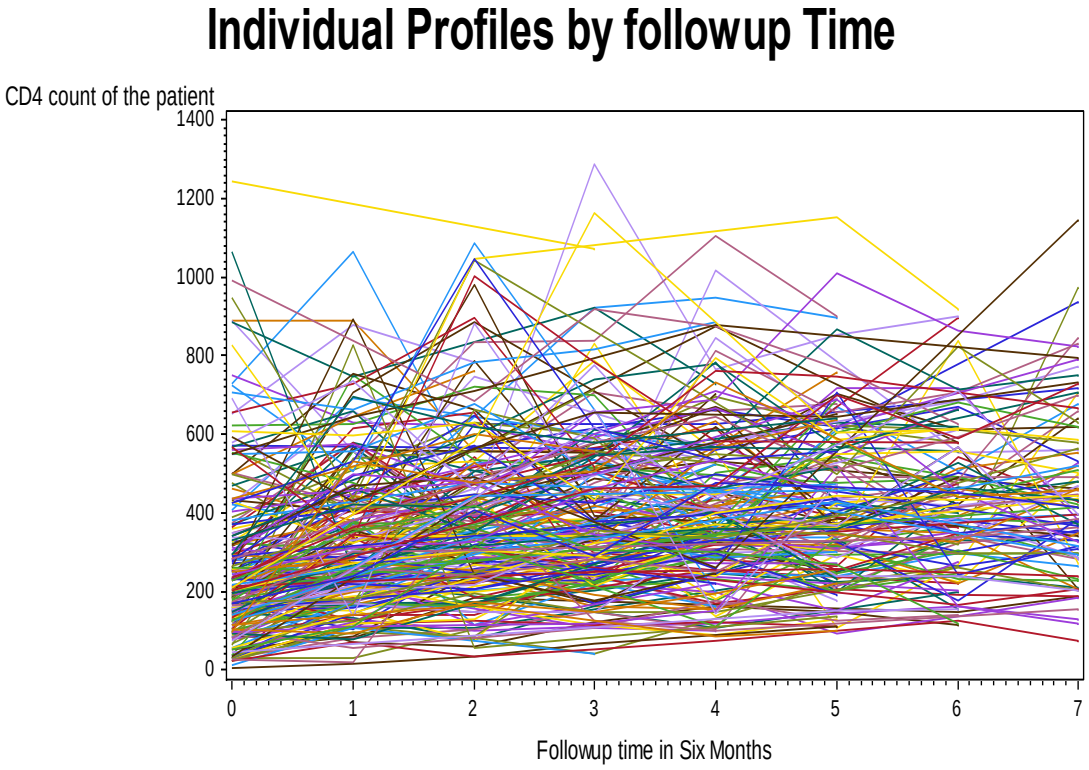


Figure 4.4. Individual Profile plot of $CD4^+$ count by Time

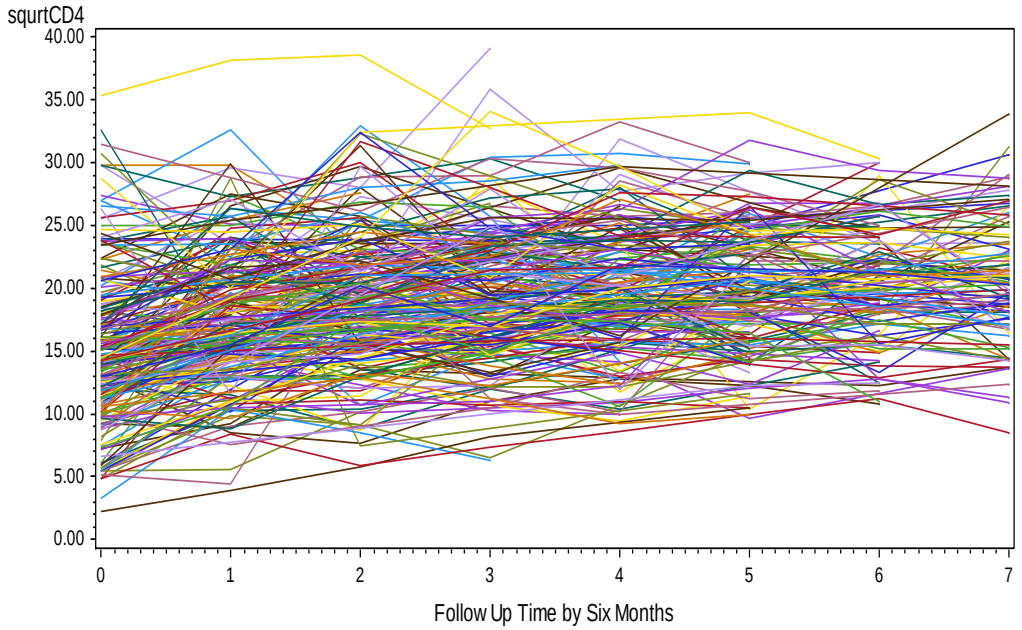


Figure 4.5: Individual Profile Plot of Square Root $CD4^+$ Count by Time

Figure 4.5 depicts the individual patient profile plot of $CD4^+$ counts by sex. The plot shows that there is an observed variability in $CD4^+$ count over time within patients and between patients. In the graph there seems almost the same variability in the progression of $CD4^+$ count of patients over time by sex group. But, it shows us that there is variability in $CD4^+$ count progression within males and females.

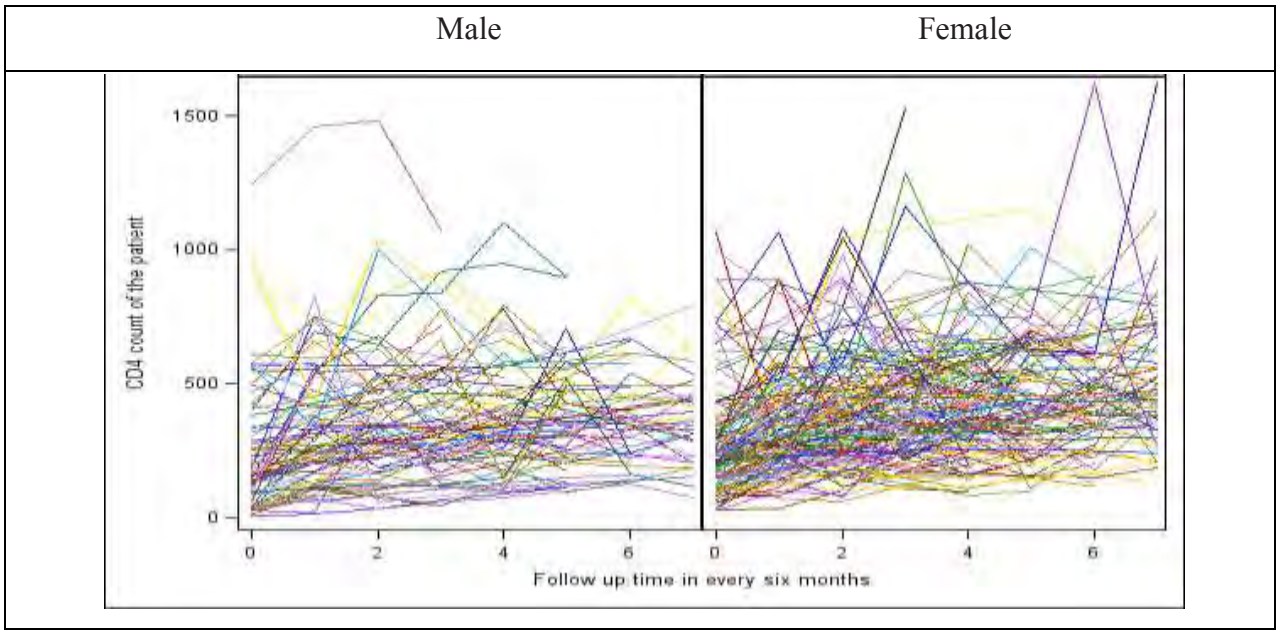
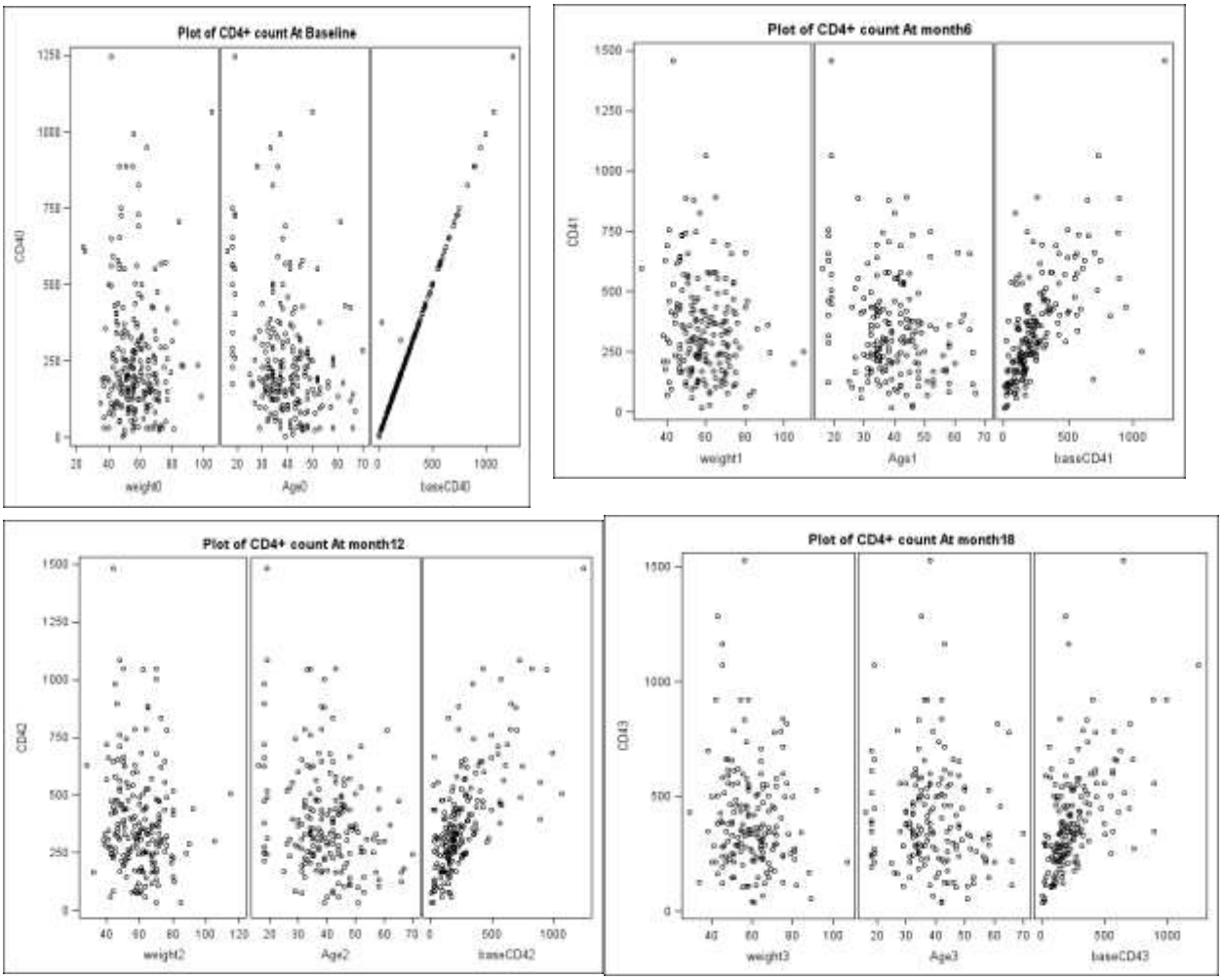


Figure 4.6. Individual Profile plots of the response by Time and Sex

4.3.6.1. **Scatter plots for CD4$^+$ count**

Several relationships between the response variable versus continuous predictors were investigated. The predictor's weight, age and baseline CD4$^+$ count were plotted against CD4$^+$ count to assess relationship between these measured covariate variables. The scatter plots for CD4+ counts were plotted at different follow up as shown in Figure 4.6.The plots suggested a strong correlation between baseline CD4+ count and CD4+ count at all follow up times. We can also observe that, weak negative correlation existed between the baseline age and CD4+ count. The correlation between the plots of CD4+ count and weight seemed to have weak negative correlation at all-time points.



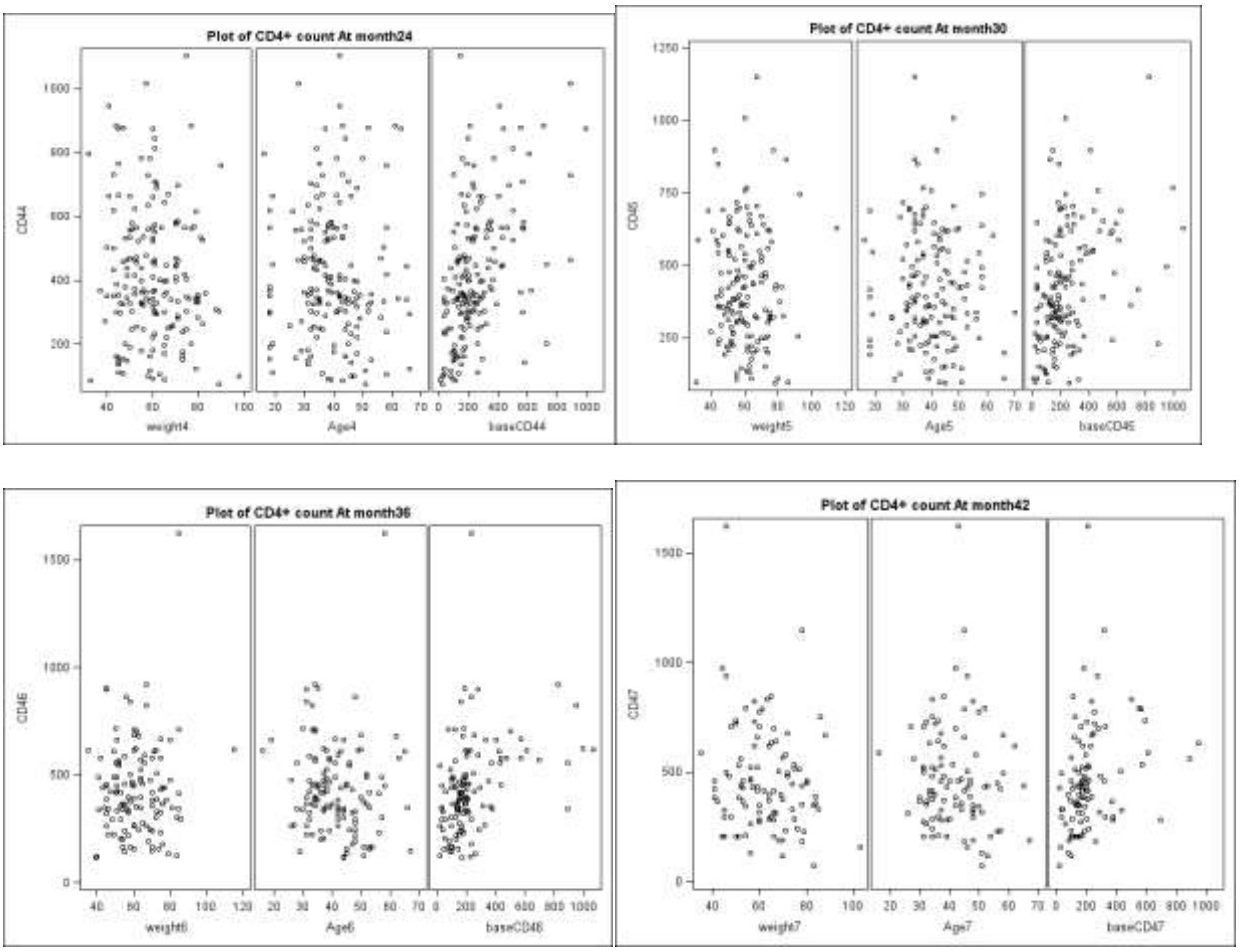


Figure 4.7. Scatter plots of $CD4^+$ count against Covariates by time

4.3.7. Exploring the Mean Profile

Figures 4.7 show the average evolution of $CD4^+$ counts of patients initiated on ART over time. From the Figures, we observed almost a sharp linear increase up to time two and then goes by following some slight increase in mean $CD4^+$ counts up to the last time of follow up. The mean profiles of CD4+ count of female patients was higher than males at baseline, the same at time1 and mean profiles of female patients was higher than males throughout the other follow up times. Generally, the mean $CD4^+$ count of patients who initiated on ART seems to show an increasing progress over time for all covariates even if the amount varies from group to group with in a covariate.

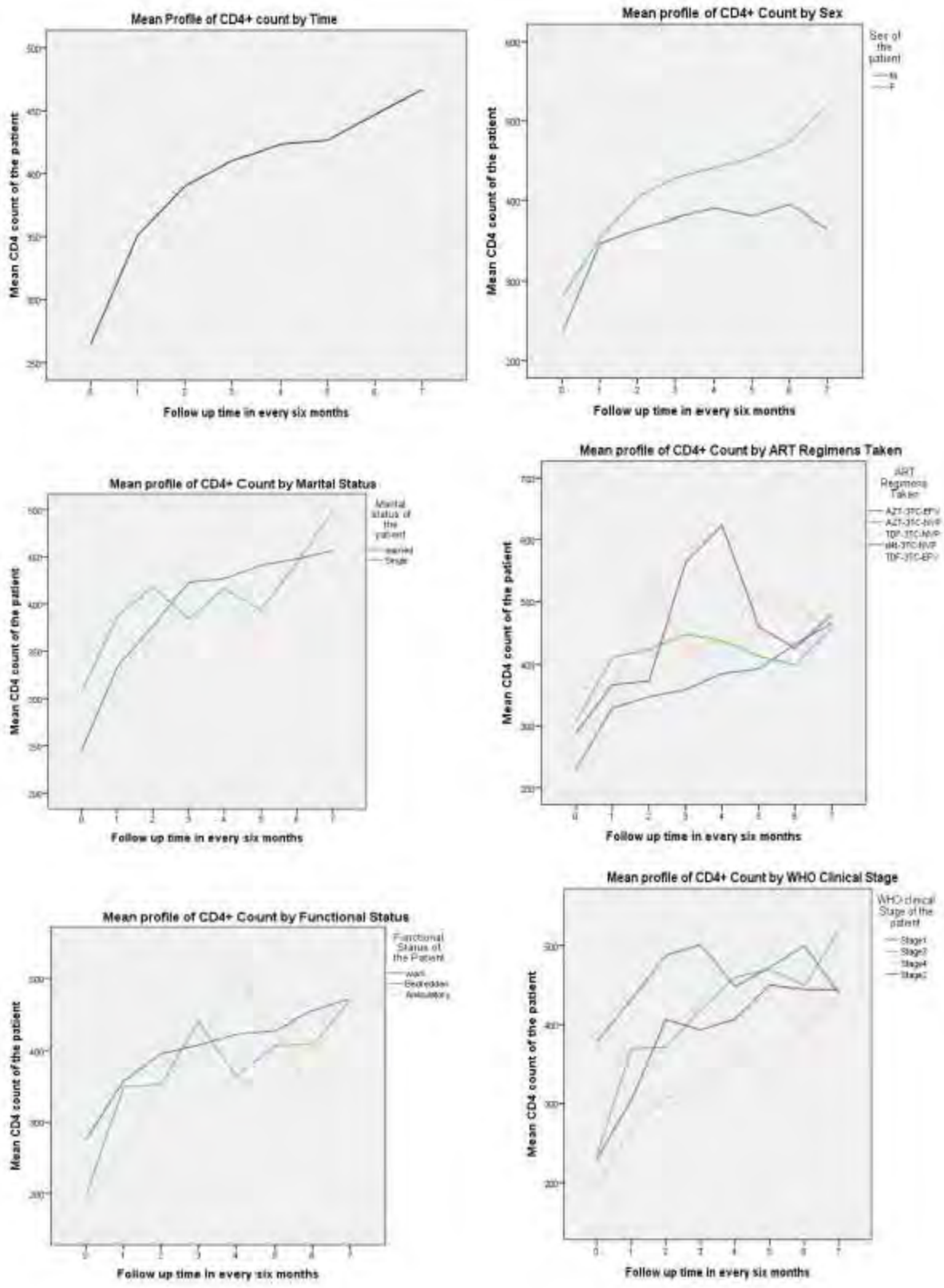


Figure 4.8: Mean Profile of Patients $CD4^{+}$ count of Categorical Covariates

4.3.8. Exploring Variance profiles

From Figure 4.8, it can be observed that the variance of patients $CD4^+$ count increases at a decreasing rate until time one, increases at an increasing rate to time three and then decreases with an increasing rate up to time five and then increases at an increasing rate to time seven. This implies that, spatial variance structures would be needed in modelling of patients $CD4^+$ count over time rather than the constant one. The variability of patients $CD4^+$ count for males is higher than that of females around the beginning of the follow up time. At time three, the variability is almost the same for both genders and stayed the same until time and then from time five until the last follow up time the variability of patients $CD4^+$count was higher for females than males. Overall, the variance of patients $CD4^+$ count fluctuates from time to time in different covariates.

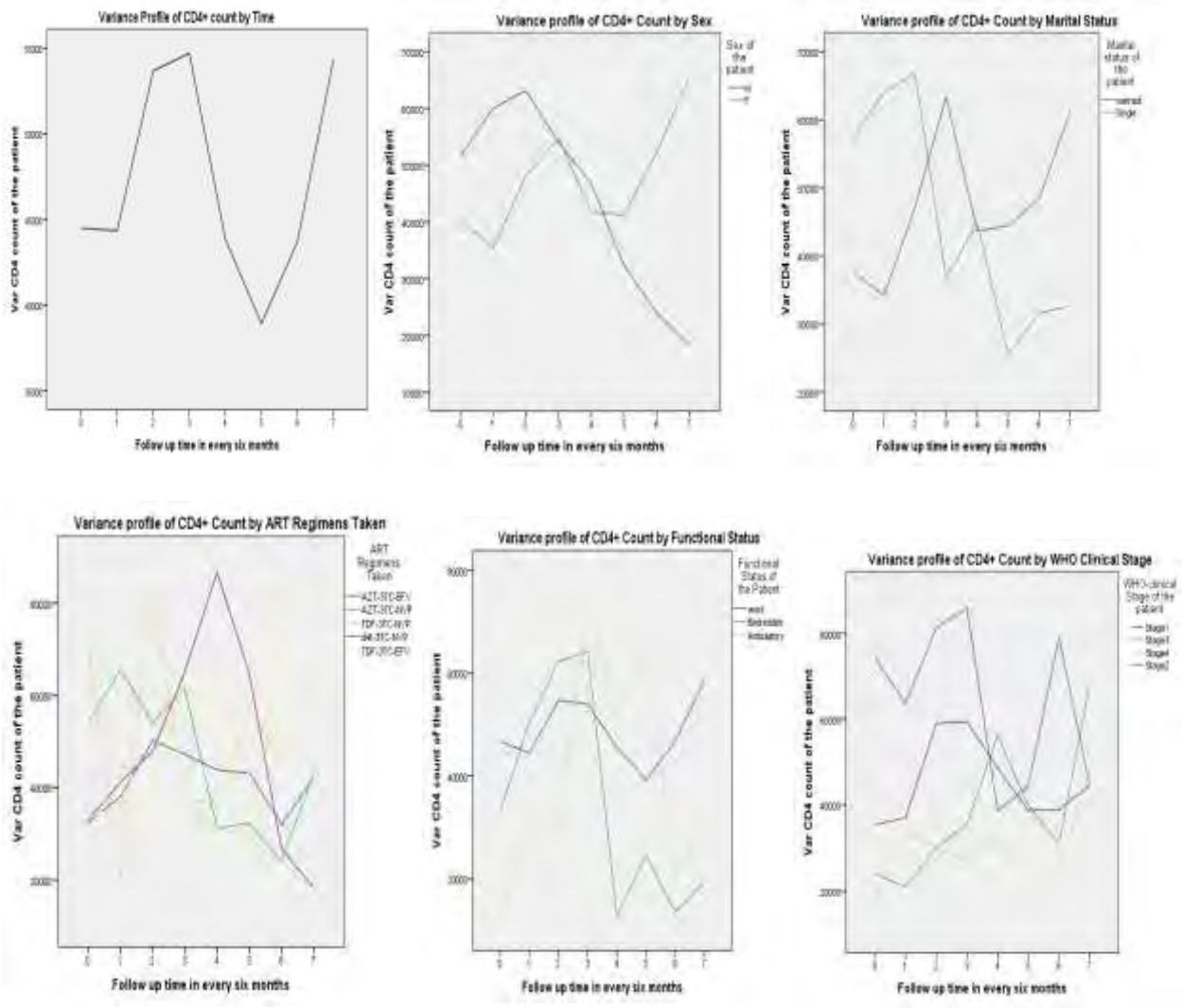
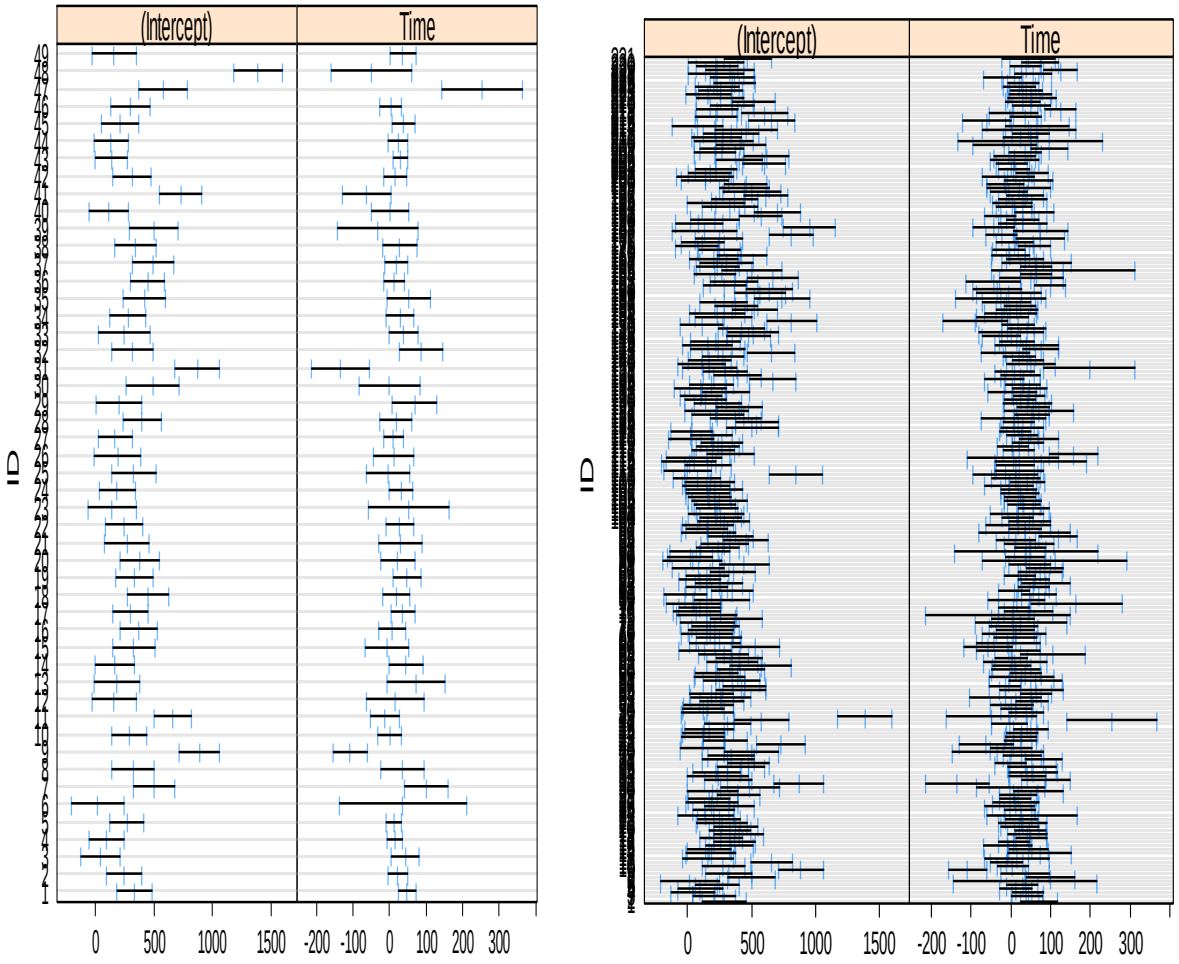


Figure 4.9. Variance Profile of $CD4^+$ count by Categorical Covariates over Time

4.3.9. Exploring the Random Effect

The main purpose of exploring the random effect is to get clues of what random effect structure to be used in modelling the linear mixed effect models. To explore these effects, we used the estimation of confidence intervals to draw the interval plot for estimated coefficients. As shown inFigure4.9A and Figure 4.9B, for sample of patients and the total number of patients respectively, the intercept and slop are random.



A. B.

Figure 4.10 Confidence Interval Plot of random intercept and Linear Time Effect

4.3.10. Exploring correlation structure and Scatter plot of Correlation matrix

The correlation matrix can be used to show the correlation between repeated measurements. The correlation in the presence of dropouts and missing's could be computed in various ways. In this study, we used the available data for computing the correlation matrix. The correlation matrix shown in Table 4.5 revealed a positive correlation between any two repeated measurements except for the last three time points when the correlation deemed decreasing over time.

Table 4.5: Correlation Matrix of $CD4^+$ count over Time.

	Baseline	**Time1**	**Time2**	**Time3**	**Time4**	**Time5**	**Time6**	**Time7**
Baseline	1.000	0.8383	0.7455	0.4858	0.1812	-0.0265	-0.1189	0.0576
Time1		1.000	0.6029	0.5438	0.1251	-0.0414	-0.0696	0.2061
Time2			1.000	0.5765	0.2246	0.0536	-0.1577	-0.1577
Time3				1.000	0.1548	-0.2385	0.0143	-0.0995
Time4					1.000	0.1493	0.0252	-0.1177
Time5						1.000	-0.0899	0.1731
Time6							1.000	0.0374
Time7								1.0000

Pair wise scatter plots were used for exploring the correlation between any two repeated measurements of CD4+ count of patients and it appears that, there was a positive and linear relationship between CD4+ counts at different time points (Figure4.10).

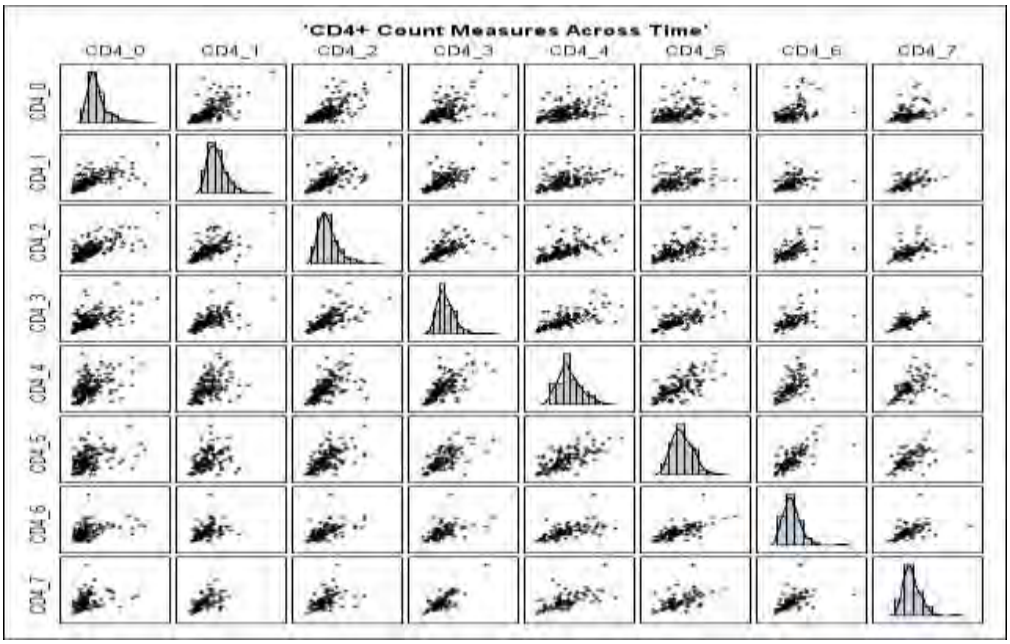


Figure 4.11: Scatter plot of correlation matrix

4.4. Linear Mixed Models

4.4.1. Univariate Marginal Models

We started by fitting univariate marginal models to explore the relationship between square root of $CD4^+$ count and each of the covariates. For this analysis six categorical and three continuous variables were considered. Variables significant at 0.25 level of significance were time, sex, age, marital status, baseline $CD4^+$ count, regimen class, functional status, who stage and the interaction terms of age, marital status, baseline $CD4^+$ count, regimen class, functional status and WHO clinical stage with time.

4.4.2. The full Marginal Model Including Variable selection

This section is devoted to selection of correlation structure and variables using REML and ML method of estimation respectively.

4.4.2.1.Selection of Covariance structure

Covariance structures should be carefully selected in order to obtain valid inferences for the parameters of fixed effects in the model. If one ignores important correlations by using a model that is too simple, one risks increasing type I error rate and underestimates standard errors (Little*et al.*, 2006). On the other hand, if the model is too complex, the power and efficiency is sacrificed (Little *et al.*, 2006). The covariance structure that fits data best based on the AIC was used to estimate the fixed effect parameters. In fitting the linear mixed model, a series of correlation structures of the longitudinal data of $CD4^+$ counts on HIV positive patients were considered. From the possible covariance structures in fitting the model, the one with the smallest AIC and convergence of the model in REML and ML was considered. Hence, AR (1) was selected for assessing the progression of CD4+ counts by linear mixed model.

4.4.2.2. Variable selection

The model development process was carried out by selecting covariates that have the potential to be included in the linear mixed models. Variable selection was done using the manual backward selection criteria. It was carried out based on maximum likelihood method of estimation and using the selected covariance structure, AR (1), as discussed under section 4.3.2.1.Accordingly, the selected variables were: time, age, baseline $CD4^+$ count, WHO clinical stage and the interaction terms of sex, age, baseline $CD4^+$ count and regimen class with time.

4.4.2.3.Correlation Structure Checking

After selecting variables of linear mixed models as discussed in Section 4.3.2.2, we refitted the linear mixed effects model of Square root transformed $CD4^+$ count under all possible and proposed correlation structures. So, in fitting the model by REML method of estimation, the one with the smallest AIC was considered. Accordingly, the selected correlation structure was UN.

Table 4.6: Fit Statistics for checking correlation Structure

Information Criteria	Correlation Structure	
	AR(1)	UN
-2loglikelihood	6918.3	6598.0
AIC	6924.3	6670.0
AICC	6924.3	6672.1
BIC	6934.4	6791.6

Possible two way interaction terms were considered in the refined model. From the considered interaction terms sex, age, baseline $CD4^+$ count and regimen class were found to have significant interaction with time. The insignificant variables such as, sex, age and regimen class are included in the model since their interaction terms with time are statistically significant.

The interaction term of sex with time is significant which indicates that females have a significant higher progression in square root $CD4^+$ count than males over time which coincides with the mean profile plot shown in Figure 4.8. Age was found to have a significant interaction effect with time. Parameter estimates of the interaction term of age with time is positive which indicates that the rate of progression in square root $CD4^+$ count for older patients is higher than that of younger ones. Baseline $CD4^+$ count had statistically significant interaction with time and the parameter estimate for this interaction term is negative indicating that patients that were initiated on ART at minimum $CD4^+$ counts have higher progression in square root $CD4^+$ count than those that starts at higher baselines. Among the interaction of first line ART regimen classes with time, only AZT-3TC-NVP with time was found to have a significant difference from the reference group TDF-3TC-EFV with time. Among WHO clinical stages, only stage2 was found to have a significant difference from the reference group stage4.

The final model is stated as follows.

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1 A_{ij} + \beta_2 B_i + \beta_3 WH_{0i} + \beta_4 WH_{1i} + \beta_5 WH_{2i} + (\beta_6 S_i + \beta_7 A_{ij} + \beta_8 B_i + \beta_9 R0i + \beta_{10} R_{1i} + \beta_{11} R_{2i} + \beta_{12} R_{3i} + b_{1i})\, t_{ij} + \varepsilon_{ij} \quad \ldots\ldots(4.1)$$

where, A_{ij}, B_i, WH_{0i}, WH_{1i}, WH_{2i}, S_i, A_{ij} , R_{0i}, R_{1i}, R_{2i}, $R_{3i}$are age, baseline $CD4^+$ count, WHO clinical stage1, WHO clinical stage2, WHO clinical stage3, and first line ART regimen classes AZT-3TC-EFV, AZT-3TC-NVP, TDF-3TC-NVP and d4t-3TC-NVP respectively.

Table 4.7: Parameter Estimates for full linear Mixed Model

Effect	Estimate	St.Err	DF	T.val	Pr>|t|
Intercept	8.3364	0.5502	211	15.15	<.0001
Sex	-0.6568	0.2537	211	-2.59	0.0103
AGE	-0.00810	0.009591	211	-0.85	0.3991
BASECD4	0.02666	0.000938	211	28.42	<.0001
REGIMENCLASS (AZT-3TC-EFV)	0.1626	0.2686	211	0.61	0.5456
REGIMENCLASS (AZT-3TC-NVP)	0.1890	0.2637	211	0.72	0.4744
REGIMENCLASS (TDF-3TC-NVP)	0.1326	0.3424	211	0.39	0.6990
REGIMENCLASS (d4t-3TC-NVP)	0.3678	0.5356	211	0.69	0.4930
WHOSTAGE1	0.2767	0.3154	211	0.88	0.3813
WHOSTAGE2	0.6282	0.3164	211	1.99	0.0484
WHOSTAGE3	-0.02595	0.3471	211	-0.07	0.9405
TIME	1.3320	0.2096	211	6.35	<.0001
TIME*SEX(Male)	-0.2144	0.09904	211	-2.16	0.0315
AGE*TIME	0.009795	0.004708	211	2.08	0.0387
BASECD4*TIME	-0.00248	0.000252	211	-9.83	<.0001
TIME*REGIMENCLASS (AZT-3TC-EFV)	-0.2382	0.1241	211	-1.92	0.0562
TIME*REGIMENCLASS (AZT-3TC-NVP)	-0.2697	0.1309	211	-2.06	0.0405
TIME*REGIMENCLASS (TDF-3TC-NVP)	-0.2592	0.1803	211	-1.44	0.1522
TIME*REGIMENCLASS(d4t-3TC-NVP)	0.3060	0.1994	211	1.53	0.1263

Table 4.8 shows type three tests of fixed effects. From the table, we can observe that the overall main effect of base line CD4+ count and WHO clinical stage and the overall interaction effects of sex, age, baseline CD4+ count and regimen class with time are statistically significant.

Table 4.8: Type 3 Tests for full linear Mixed Model

Effect	Num.DF	Den.DF	F Value	Pr> F
Sex	1	211	6.70	0.0103
AGE	1	211	0.71	0.3991
BASECD4	1	211	807.49	<.0001
REGIMENCLASS	4	211	0.19	0.9441
WHOSTAGE	3	211	3.28	0.0220
TIME	1	211	31.68	<.0001
TIME*SEX	1	211	4.69	0.0315
AGE*TIME	1	211	4.33	0.0387
BASECD4*TIME	1	211	96.54	<.0001
TIME*REGIMENCLASS	4	211	3.30	0.0120

4.4.3. The random effects model

The random effects model was fitted by considering the covariates selected under the variable selection step in both univariate and full random effects model. These variables were Age, Baseline $CD4^+$ count, WHO Clinical Stage, Time and the interaction terms of sex, age, base line $CD4^+$ count and first line ART regimen class with time.

4.4.3.1.Univariable Random effects Model

The set of models similar to the marginal models covered above was fitted except that in this modelling approach, we allow the intercept and slope to account for subject to subject variability through the subject specific random effects. In random intercept and slope effect models, patients vary not only in their baseline level of response, but they also vary in terms of their progression of the mean response over time.

In fitting the univariable random effects model, covariates time, age, base line $CD4^+$ count, WHO clinical stage and the interaction of age, base line $CD4^+$ count with time and only the interaction of ART regimen class AZT-3TC-EFV with time were found to have a statistically significant effect on CD4+ count progression over time at 0.25 level of significance.

4.7.3.2.The random intercepts and slope model

The random effects model or subject specific model assumes that extra correlation arises among the longitudinal response as the parameter estimates in the random effects model vary across individuals (Diggle*et al*., 2002). In the univariable random effects model, we have seen that the

variances for the random effects are statistically significantly different from zero indicating that there is between patient variations. This variation does not only exist at baseline, but it exists over the follow up time as well implying that the random intercept and slope should be included in the linear mixed model.

In the exploratory data analysis section for exploring the random effect in Figure 4.9, we have seen that both the intercept and slope terms are random. Hence, in this section we have fitted the full random effects model with both random intercept and slope. So, the model fitted is:

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1 A_{ij} + \beta_2 M_i + \beta_3 B_i + \beta_4 WH_{0i} + \beta_5 WH_{1i} + \beta_6 WH_{2i} + (\beta_7 S_i + \beta_8 A_{ij} + \beta_9 M_i + \beta_{10} + \beta_{11} R0i + \beta_{12} R_{1i} + \beta_{13} R_{2i} + \beta_{14} R_{3i} + b_{1i})\, t_{ij} + \varepsilon_{ij} \quad \ldots\ldots(4.2)$$

where the notations are all the same as that we have used in Table 4.7. This model was fitted under REML method of estimation and by the AR (1) covariance structure since the proposed covariance structures UN, CS and TOEP and any of the spatial covariance structures had a convergence problem.

Table 4.9: Fit Statistics for random intercepts and slope model

-2 Res Log Likelihood	6887.3
AIC (smaller is better)	6891.3
AICC (smaller is better)	6891.3
BIC (smaller is better)	6898.1

The covariance parameter estimates here show that there exists both within patient and between patient variations. The AIC value of this model is 6891.3 which is greater than the AIC value of the marginal model (6670.0). The significance of the covariance estimates for the random effects associated with the intercept and the linear time effect suggests that the intercepts and slopes vary across patients which is in line with the univariable random effects model.

Table 4.10: Covariance Parameter Estimates

ObsCovParm	Subject	Estimate	Std.Err	Z.Value	ProbZ
1	Variance	ID	0.4002	0.05124	7.81	<.0001
2	AR (1)	ID	1.000	0	.	.
3	Residual		9.5981	0.4226	22.71	<.0001

As depicted in Table 4.11, age had a significant interaction effect with time indicating that there is a significant difference between younger and older patients in the progression of $CD4^+$ count at a square root level over time which coincides with the results of the marginal linear mixed model in equation 4.1. The progression of square root $CD4^+$ count over time for younger patients was

high as compared to older ones taking the other variables in the model. Sex was found to have a significant interaction effect with time. The rate of increase in square root $CD4^+$ count for female patients was higher than that of male patients over time taking the other variables in the model. This coincides with the mean profile of $CD4^+$ count over time in Figure 4.7.

Table 4.11: Parameter Estimates for Random Slope and Intercept Model

Effect	Estimate	St.err	DF	T-val	Pr>|t|
Intercept	11.0034	0.9994	209	11.01	<.0001
SEX	0.08245	0.3646	831	0.23	0.8211
AGE	-0.02129	0.01728	831	-1.23	0.2182
BASECD4	0.02213	0.001409	831	15.71	<.0001
REGIMENCLASS (AZT-3TC-EFV)	0.09267	0.4773	831	0.19	0.8461
REGIMENCLASS (AZT-3TC-NVP)	0.5077	0.4733	831	1.07	0.2837
REGIMENCLASS (TDF-3TC-NVP)	-0.00326	0.5985	831	-0.01	0.9957
REGIMENCLASS(d4t-3TC-NVP)	-0.1583	0.6475	831	-0.24	0.8069
WHOSTAGE1	0.4464	0.5204	831	0.86	0.3912
WHOSTAGE2	0.9392	0.5108	831	1.84	0.0663
WHOSTAGE3	-0.2771	0.4968	831	-0.56	0.5772
TIME	1.1441	0.2282	212	5.01	<.0001
TIME*SEX(Male)	-0.2518	0.1092	831	-2.30	0.0214
AGE*TIME	0.01136	0.004678	831	2.43	0.0153
BASECD4*TIME	-0.00205	0.000306	831	-6.68	<.0001
TIME*REGIMENCLASS (AZT-3TC-EFV)	-0.2543	0.1384	831	-1.84	0.0665
TIME*REGIMENCLASS (AZT-3TC-NVP)	-0.2892	0.1430	831	-2.02	0.0434
TIME*REGIMENCLASS (TDF-3TC-NVP)	-0.2103	0.1844	831	-1.14	0.2542
TIME*REGIMENCLASS(d4t-3TC-NVP)	0.4456	0.2275	831	1.96	0.0505

The baseline $CD4^+$ count is significant both at baseline and throughout the follow up period of time. The progression of square root $CD4^+$ count is high for those patients whose baseline $CD4^+$ count at initiation of ART was minimum as compared to those have maximum baseline $CD4^+$ count. Over all WHO clinical stage had found to have statistically significant effect on $CD4^+$ count at a square root level.

Table 4.12: Type 3 Tests of Random Slope and Intercept Model

Effect	Num.DF	Den.DF	F Value	Pr> F
SEX	1	831	0.05	0.8211
AGE	1	831	1.52	0.2182
BASECD4	1	831	246.87	<.0001
REGIMENCLASS	4	831	0.56	0.6940
WHOSTAGE	3	831	2.94	0.0322
TIME	1	212	20.40	<.0001
TIME*SEX	1	831	5.31	0.0214
AGE*TIME	1	831	5.90	0.0153
BASECD4*TIME	1	831	44.60	<.0001
TIME*REGIMENCLASS	4	831	3.79	0.0046

From the random effect models, the random intercept only model was fitted lastly. The model thus fitted was,

$$Y_{ij} = \beta_0 + b_{0i} + \varepsilon_{ij} \ldots\ldots\ldots\

4.7.4. Mixture Chi-square Distribution

All observed values for -2ln (Λ_N) or a mixture of chi-square values are larger than 90, yielding p-values smaller than 0.0001(Verbeke and Molenberghs 2000). The need for random intercept was assessed using a mixture of chi-squared distributions. That is: $\chi^2_{1:2}$ = 623.1 with p < 0.0001. This test indicates that, the random intercept should be included in the model. The importance of the random slope was tested by a mixture of chi-squared distributions $\chi^2_{0:1}$ = 912.4 with a p-Value <0.0001 resulted as shown in Table 4.16 indicating that the covariance structure should not be simplified by deleting the random slopes from the model. So, including the random intercept and slope in the model fits the data well relative to the marginal one.

Table 4.16: Mixture Chi-square Distribution

Parameter	**No random effects**	**Random Intercept**	**Random Intercept andslope**
	Estimate (Std Error)	**Estimate (Std Error)**	**Estimate (Std Error)**
σ^2_{int}		15.1671 (0.6620)	
$\sigma^2_{int, Time}$			9.5981 (0.4226)
σ^2_{Time}	2.1292 (0.2114)		
-2 REML Log Likelihood	6598.0	7510.4	6887.3
*$\chi^2_{0:1}$ = 912.4 P-Value <0.0001 $\chi^2_{1:2}$ = 623.1 p < 0.0001			

4.8. Handling of Missing Data.

4.4.1. Maximum Likelihood

Direct likelihood method of handling missing data is the output of the statistical models applied to the data without further considerations about the missing data. Therefore, there are no additional computations or data manipulations about the missing data. The underlining assumption is that the data was missing at random (MAR). This was shown in the missing pattern created under the multiple imputation with null imputations. MAR assumes in the analysis that, given observed data, the missing ness mechanism does not depend on the unobserved data. The likelihood inference based on direct likelihood approach is valid under the assumption that the missing data mechanism is missing at random (MAR).

4.4.2. Multiple Imputations

The focus of this section was in performing multiple imputations to account for missing values in the $CD4^+$ count data of patients on ART. The underlining assumption for the data generating mechanism was missing at random (MAR) as shown in the missing pattern in Table A_2 on the appendix. The multiple imputation inference have been conducted in three distinct phases. Firstly, the missing data were filled in 10 times to generate ten complete data sets. Then, the ten data sets were analyzed using linear mixed models. Lastly, the results from the ten complete data sets were combined to obtain the following results. In the variance information result from MI, the between variance measures the variability in the parameter estimates obtained from each of the ten imputed data sets. This variability tells the additional variation or uncertainty that results from missing data. Detail of the variance information are shown in Table A_4 on the appendix. According to the result by multiple imputation, the parameter estimates for time, WHO stage3 and the interaction effects of sex, baseline $CD4^+$ count and regimen class AZT-3TC-NVP with time are found to have a significant effect on $CD4^+$ count progression as depicted in Table 4.17.

Table 4.17: Parameter Estimates for Multiple Imputation

	Parameter Estimates (10 Imputations)						
Parameter	**Estimate**	**Std Error**	**95% Confidence Limits**		**DF**	**t for H0: Parameter =Theta0**	**Pr > |t|**
Intercept	10.138	1.1421	7.8972	12.3798	926.46	8.88	<.0001
Age	-0.0097	0.0173	-0.0439	0.0245	3399.2	-0.56	0.5779
baseCD4	0.0215	0.0013	0.0190	0.0240	12153	16.94	<.0001
Sex (Female)	0.0180	0.3991	-0.7666	0.8025	392.93	0.05	0.9640
Regimenclass0	-0.0121	0.5155	-1.0247	1.0006	541.35	-0.02	0.9813
Regimenclass1	0.4652	0.4982	-0.5121	1.4425	1519.1	0.93	0.3506
Regimenclass2	0.2446	0.6169	-0.9686	1.4578	365.85	0.40	0.6920
Regimenclass3	-0.1108	0.8636	-1.8061	1.5844	767.67	-0.13	0.8979
WHOStage1	0.6972	0.4837	-0.2570	1.6515	187.04	1.44	0.1512
WHOStage2	0.6905	0.4849	-0.2682	1.6492	140.24	1.42	0.1567
WHOStage3	1.5163	0.4241	0.6834	2.3492	600.38	3.58	0.0004
Time	1.0739	0.2679	0.5346	1.6131	46.127	4.01	0.0002
Time*sex F	0.2229	0.1049	0.0134	0.4324	62.894	2.13	0.0375
Age*Time	0.0040	0.0050	-0.0061	0.0142	42.082	0.80	0.4299
baseCD4*Time	-0.0013	0.0003	-0.0019	-0.0008	102.53	-5.09	<.0001
Time*Regimenclass0	-0.1735	0.1261	-0.4233	0.0763	115.69	-1.38	0.1716
Time*Regimenclass1	-0.2875	0.1292	-0.5431	-0.0320	136.61	-2.23	0.0277
Time*Regimenclass2	-0.3398	0.1819	-0.7029	0.0234	67.977	-1.87	0.0662
Time*Regimenclass3	0.0423	0.2122	-0.3778	0.4624	122.78	0.20	0.8423

4.9. Model comparison

The primary objective of model comparison is to choose the simplest model that provides the best fit to the data. We have used Information criteria followed by Likelihood ratio test to select the best model from the fitted ones. The models to be compared in this study were the full marginal model, the random intercept and slope model and the random intercept only model. So, by the information criteria and likelihood ratio test from the models the full linear mixed model we have fitted in Table 4.7 is best relative to the others.

According to the AIC's of the three models, the full linear mixed model has minimum AIC (6670.0) than the model with random intercept and slope models AIC (6891.3) obtained and random intercept only model AIC (7514.4). So, the selected model best describes the square root $CD4^+$ count.

When we compares the parameter estimates of ML and MI results, we have observed it to be comparable. The variables baseline $CD4^+$ count, WHO clinical stage3, time and the interaction terms of baseline $CD4^+$ count with time were statistically significant in both models and the standard errors of these estimates showed slight deflation from the full likelihood to the result by MI.

4.10. Model Diagnostics

Residuals are frequently used to evaluate the validity of the assumption of models. In normal linear models residuals are used to verify linearity of effects, normality, independence and homoscedasticity of the errors and presence of outliers or influent observations. Since mixed models have two sources of variation (within and between), different types of residuals may be defined and the corresponding analysis is more complex., In linear mixed models, there are three types of residuals that accommodate the extra source of variability present in the model (Nobre and Singer (2007)).

i. Marginal residuals, $\hat{\varepsilon} = y - x\,\hat{\beta}$that predict the marginal errors $\varepsilon = y - IE[y] = y - X\beta$.
ii. Conditional residuals, , $\hat{\varepsilon} = y - x\,\hat{\beta} - z\,\hat{b}$, that predict the conditional errors $\varepsilon = y - E[y/b] = y - X\beta - z\,b$
iii. The best linear unbiased predictor BLUP, $z\hat{b}$, that predicts the random effects, $Zb = E[y/b] - E[y]$

These three residual plot types are useful to evaluate and test the assumptions of linear mixed models we have fitted in earlier sections. The marginal residuals are used to assess the fixed effects specification of the model and the conditional residuals are used to assess the outlier subjects and whether the random effects are proper or not. In the linear mixed model we have fitted, the assumption is that, residuals and random effects are normally distributed. Hence, the violation of this assumption does affect the parameter estimates and standard errors of the residual effects.

Firstly we have seen the plot of marginal residuals and looks it as follows. Hence the plot of marginal residuals below can be used to assess linearity of the response with respect to explanatory variables. Figure 4.11 shows the plot of marginal residuals in which a random behaviour scatter

plot around zero is a sign of linearity. There is no indication that linearity and/or within-patients square root $CD4^+$ counts covariance structure might not be appropriate for patients. The histogram plot of the residuals for the marginal model shows that the residuals are normally distributed and thus, the normality assumption is not violated under the linear mixed models as the histogram is indicative of a bell shaped distribution.

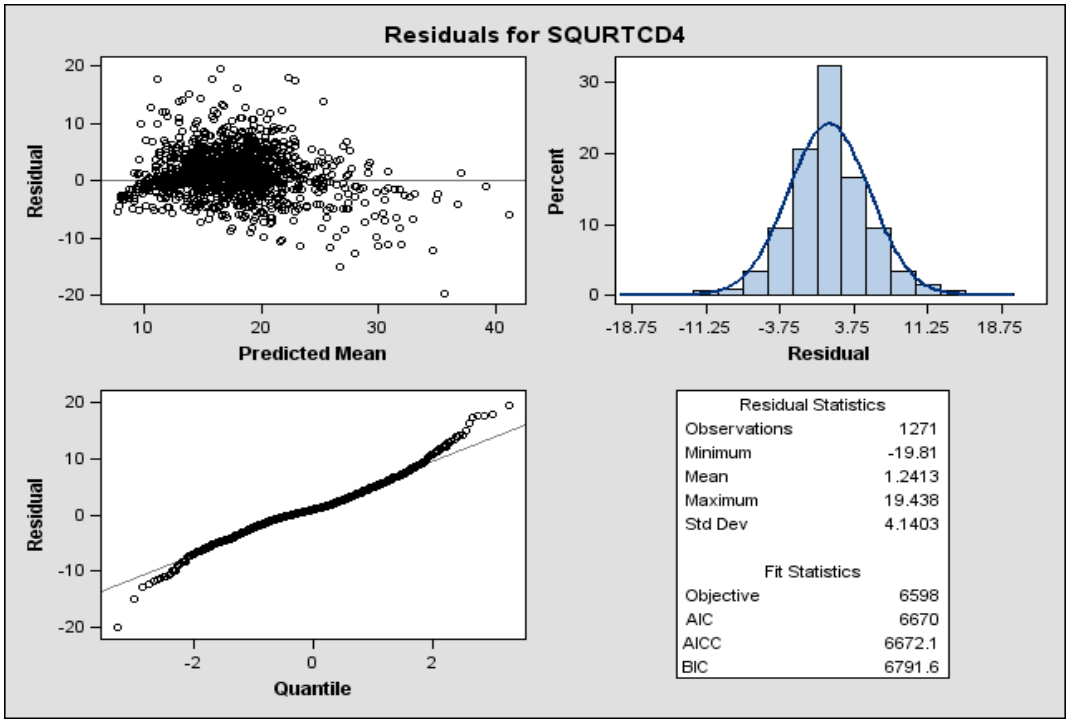


Figure 4.12 Residuals plot for square root $CD4^+$ count

Conditional residualPlots of estimates against Y in figure 4.12 was used to assess homogeneity of the variances as well as normality. These plot of standardized conditional residuals against the predicted mean shows that the plots are scattered around zero and suggests existance of some possible outlying values. But, the histogram shows normality of the conditional error terms for the $CD4^+$ count data which approaches with the quantile plot against these standardized residualsin the same figure. This indicates that normality assumption of the error terms are meet and not violated for the fitted linear mixed effects model.

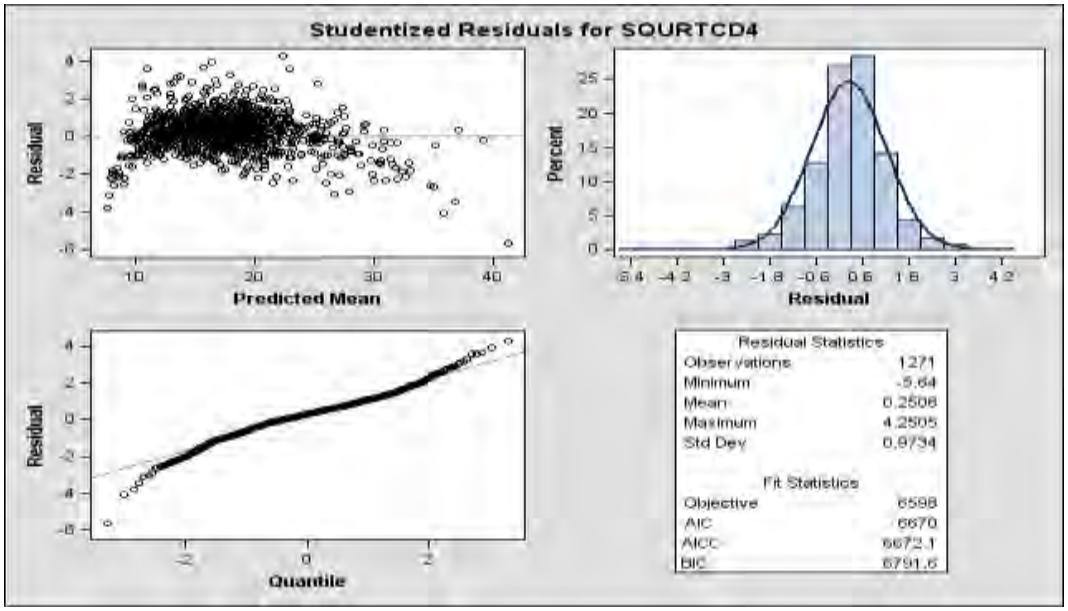


Figure 4.13 Studentized residual plots of square root CD4$^+$ count

The normality assumtion of the random intercept and slope still shows to hold implying that he inference for the random effects are not affected. This is also well shown by the histogram of pearson residuals for the square root transformed CD4$^+$ count data in graph 4.13. The quantile plot, the predicted mean against the pearson residual and the histogram plot all shows that normality assumtion of the random effects are not violated too.

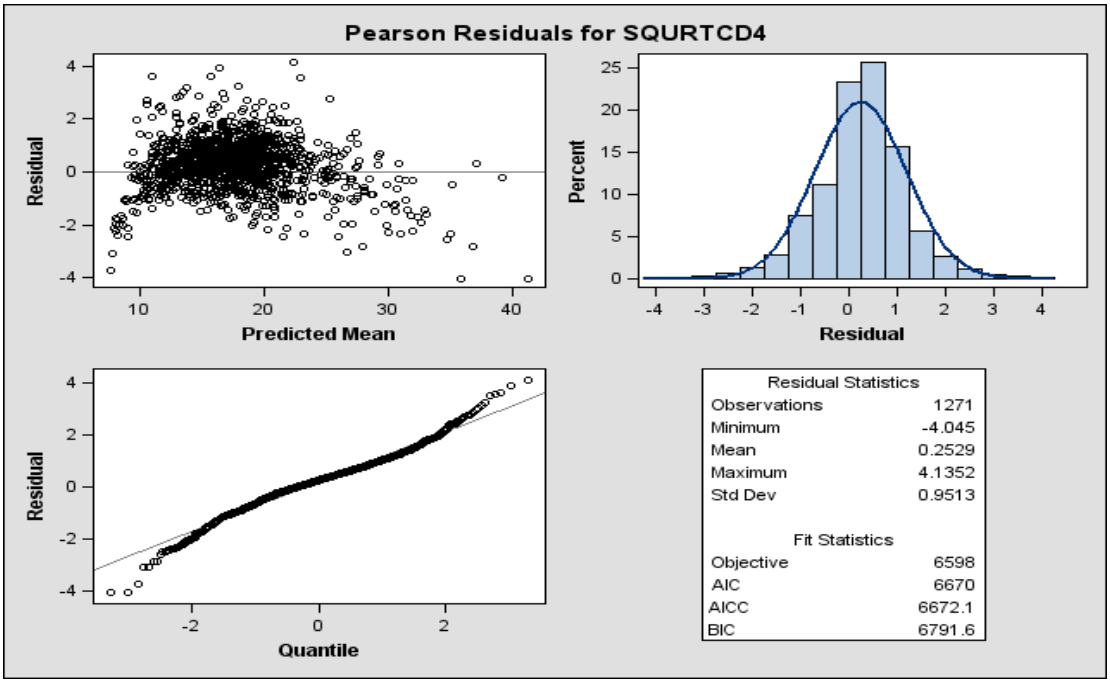


Figure 4.14 Pearson Residual Plots of Square Root CD4$^+$Count

4.11. Discussion of the results

This study attempted to examine the impact of some demographic and clinical factors for the progression of $CD4^+$ count over time in Yekatit 12 Referral Hospital. The results showed that there was a significant linear time effect on the progression of $CD4^+$ counts of patients on ART which coincides with the results of a study by Reda*et al*. 2013.

When selecting a covariance structure for a model, only structures that make sense for the data should be considered (Hallahan, 2003). In the linear mixed model, AR (1) was selected after considering all possible covariance structures due its smallest value. Then, variables selection was done using the manual backward selection criteria. It was carried out based on maximum likelihood method of estimation and using the selected covariance structure and hence, the selected variables were: time, age, baseline $CD4^+$ count, WHO clinical stage and the interaction terms of sex, age, baseline $CD4^+$ count and regimen class with time.

The linear mixed model showed that, the interaction term of sex with time was significant which indicates that females had a significantly higher progression in square root $CD4^+$ count than males. This is in contrast to a study conducted by Adams and Luguterah (2013).

In this study, the interaction term of age with time was found to have significant positive effect implying that, the rate of progression in square root $CD4^+$ count for older patients was higher than that of younger ones which was also in contrast with a study by Picat*et al.* 2013.

Baseline $CD4^+$ count had statistically significant interaction with time and the parameter estimate for this interaction term is negative indicating that patients that have initiated ART at minimum $CD4^+$ counts have higher progression in square root $CD4^+$ count than those that start at higher baselines.

Our study revealed that baseline CD4+ count was significant predictor of $CD4^+$ count progression which is in line with a retrospective cross-sectional study on $CD4^+$cell count trends after commencement of ART among HIV infected patients in Tigray, Northern Ethiopia (Addisu*et al*., 2015).

The study found the interaction of female sex group with time to have statistically significant difference from the reference group male with time for the progression of $CD4^+$ counts over time. This implies that the progression of CD4+ counts for female patients are more likely higher than their female counterparts over time.

Among the interaction of first line ART regimen classes with time, only AZT-3TC-NVP with time was found to have a significant difference from the reference group TDF-3TC-EFV with time. Among WHO clinical stages, only stage 3 was found to have a significant difference from the reference group stage 4. This implies that the progression of $CD4^+$ counts of patients started ART at stage three was more likely higher than those started ART at stage four.

In a retrospective longitudinal study on predictors of change in CD4 Lymphocyte count and weight among HIV infected patients on ART in Ethiopia by Ayalu *et al*., (2013), functional status had a significant effect on $CD4^+$ lymphocyte count. In our study functional status was not significantly associated with the progression of $CD4^+$ counts over time.

Lastly, we conducted Multiple Imputation (MI) and found that the results of the full likelihood and MI approaches were comparable for covariates that are significant. This coincides with the simulation study result reported in Verbeke and Molenberghs, 2001 and Kenward and Molenberghs, 2007.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.3.Conclusion

$CD4^+$ counts are the primary laboratory markers used to track the progression of HIV to AIDS. It is the strongest predictor for risk of death and AIDS (DHHS, 2011) and used to decide when to start antiretroviral therapy (ART).The objective of this study was to assess the progression of $CD4^+$ counts of patients and tried to identify determinant factors for the change in $CD4^+$ counts of patients initiated on ART. This study therefore tried to see the progression of $CD4^+$ count and identified the factors affecting $CD4^+$ count of patients using linear mixed models.

The results of this study showed that age, baseline CD4+ count, WHO clinical stage, time and the interaction effects of age, sex, baseline CD4+ count and regimen class with time are significantly associated with the progression of CD4+ counts over time.

5.4.Recommendation

This study on longitudinal modelling $CD4^+$ counts of HIV positive patients has used $CD4^+$ count data of patients on ART in Yekatit 12 Referral Hospital. From the results of this study, the following recommendations are forwarded for clinicians on HIV/AIDS and patients as well.

- ✓ Comparing the progression of CD4+ counts, males were found to have the worst $CD4^+$ count progression compared to females. Hence, actions are needed to increase male patients $CD4^+$ count progression.
- ✓ HIV positive patients at younger age have to start treatment early to upgrade their $CD4^+$ count progress over time.
- ✓ Based on the findings of our study and WHO recommendation, patients should initiate ART regimens at an early WHO clinical stage.
- ✓ ART regimens should be taken continuously to increase the progress of $CD4^+$ count over time.
- ✓ Other factors that were not considered in this study might have association with the progression of $CD4^+$ counts over time. So, further studies in the area should consider possible covariates such as educational status, ART second line regimens and other socioeconomic factors.

References

1. AddisuAsfaw, Dagim Ali, TadeleEticha, AdissuAlemayehu, MussieAlemayehu, FilmonKindeya. CD4 Cell Count Trends after Commencement of Antiretroviral Therapy among HIV Infected Patients in Tigray, Northern Ethiopia: A Retrospective Cross-Sectional Study, *PLOS ONE*, 2015; 1 DOI:10.1371/journal.pone.0122583.
2. AgrestiA . (1996): *An introduction to Categorical Data Analysis*. Wiley Series in Probability and Statistics.
3. AkinsegunAkinbami, AdedoyinDosunmu, etal. "CD4 Count Pattern and Demographic Distribution of Treatment-Na¨ıve HIV Patients in Lagos, Nigeria". *Hindawi Publishing Corporation,* 2012, Article ID 352753 doi:10.1155/2012/352753
4. Allison A. (2006). *HIV Laboratory Monitoring*, www.aidsalliance.org/publications.
5. Bennell P, Hyde K, Swainson N. The impact of HIV/AIDS in sub-Saharan Africa: a synthesis of a three-country study .Brighton, Centre for International Education 2002.
6. C. Bijleveld, L. van der Kamp, A. Mooijaart, W. van der Kloot, R. van der Leeden, *Longitudinal Data Analysis*: Designs, Models & Methods, Sage publications, 1999.
7. Cross C. Sinking deeper down: HIV/AIDS as an economic shock to rural Households. Society in transition. 2001; 32(1).
8. Der, G. and Everitt, B. S. (2006). *Statistical Analysis of Medical Data using SAS*: Taylor & Francis Group, LLC.
9. Duncan, S., Ruy, M. and Martin A.(1999).Virus phenotype switching and disease progression in HIV-1 infection. *Royal society* vol.266, pp.2523-2530

   Dinberu S., *et al*2016, Predicting AIDS disease progression using longitudinal CD4 count among adult HIV/AIDS patients in Southwest Ethiopia: Application of semi markov process, International *Journal of Computational Bioinformatics and In Silico Modeling*, Vol. 5, 808-814.
10. Ethiopian Public Health Institute (EPHI), Federal Ministry of Health. HIV Related Estimates and Projections for Ethiopia July, 2014, Addis Ababa, Ethiopia.
11. FDRE, country progress report on the HIV response, 2014
12. Gazzard BG on behalf of the BHIVA Treatment Guidelines Writing Group. British HIV Association guidelines for the treatment of HIV-1-infected adults with antiretroviral therapy

(2008). Available at http://www.bhiva.org/files/file1030835.pdf. [Accessed 26st November 2015] [Context Link]

13. Getachew T, 2016. Statistical Analysis of CD4+ Cell Counts progression of HIV-1-positive Patients enrolled in Antiretroviral Therapy at Hossana District Queen ElleniMohamad Memorial Hospital, South Ethiopia, *Biometrics & Biostatistics International Journal*, volume 3, 2 – 10.
14. Habtamu, A. and Eshetu, W. (2012). Factors affecting the survival of HIV-infected children after ART initiation in Bahir Dar, Ethiopia. *Ethiop. J. Health Dev.* 2012;26 (3) 193-199.
15. Hallahan, C. (2003). Longitudinal Data Analysis with Discrete and Continuous responses using Proc Mixed. Maintained at: http://www.cpcug.org/user/sigstat/PowerPointSlides.
16. Hedeker, D. (2004). An introduction to growth modeling. In D. Kaplan (Ed.), Quantitative Methodology for the Social Sciences: Thousand Oaks CA: Sage Publications.
17. Holmberg SD, Palella FJ, Lichtenstein KA, Havlir DV. The case for earlier treatment of HIV infection. Clin Infect Dis 2004; 39:1699-1704. Phillips AN, Gazzard BG, Clumeck N, *et al.* When should antiretroviral therapy for HIV be started? *Br Med J* 2007; 334:76-78.
18. Jayasuriya A, Robertson C, Allan PS. Twenty-five years of HIV management. *J Roy Soc* Med 2007; 100:363-366
19. Johansson et al. further benefits by early start of HIV treatment in low income countries Survival estimates of early versus deferred antiretroviral therapy. January 2010, 7:3
20. J. W. Mellors, A. Munoz, J. V. Giorgi et al., "Plasma viral load and CD4+ lymphocytes as prognostic markers of HIV-1 infection," *Annals of Internal Medicine*, vol. 126, no. 12, pp. 946– 954, 1997.
21. Laird, N. and Ware, J. (1982). *Random-effects models for longitudinal data*. Biometrika 38, 963 -974
22. Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D. and Schabenberger, O. (2006). *SAS for Mixed Models*: Second Edition. Cary: SAS Institute Inc., Cary, NC, USA.
23. Lorna Renner, Meghan Prin, Fang-Yong Li, BamenlaGoka, Veronika Northrup, and Elijah Paintsil, Time to and Predictors of CD4+ T-Lymphocytes Recovery in HIV-Infected Children Initiating HAART in Ghana. 2011; Hindawi Publishing Corporation AIDS Research and Treatment, Article ID 896040, 9, doi:10.1155/2011/896040.

24. Malone, J. L., Simms, T. E., Gray, G. C., Wagner, K. F., Burge, J. R. and Burke, D. S. (1990). Sources of Variability in Repeated T-Helper Lymphocyte Counts from Human Immunodeficiency Virus Type 1-Infected Patients - Total Lymphocyte Count Fluctuations and Diurnal Cycle Are Important. Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology 3, 144-151.
25. Marlon Bwalya, factors affecting cd4+t-lymphocyte count response to HAART in HIV/AIDS patients within 24 months of treatment at Chreso ministries art centers'. 2010
26. McCulloch, C. E., Searle, S. R. and Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models*: John Wiley & Sons, Inc.
27. M. Egger, M. May, G. Chˆene et al., "Prognosis of HIV-1- infected patients starting highly active antiretroviral therapy: a collaborative analysis of prospective studies," *The Lancet*, vol. 360, no. 9327, pp. 119–129, 2002
28. Mocroft A, Ledergerber B, Katlama C, *et al*. Decline in the AIDS and death rates in the EuroSIDA study: an observational study. Lancet 2003; 362:22-29.
29. Molenbergs, G. and Verbke, G. (2005), *Models for Discrete Longitudinal Data*. New York, Springer-Verlag .Pp 265-280
30. Obeagu Emmanuel Ifeanyi, OkoroiwuIjeoma Leticia *et al*. CD4 Count of Art and Non-Art Hiv Positive Patients Attending Living Word Mission Hospital Abayi Aba, Abia State, Nigeria**:** *Journal of Dental and Medical Sciences*, *2014, 7, PP 41-46*
31. Palella FJ, Baker RK, Moorman AC, *et al*. Mortality in the highly active antiretroviral therapy era. Changing causes of death and disease in the HIV Outpatient Study. J Acquir Immune DeficSyndr 2006; 43:27-34
32. Picat, M. Q., Lewis, J., Musiime, V., Prendergast, A., Nathoo, K., Kekitiinwa, A., Ntege, P. N., Gibb, D. M., Thiebaut, R., Walker, A. S., Klein, N., Callard, R., the ARROW Trial Team. (2013). Predicting Patterns of Long-Term CD4 Reconstitution in HIV-Infected Children Starting Antiretroviral Therapy in Sub-Saharan Africa: A Cohort-Based Modelling Study. *PLoS Med,* **10(10)**: e1001542.
33. Reda, A. A., Biadgilign, S., Deribew, A., Gebre, B., Deribe, K. (2013). Predictors of Change in CD4 Lymphocyte Count and Weight among HIV Infected Patients on Anti-Retroviral Treatment in Ethiopia: A Retrospective Longitudinal Study. *PLoS ONE*, **8(4):** e58595.

34. Smith *et al,* 2004. Factors Influencing Increases in CD4 Cell Counts of HIV-Positive Persons Receiving Long-Term Highly Active Antiretroviral Therapy. *Journal of Infectious Diseases*, 2004; 8: 190:1860,
35. Stohr w., *et al* (2007). HIV medicine journal. Volume 8, *Blackwell publishing*, uk.pp135-141(7).
36. UNAIDS. How AIDS Changed Everything; 2015.
37. *U.S DHHS,* Panel on Antiretroviral Guidelines for Adults and Adolescents. Guidelines for the Use of Antiretroviral Agents in HIV-1-Infected Adults and Adolescents, Department of Health and Human Services, 2011, http://www.aidsinfo.nih.gov/.
38. Verbeke, G. &Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York, Springer-Verlag. No.2.
39. Wellington Murenjekwa**,** Comparison of cd4+ t-cell changes in response to highly active antiviral therapy (HAART) in adolescents and children enrolled at parirenyatwa hospital family care centre (2005-2010) - secondary data analysis, 2013
40. WHO/UNAIDS. Progress on global access to HIV antiretroviral therapy: a report on "3 by 5" and beyond. Geneva, World Health Organization/United Nations Joint Programme on AIDS, 2006.
41. World Health Organization, Antiretroviral Therapy for HIV infection in adults and adolescent Recommendation for public health approach, revision, 2013
42. World Health Organization. Technical Brief in CD4Technologies, 2012 [cited, Available from, http.who.int/hiv/topics/treatment/tech_brief_20100501_en pdf.
43. Allison, P. D. (2001). *Missing Data*: Quantitative Applications in the Social Sciences. Thousand Oaks, CA; Sage publications, Inc.
44. Schafer, J. L., Graham, J. W. (2002). *Missing Data*: Our View of the State of Art. Psychological Methods, 7, 147-177.
45. Y. C. Yuan, *Multiple Imputation for Missing Data*: Concepts and New Development, Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference, 267-25, SAS Institute, Cary NC, 2000
46. Rubin, D.B. (1977). Formalizing subjective notions about the effect of non-respondents in sample surveys. *Journal of the American Statistical Association*, 72, 359, 538-543.
47. UNAIDS. Fact sheet 2014: Global statistics, 2014.

48. Yarchoan R, Venzon DJ, Pluda JM, et al: CD4 count and the risk for death in patients infected with HIV receiving antiretroviral therapy. *Ann Intern Med* 1991; 115: 184-9.
49. World Bank. Intensifying action against HIV/AIDS in Africa: Responding to a Development Crisis, 1999.

Appendix A

Table A_1: Explanatory Variables with their Codlings

No	Variable	Codlings (if any)	Category (if any)
1.	Sex	0	Male
		1	Female
2.	Age	--	--
3.	Marital Status	0	Married
		1	Single (Includes Unmarried, Divorced and Widowed
4.	Basline CD4+ count	---	---
5.	Weight in kgs	---	---
6.	ART Regimen Class (1st line regimens)	0	AZT-3TC-EFV
		1	AZT-3TC-NVP
		2	TDF-3TC-NVP
		3	d4t-3tc-NVP
		4	TDF-3TC-EFV
7.	Functional Status	0	Working
		1	Bed redden
		2	Ambulatory
8.	WHO Clinical Stage	0	Stage1
		1	Stage2
		2	Stage3
		3	Stage4
9.	Time in Six Months	---	--

Table A_2: Descriptive Statistics of CD4+ count by Categorical Covariates over Time

Time		Baseline	1	2	3	4	5	6	7
Male	N	76	67	69	58	59	52	40	33
	Mean (Std.dev)	235.4474 (227.0328)	347.358 (245.0609)	365.0289 (251.2071)	378.2414 (232.3719)	390.5085 (216.3424)	380.3462 (179.6460)	395.0250 (155.0554)	365.7879 (135.7143)
Femle	N	141	112	122	104	109	89	76	64
	Mean (Std.dev)	280.7518 (200.7598)	353.7411 (188.1414)	404.3607 (219.6902)	428.1346 (233.9253)	441.3028 (204.3312)	453.7079 (202.9959)	474.2237 (228.7666)	518.7656 (255.7438)
Married	N	148	120	129	109	116	98	86	74
	Mean (Std.dev)	245.0878 (194.0114)	333.8833 (184.9362)	376.6357 (217.3899)	422.9908 (251.7142)	426.7155 (209.0640)	441.2755 (210.7395)	447.6977 (219.9233)	456.7703 (247.4627)
Single	N	69	59	62	53	52	43	30	23
	Mean (Std.dev)	307.3478 (239.3072)	386.8814 (253.0284)	418.2742 (258.5447)	384.1132 (191.6521)	416.2115 (212.0014)	393.3256 (159.7775)	444.6667 (177.7013)	498.7391 (180.6771)
AZT-3TC-EFV	N	76	63	70	60	62	50	39	26
	Mean (Std.dev)	229.6711 (180.3601)	329.5238 (195.0895)	347.3143 (223.9545)	359.1667 (217.4412)	385.3709 (209.3948)	392.4000 (207.7893)	433.3077 (78.9422)	467.3462 (205.3337)
AZT-3TC-NVP	N	56	49	51	43	49	36	29	27
	Mean (Std.dev)	307.0714 (232.3106)	411.6939 (256.2129)	424.2549 (232.6862)	449.4419 (247.6563)	438.5714 (176.7906)	413.6944 (179.6135)	399.379 (155.2186)	459.2222 (207.5217)
TDF-3TC-NVP	N	25	19	19	17	18	16	15	16
	Mean (Std.dev)	364.520 (265.6520)	352.2105 (143.9539)	528.5263 (287.2638)	429.1765 (194.3480)	436.0000 (190.9169)	514.0000 (238.5112)	495.2667 (214.1053)	454.2500 (255.6739)
d4t-3TC-NVP	N	11	10	7	4	4	6	3	3
	Mean (Std.dev)	288.9091 (181.7424)	367.000 (202.8754)	372.8571 (218.8458)	562.7500 (255.1214)	623.5000 (294.2771)	461.1668 (254.3127)	426.000 (163.6979)	481.6667 (135.3563)
TDF-3TC-EFV	N	49	38	44	38	35	33	30	25
	Mean (Std.dev)	215.0612 (183.0140)	305.1842 (190.9019)	361.7727 (197.0439)	422.1316 (250.1081)	440.4857 (242.1756)	444.0606 (160.4106)	488.4667 (280.3582)	480.3600 (290.0557)
Working	N	189	153	163	145	150	121	103	85
	Mean (Std.dev)	275.4815 (216.2098)	358.0458 (210.9718)	396.2269 (233.7808)	408.5862 (232.4841)	423.6467 (213.5437)	428.3636 (198.1053)	455.1068 (216.3914)	471.4941 (242.2546)
Bed redden	N	17	15	17	10	11	12	8	6
	Mean (Std.dev)	197.2353 (181.9687)	349.9333 (224.5336)	353.4118 (249.5013)	441.8000 (253.4223)	364.3636 (113.1621)	408.3333 (157.2614)	409.1250 (116.0953)	469.0000 (139.6825)
Ambulatory	N	11	11	11	7	7	8	5	6
	Mean (Std.dev)	187.3636 (112.4627)	260.1818 (180.2508)	356.9091 (175.5696)	400.1429 (269.1520)	512.4286 (225.8730)	428.2500 (257.1485)	338.6000 (141.0401)	396.8333 (179.3749)

Stage1	N	66	57	58	47	49	43	34	27
	Mean (Std.dev)	378.2273 (272.5665)	431.5439 (252.6408)	487.6724 (285.1185)	501.1489 (293.3276)	448.1837 (196.7091)	472.2093 (210.9839)	499.7353 (281.1449)	439.5926 (211.6233)
Stage2	N	58	44	48	48	49	40	36	35
	Mean (Std.dev)	233.4310 (154.7803)	369.6364 (144.6783)	371.0208 (173.7646)	419.3750 (188.4632)	460.4082 (237.0825)	469.4500 (200.0213)	449.6111 (175.7360)	519.4000 (261.4163)
Stage3	N	66	52	62	47	48	43	39	26
	Mean (Std.dev)	193.6061 (136.9980)	271.3269 (183.7103)	307.7097 (174.1872)	317.3404 (164.5989)	368.0625 (178.3526)	333.1628 (148.7623)	398.8205 (155.4943)	431.8461 (221.0313)
Stage4	N	27	26	23	20	22	15	7	9
	Mean (Std.dev)	229.6296 (188.2241)	304.6538 (192.6472)	406.3913 (242.9930)	393.2500 (243.8092)	407.0000 (220.6095)	449.9333 (196.9291)	444.4286 (197.3236)	444.0000 (211.2155)

Table A_3: Missing Data Pattern for Multiple Imputation

Missing Data Patterns

Group	SEX	AGE	TIME	BASECD4	REGIMENCLASS	WHOSTAGE	SQURTCD4	Freq	Percent
1	XXXXXXX			1271	73.21				
2	XXXXXX		.		465	26.79			

Missing Data Patterns

Group	SEX	AGE	TIME	BASECD4	REGIMENCLASS	WHOSTAGE
	Group Means					
1	0.642801	39.505114	2.988985	255.140834	1.527931	1.236035
2	0.668817	38.027957	4.896774	288.541935	1.479570	1.283871

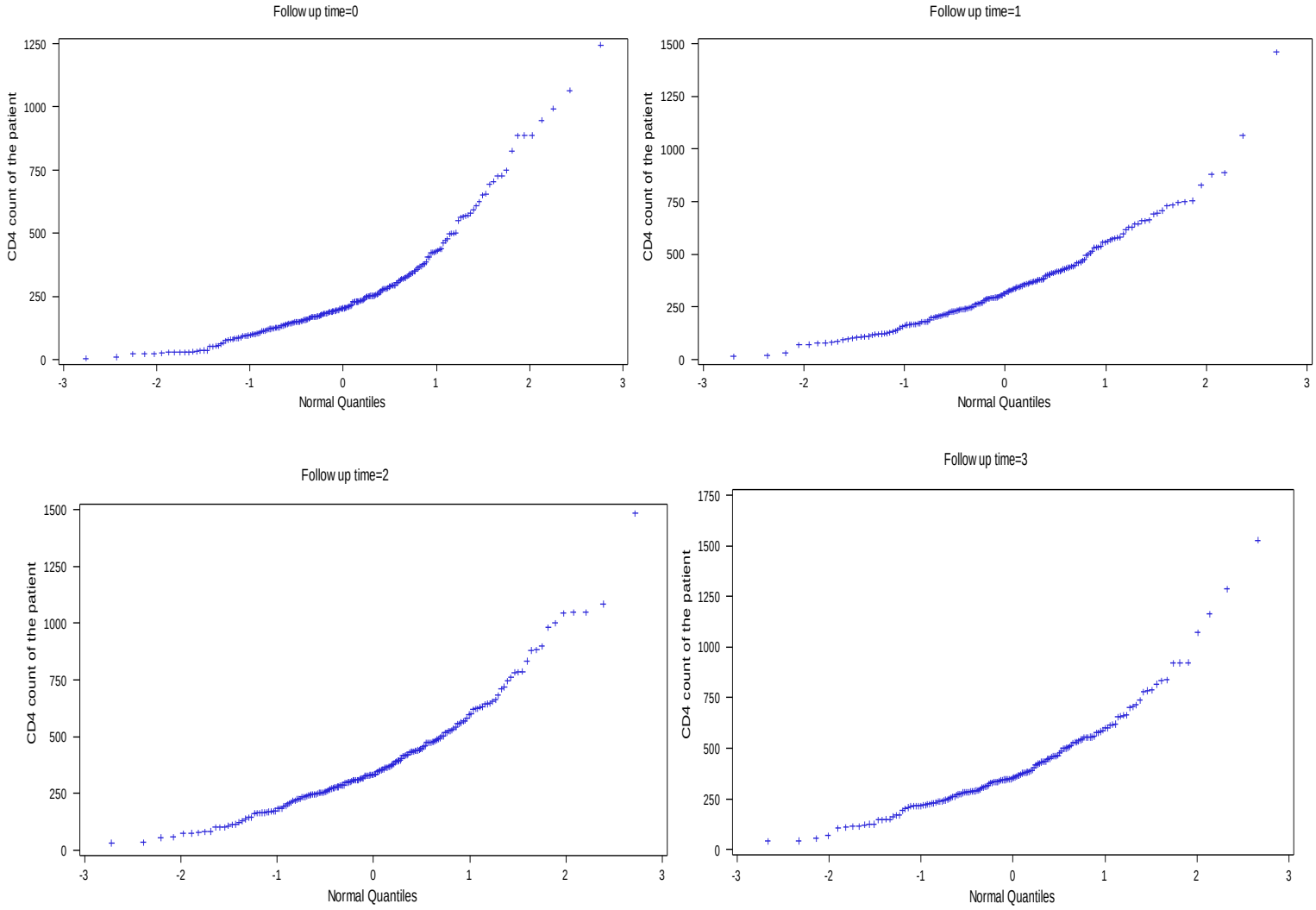
Missing Data Patterns

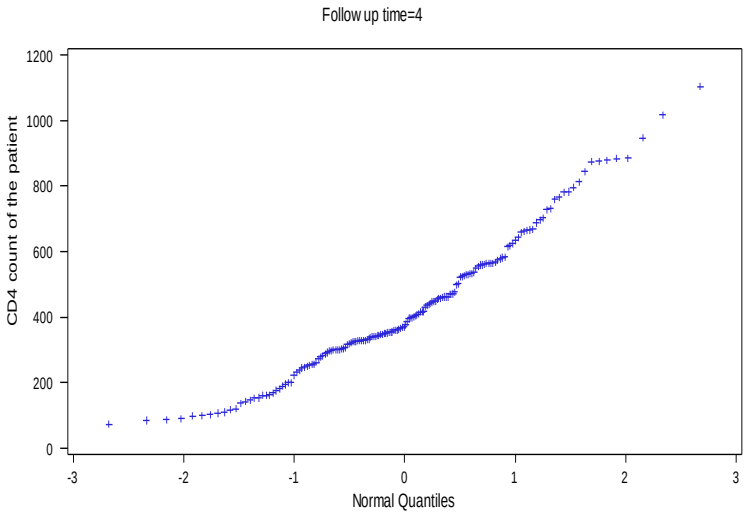
Group	SQURTCD4
	Group Means
1	18.789890
2	.

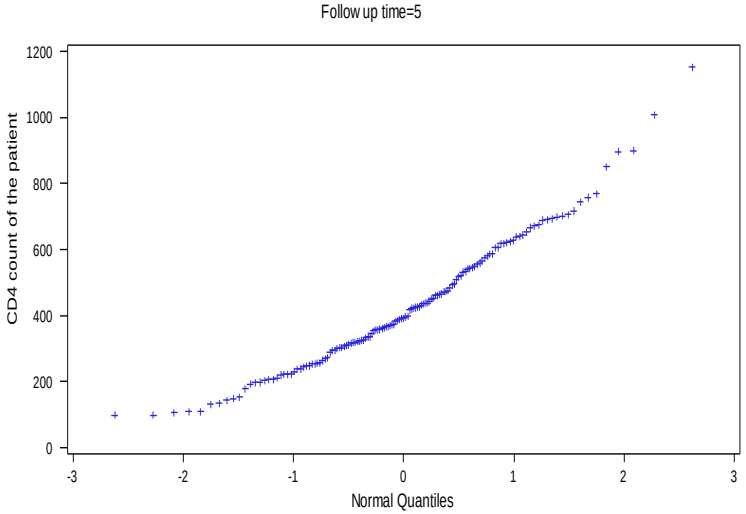
Table A_4: Variance Information for Multiple Imputations

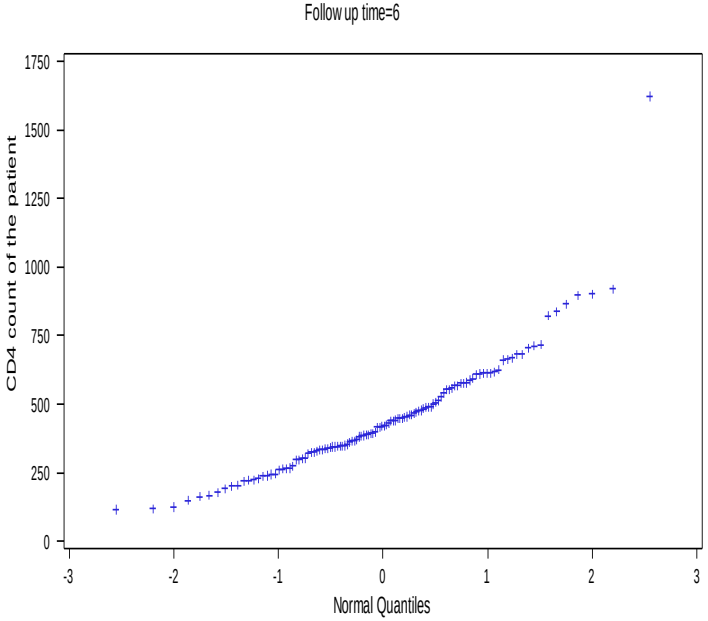
Variance Information (10 Imputations)							
Parameter	**Variance**			**DF**	**Relative Increase in Variance**	**Fraction Missing Inform**	**Rel. Efficiency**
	Between	**Within**	**Total**				
Intercept	0.1169	1.1757	1.3043	926.46	0.1093	0.1005	0.9901
Age	0.00001	0.0003	0.0003	3399.2	0.0542	0.0520	0.9948
baseCD4	3.98E-8	0.0000016	0.000002	12153	0.0279	0.0274	0.9973
sex	0.0219	0.1352	0.1593	392.93	0.1783	0.1556	0.9847
Regimenclass0	0.0312	0.2315	0.2658	541.35	0.1480	0.1321	0.9869
Regimenclass1	0.0174	0.2291	0.2482	1519.1	0.0834	0.0782	0.9922
Regimenclass2	0.0543	0.3209	0.3806	365.85	0.1860	0.1614	0.9841
Regimenclass3	0.0734	0.6650	0.7458	767.67	0.1214	0.1106	0.9891
WHOStage1	0.0467	0.1827	0.2339	187.04	0.2810	0.2276	0.9777
WHOStage2	0.0542	0.1756	0.2352	140.24	0.3393	0.2638	0.9743
WHOStage3	0.0201	0.1578	0.1799	600.38	0.1395	0.1253	0.9876
Time	0.0288	0.0401	0.0718	46.127	0.7912	0.4644	0.9556
Time*sexF	0.0038	0.0068	0.0109	62.894	0.6084	0.3972	0.9618
Age*Time	0.00002	0.000014	0.00003	42.082	0.8603	0.4863	0.9536
baseCD4*Time	1.83E-8	4.797E-8	6.817E-8	102.53	0.4209	0.3096	0.9699
Time*Regimenclass0	0.0040	0.0115	0.0159	115.69	0.3868	0.2910	0.9717
Time*Regimenclass1	0.0039	0.0124	0.0167	136.61	0.3453	0.2673	0.9739
Time*Regimenclass2	0.0109	0.0211	0.0331	67.977	0.5719	0.3818	0.9632
Time*Regimenclass3	0.0111	0.0328	0.0450	122.78	0.3713	0.2823	0.9725

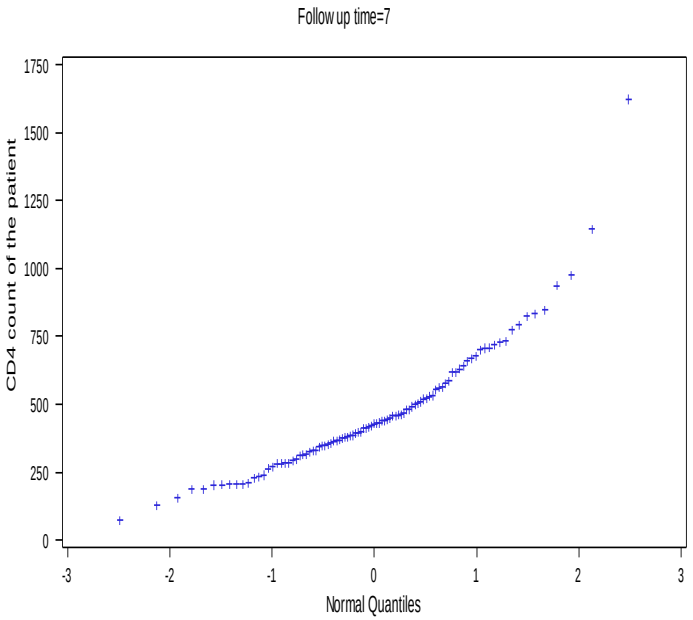
Figure A_1 : Normal Q-Q plots of $CD4^+$ counts by Time

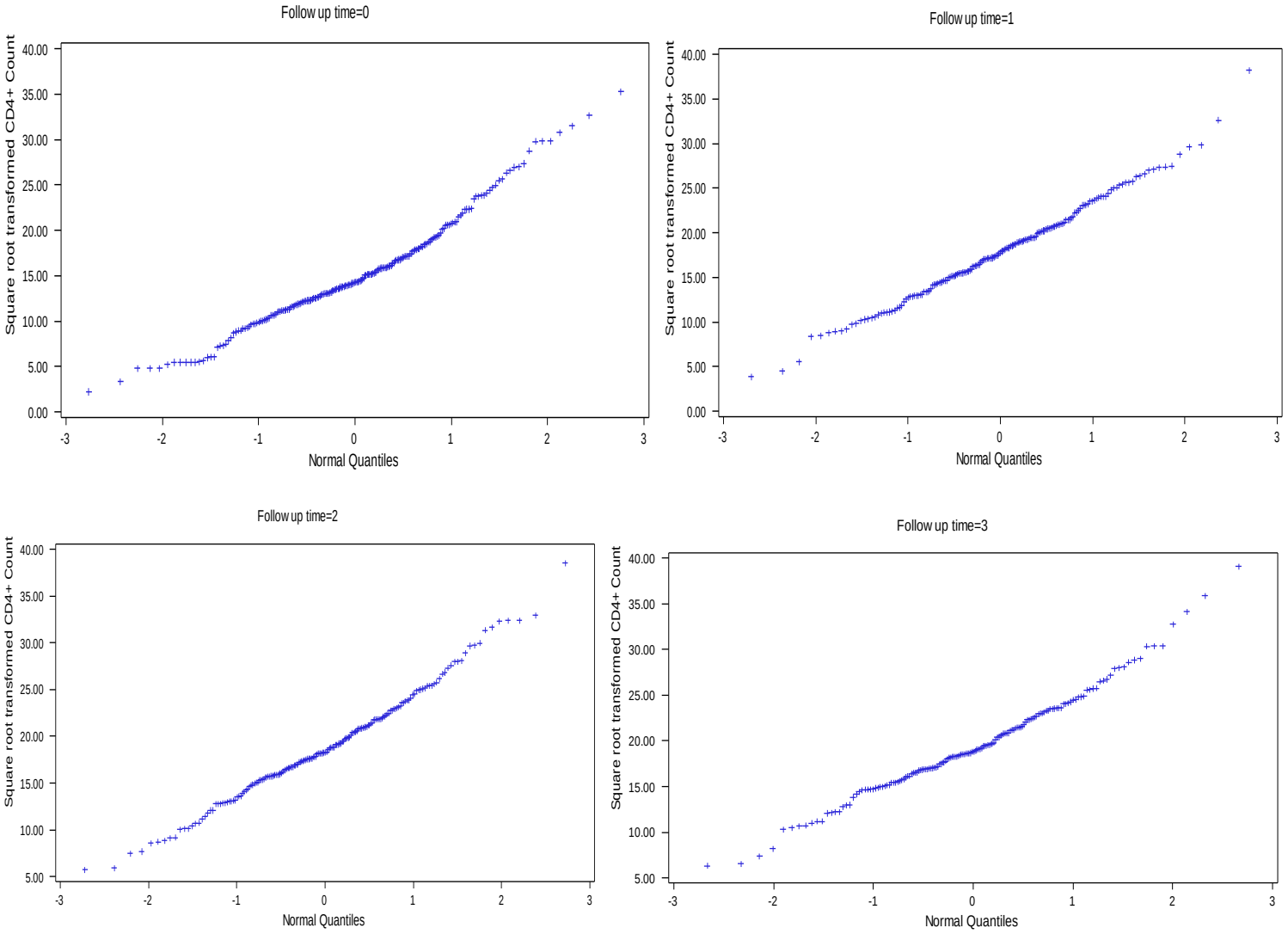


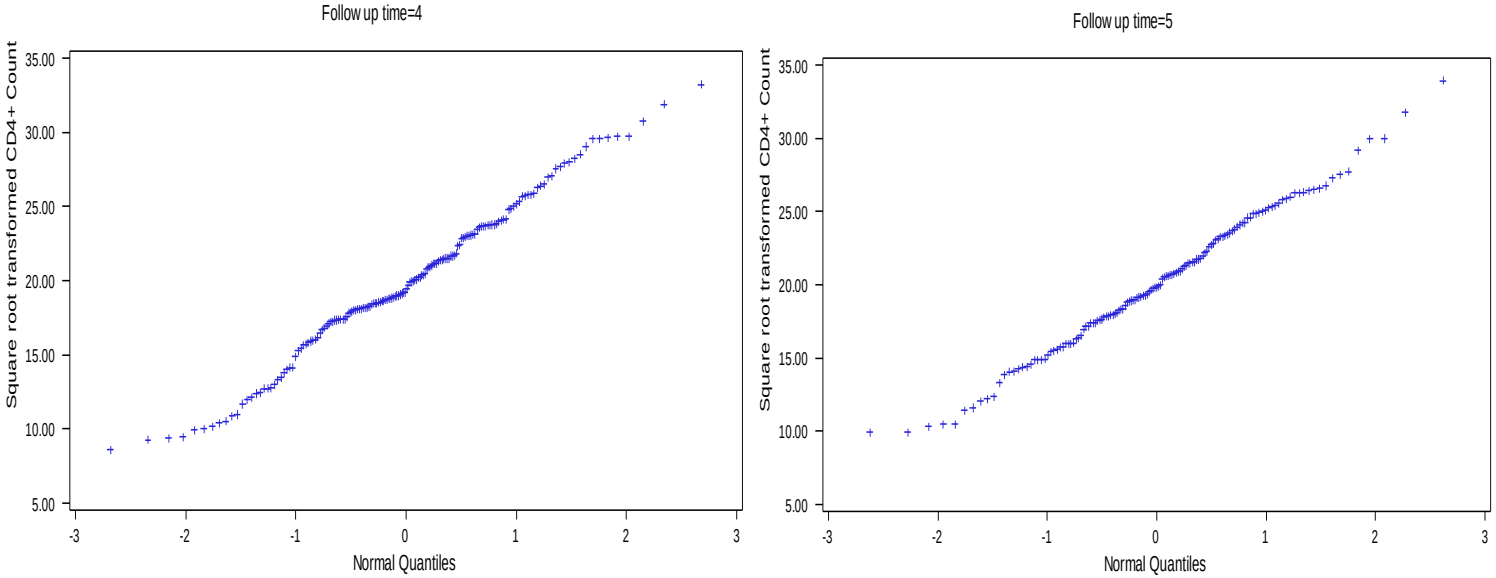

Follow up time=4
CD4 count of the patient
Normal Quantiles
0
200
400
600
800
1000
1200
-3
-2
-1
0
1
2
3


Follow up time=5
CD4 count of the patient
Normal Quantiles
0
200
400
600
800
1000
1200
-3
-2
-1
0
1
2
3

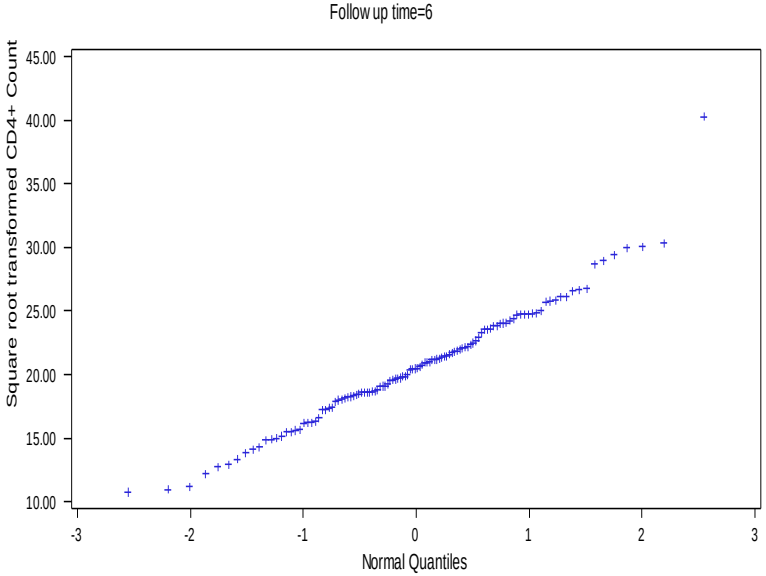

Follow up time=6
CD4 count of the patient
Normal Quantiles
0
250
500
750
1000
1250
1500
1750
-3
-2
-1
0
1
2
3

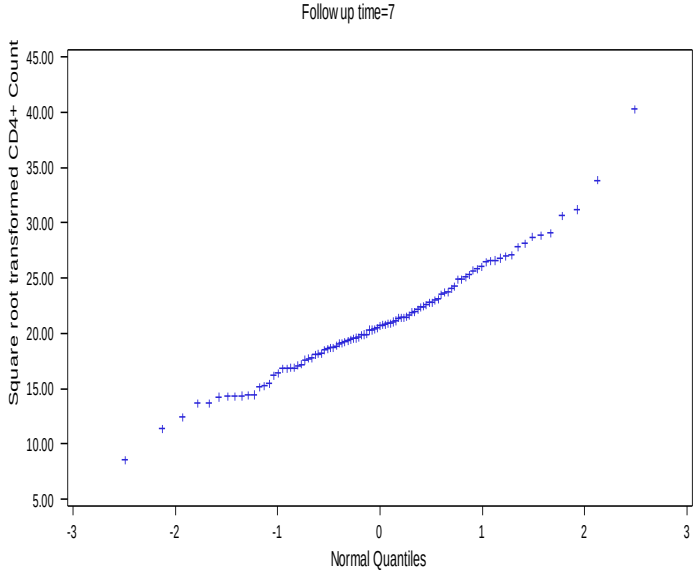

Follow up time=7
CD4 count of the patient
Normal Quantiles
0
250
500
750
1000
1250
1500
1750
-3
-2
-1
0
1
2
3

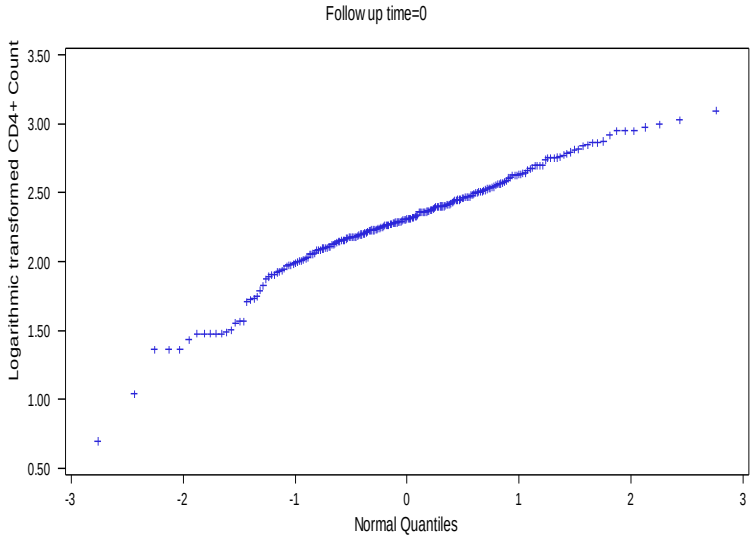

Follow up time=0
Square root transformed CD4+ Count
Normal Quantiles
Follow up time=1
Square root transformed CD4+ Count
Normal Quantiles
Follow up time=2
Square root transformed CD4+ Count
Normal Quantiles
Follow up time=3
Square root transformed CD4+ Count
Normal Quantiles

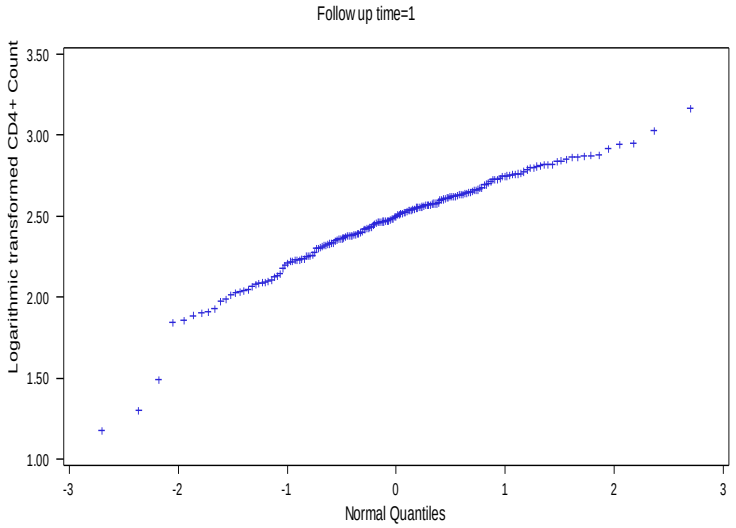

Follow up time=4
Square root transformed CD4+ Count
35.00
30.00
25.00
20.00
15.00
10.00
5.00
-3
-2
-1
0
1
2
3
Normal Quantiles

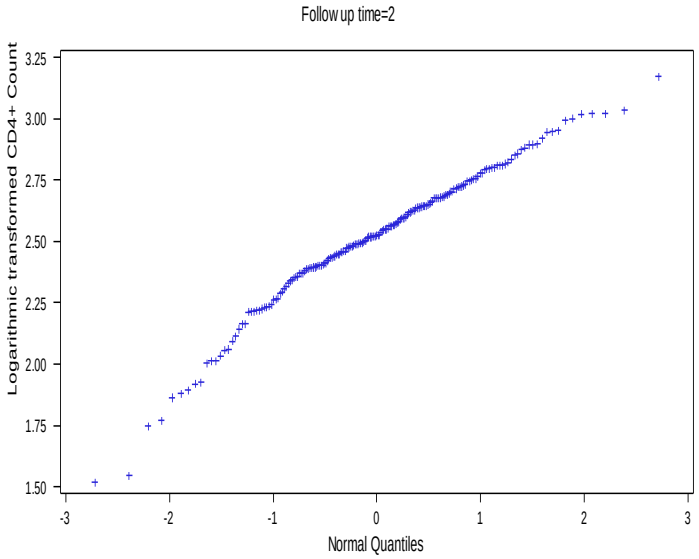
Follow up time=5
Square root transformed CD4+ Count
35.00
30.00
25.00
20.00
15.00
10.00
5.00
-3
-2
-1
0
1
2
3
Normal Quantiles

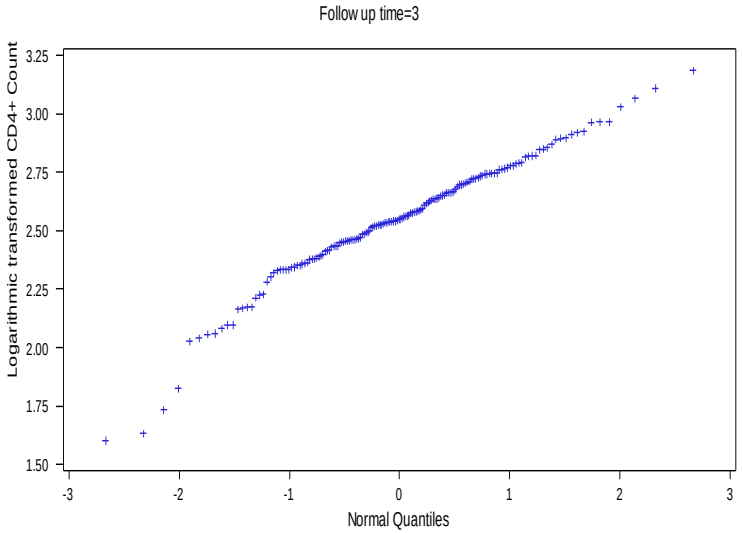

Follow up time=6
Square root transformed CD4+ Count
45.00
40.00
35.00
30.00
25.00
20.00
15.00
10.00
-3
-2
-1
0
1
2
3
Normal Quantiles

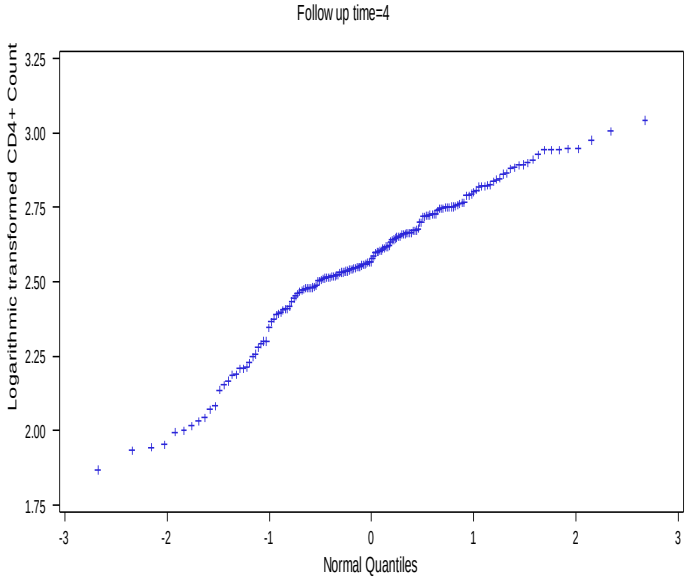

Follow up time=7
Square root transformed CD4+ Count
45.00
40.00
35.00
30.00
25.00
20.00
15.00
10.00
5.00
-3
-2
-1
0
1
2
3
Normal Quantiles

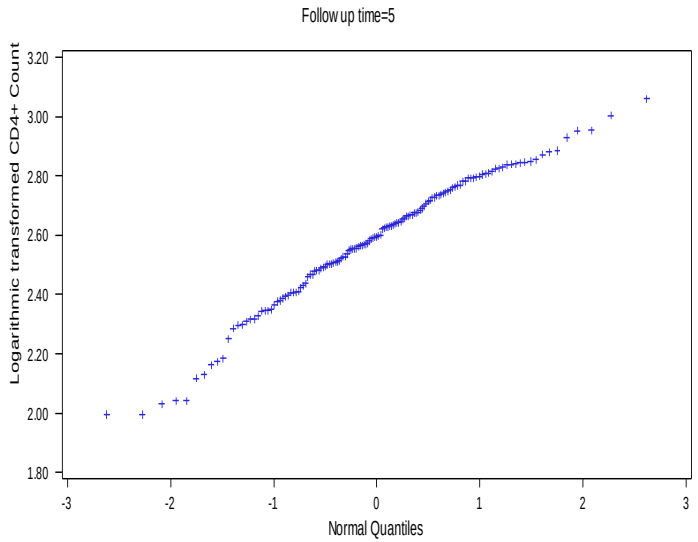

Follow up time=0
Logarithmic transformed CD4+ Count
3.50
3.00
2.50
2.00
1.50
1.00
0.50
-3
-2
-1
0
1
2
3
Normal Quantiles

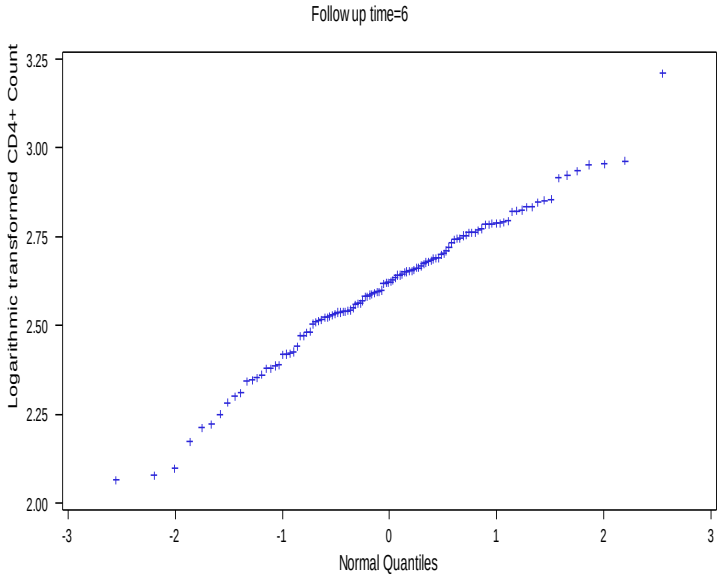

Follow up time=1
Logarithmic transformed CD4+ Count
3.50
3.00
2.50
2.00
1.50
1.00
-3
-2
-1
0
1
2
3
Normal Quantiles

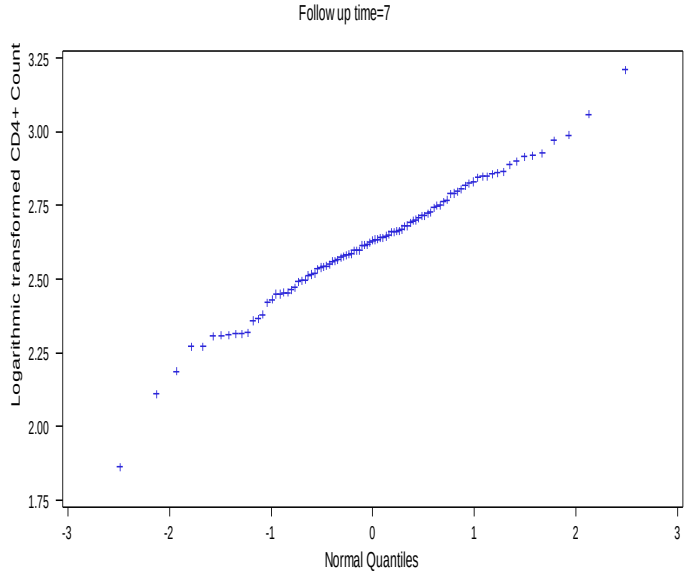

Follow up time=2
Logarithmic transformed CD4+ Count
3.25
3.00
2.75
2.50
2.25
2.00
1.75
1.50
-3
-2
-1
0
1
2
3
Normal Quantiles


Follow up time=3
Logarithmic transformed CD4+ Count
3.25
3.00
2.75
2.50
2.25
2.00
1.75
1.50
-3
-2
-1
0
1
2
3
Normal Quantiles


Follow up time=4
Logarithmic transformed CD4+ Count
3.25
3.00
2.75
2.50
2.25
2.00
1.75
-3
-2
-1
0
1
2
3
Normal Quantiles


Follow up time=5
Logarithmic transformed CD4+ Count
3.20
3.00
2.80
2.60
2.40
2.20
2.00
1.80
-3
-2
-1
0
1
2
3
Normal Quantiles


Follow up time=6
Logarithmic transformed CD4+ Count
3.25
3.00
2.75
2.50
2.25
2.00
-3
-2
-1
0
1
2
3
Normal Quantiles


Follow up time=7
Logarithmic transformed CD4+ Count
3.25
3.00
2.75
2.50
2.25
2.00
1.75
-3
-2
-1
0
1
2
3
Normal Quantiles