# MODELING THE PROGRESSION OF HIV INFECTION USING LONGITUDINALLY MEASURED CD4 COUNT FOR HIV POSITIVE PATIENTS FOLLOWING HIGHLY ACTIVE ANTIRETROVIRAL THERAPY

By:

**Taye Abuhay Zewale**

**A Thesis Submitted to the department of Statistics, School of Graduate Studies, College of Natural Science, Jimma University in Partial Fulfillment for the Requirements of Master of Science (MSc.) Degree in Biostatistics**

**August, 2012**

**Jimma, Ethiopia**

# MODELING THE PROGRESSION OF HIV INFECTION USING LONGITUDINALLY MEASURED CD4 COUNT FOR HIV POSITIVE PATIENTS FOLLOWING HIGHLY ACTIVE ANTIRETROVIRAL THERAPY

**M.SC. THESIS**

**TAYE ABUHAY ZEWALE**

**MAIN ADVISOR: WONDWOSEN KASSAHUN (PhD scholar)**

**CO-ADVISOR: TADELE AKEBA (MSC.)**

**AUGUST 2012**

**JIMMA UNIVERSITY**

**DEPARTMENT OF STATISTICS, SCHOOL OF GRADUATE STUDIES**

**JIMMA UNIVERSITY**

As thesis research advisors, we herby certify that we have read and evaluated the thesis prepared by **TAYE ABUHAY** under our guidance, which is entitled **Modeling the progression of HIV infection using longitudinally measured CD4 count for HIV positive patients following Highly Active Anti-Retroviral Therapy**. We recommend that the thesis be submitted as it fulfills the requirements for the degree of Master of Science.

**Wondwosen Kassahun (PhD scholar) -- --------------------** --- --------------------
Main advisor Signature date

**Tadele Akeba (MSc.)** --------------------- -------------------
Co-advisor signature date

As the members of the board of examiners of MSc. thesis open defense examination of
**TAYE ABUHAY ZEWALE**, we certify that we have read and evaluated the thesis and examined the candidate. We recommend that the thesis be accepted as it fulfills the requirements for the degree of Master of Science in Biostatistics.

--------------------------------------------------- ------------------------------ ---------------------------
Name of chairman Signature Date

---------------------------------------------------- -------------------------------- ----------------------
Name of Main Advisor Signature Date

--------------------------------------------------- ------------------------------------- --------------------
Name of Co-advisor Signature Date

----------------------------------------------------- ------------------------------------ -------------------
Name of internal Examiner Signature Date

--------------------------------------------------- ------------------------------------- --------------
Name of External Examiner Signature Date

# DEDICATION

This thesis is dedicated to my mother, Yigardu Kebede who was with me in her terrible time

With a great patience

## STATEMENT OF THE AUTHOR

First, I declare that this thesis is a result of my genuine work and all sources of materials used for writing it have been duly acknowledged. I have submitted this thesis to Jimma University in partial fulfillment for the Degree of Master of Science. The thesis can be deposited in the library of the university to be made available to borrowers for reference. I solemnly 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.

Brief quotations from this thesis are allowed without requiring special permission provided that an accurate acknowledgement of the source is made. Requisites for extended quotations for the reproduction of the thesis in whole or in part may be granted by the head of the department of statistics when in his or her judgment the proposed use of the material is for a scholarly interest.

In all other instances, however, permission must be obtained from the author.

Name**: TAYE ABUHAY** Signature -----------------------

Place: Jimma University, Jimma

Date of submission: ---------------------

# ACKNOWLEDGMENT

First and foremost I wish to thank God as all is and has been possible through him.

I am very thankful to my Main Advisor, Wondwosen Kasahun (PhD scholar), who provided valuable guidance and inspiration starting from selection of title up to the end of my thesis. Throughout my thesis development, he provided encouragement, sound advice and good teaching.

I would like to thank also my Co-Advisor Tadele Akeba (Msc.) who gives me all the necessary advices and comments regarding my thesis development.

Finally, I wish to thank Jimma University Specialized Hospital ART staffs who allowed and facilitate all the inputs for data collection.

# ABSTRACT

**Introduction:** 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. Though, cure or vaccine for HIV or AIDS does not currently exist, great strides have been made in treatment termed as Highly Active Anti-Retroviral therapy (HAART) which effectively lowers the concentration of the virus in the body by increasing the immune system of the body ($CD4^+$ T-cells).

**Objectives:** A longitudinal retrospective based study was conducted between January 1999 and December 2004 E.C in Jimma University Specialized Hospital. The objectives of the study were to model and study the progression of HIV infection using longitudinally measured CD4 count for HIV positive patients following Highly Active Anti-Retroviral therapy (HAART) and to identify factors predicting the progression of HIV infection.

**Subjects and methods:** A total of 1504 HIV positive individuals whose age 18 years and above were included in the study. The data was extracted from medical charts in a checklist format. Generalized linear mixed model and generalized estimating equations were used to model the progression of HIV infection.

**Results:** On average CD4 count increase in a quadratic pattern over time. In addition, the progression of CD4 count depends on patient's baseline demographic and clinical characteristics as well as GLMM fits the CD4 count data better than GEE.

**Conclusions:** On average CD4 count increases after patients initiated to the HAART program (the disease rate declines). No strong evidence suggesting that significant associations between progression of HIV infection and time by sex interaction. Progression of CD4 count differ by patients base line demographic and clinical characteristics.

# LIST OF ACRONYMS

| | |
|---|---|
| AIDS | Acquired Immunodeficiency Syndrome |
| HIV | Human Immune Deficiency Virus |
| HIV 1 | Human Immune Deficiency Virus Type 1 |
| UNAIDS | Joint United Nations Program on HIV/AIDS |
| WHO | World Health Organization |
| GEE | Generalized Estimating Equation |
| GLMM | Generalized Linear Mixed Model |
| LMM | Linear Mixed Model |
| HAART | Highly Active Anti-Retroviral Therapy |
| ART | Anti-Retroviral Therapy |
| ARV | Anti-Retroviral |

## TABLE OF CONTENT

**CONTENT** **PAGE**

## LIST OF TABLES

**Tables** **Page**

## LIST OF FIGURES

**Figures** **pages**

## LIST OF ANNEXES

# Glossary

| | |
|---|---|
| Acquired Immunodeficiency Syndrome (AIDS) | ❖ The most severe manifestation of infection with HIV. There are many opportunistic infections and cancers that constitute an AIDS diagnosis in the presence of HIV infection. |
| Antiretroviral drug | ❖ Substance used to kill or inhibit the multiplication of retroviruses such as HIV. |
| Antiretroviral therapy (ART) | ❖ Treatment regimens recommended by leading HIV experts to aggressively suppress viral replication and progress of HIV disease. The usual ART regimen combines three or more different drugs. |
| ART side-effect | ❖ An unwanted effect caused by the administration of ART. Onset may be sudden or develop over time. |
| Human immunodeficiency virus (HIV) | ❖ The retrovirus isolated and recognized as the etiologic (i.e, causing or contributing to the cause of the disease) agent of AIDS. |
| WHO clinical stages of AIDS | ❖ Classification of the stages of HIV-associated clinical disease where stage1 indicates asymptomatic disease, stage2 indicates mild disease, stage3 indicates advanced disease and stage4 indicates severe disease. |
| Ambulatory | ❖ An individual able to perform activities for daily living. |
| Bed-ridden | ❖ An individual unable to perform activities of daily living. |
| Working | ❖ An individual able to perform usual work in and out of the house, harvest, go to school for children, normal activities or playing. |

# CHAPTER ONE

## 1 INTRODUCTION

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). A cure or vaccine for HIV or AIDS does not currently exist. However, great strides have been made in treatment termed as Highly Active Anti-Retroviral Therapy (HAART). HAART consists of cocktails of at least two to three different classes of antiretroviral therapies and effectively lowers the concentration of the virus in the body by increasing the immune system which is called CD4$^+$ T cells. In most developed countries, where these drugs are available, a large reduction in HIV-associated morbidity and mortality has occurred to the extent that HIV/AIDS is now well thought-out a chronic condition.

Conceptually, the progression of HIV infection from an asymptotic stage to acquired immunodeficiency syndrome, AIDS, is associated with a gradual decline in the total number of CD4$^+$ T cells in the blood. Biologically, the decrease in the total number of CD4$^+$ T cells also correlates an increase in the number of infected T cells and an increase in the amount of free virus in the blood (Venet *et al*., 1991). CD4$^+$ T cells are the first reliable marker for disease progression since it gives an indication of how the immune system is doing and it is known to have a strong prognostic influence on progression to AIDS. These considerations motivate interests in the CD4$^+$ T-cell as a useful intermediate response variable in the assessment of progression of HIV disease (Berzuini and Allemani Reviewed work, 2004). understanding these key markers do not only let health providers to monitor a patient's health, but also allows researchers to discover essential information regarding the mechanisms of the virus and the human immune system. Besides to this, it also allows researchers to develop HIV vaccines, new efficient treatments and preventative measures such as microbicides and pre-exposure prophylaxis (PrEP). Now a day, CD4$^+$ T-cells over time is used for clinical management by repeatedly measuring it from individuals who tested HIV positive to monitor the progression of the disease.  A number of attempts have been made to determine the nature of progression of CD4$^+$ T-cells after HIV infection. Eyster, Gail, and Ballard 1987 suggested that the progression of CD4$^+$ cells follows a pattern of long period of slow decline followed by a rapid decline just before onset of AIDS. Therefore, this needs further investigation for the question what happens about the progression after the start of HAART, which was examined in this study. A recent study showed HAART brings a significance improvement to CD4$^+$ T-cells and provided further quantitative evidence about aspects of the therapy effect such as the changes in slope in CD4$^+$ cells count profile (Berzuini and Allemani, 2004). Moreover, patients who had already experienced an AIDS-defining event at the point of initiating HAART were also at higher risk of developing a new event, irrespective of their CD4$^+$ T-cells evolution during treatment. These patients might have a lower capacity to restore their immunity (Binquet *et al*. 2001).

Further study indicated that in the population model, both men and women had a significant change in CD4$^+$ cells count within 2 years after initiated to HAART. However, in the individual model, both men and women gained significant change in CD4$^+$ T-cells after 7 years of HAART initiation (Watcharathanakij, 2007). Although evidence of a beneficial gain in CD4$^+$ T-cells following administration of HAART in HIV-1 infection is available (Palella *et al*., 1998), several questions about treatment effect deserve further investigation. For example, does its effect show the same evolution in their immune system for all groups? What is the magnitude of such a change in slope? Thus this study evaluated the above questions and others.

To study the progression of HIV infection, the CD4$^+$ T-cells should be measured repeatedly per individual what is called longitudinal data. Since the measurements are correlated within individuals, the classical regression techniques couldn’t use rather the most flexible and powerful models were employed to handle such types of data. This includes generalized linear mixed model and generalized estimating equations which are capable of analyzing correlated and non normal data (i.e count in this case). Generalized linear mixed models (GLMM) (Breslow and Clayton, 1993) are obtained from generalized linear models (McCullagh and Nelder, 1989) by incorporating random effects into the linear predictors, and include the well known linear mixed models (LMMs) for normal responses (Laird and Ware, 1982) as a special case. These models are useful for modeling the dependence among response variables inherent in longitudinal or repeated measures studies, for accommodating over dispersion among binomial or Poisson responses, and for producing shrinkage estimators in multi-parameter problems. GLMMs are parametric, and are estimated using maximum likelihood theory or associated methods.

The Generalized Estimating Equations (GEE) approach projected by Liang and Zeger (1986) is a class of estimating equations which take into account the correlation arising due to a longitudinal study design, to increase efficiency of standard error estimates. As introduced by Wedderburn (1972), the GEE approach is based on quasi likelihood theory and can be used for continuous as well as for discrete outcome. The GEE method is a multivariate generalization of quasi-likelihood, and this method is mainly proposed for marginal modeling with GLM.  It avoids the use of multivariate distribution by assuming a functional form for marginal distribution at each time, making it useful for non-Gaussian outcomes. The advantage of using the GEE method is that the solutions are consistent, i.e. the estimate of parameters are nearly efficient and asymptotically Gaussian, even when the time dependence is misspecified. GEEs are semi-parametric because the parameter estimates are estimated parametrically and the variances are estimated non-parametrically. It also deals with the correlation caused by collecting numerous samples from each individual via adjusting the standard error to compensate for the lack of independence among samples.

## 1.1 Statement of the problem

$CD4^+$ T cells are cells of the immune system, which begin to deplete as the virus infects the body. These cells are considered as important biomarkers of disease progression for HIV infected individuals. Given the disease direct relation to the immune system, $CD4^+$ T cells are the primary indicator for prognostic information and a guide for antiretroviral therapy for HIV-positive individuals. Mostly, we are interested to know the number of $CD4^+$ T cells for a group of patients on their first visit in medical office to say something about the disease progress; however, since the number of $CD4^+$ T cells at one time point is not very instructive to tell about the disease status; the change in the number of cells over time is a good indicator of disease condition. A study defined a normal person has a range of 500-1,100cells per $mm^3$ $CD4^+$ T-cells, and this number decreases over time for an infected person (Goedert, 1989).

However, HAART fights against the progression of the disease by increasing the main body immune system or by decreasing RNA concentration (viral load) in the blood content. Even though, we can say HAART improves the immune system, there are many questions every one can raise about the numerical improvement. Among them how the $CD4^+$ cells count involve over time after patients initiated to HAART or do a change has different pattern depending on the patient's gender, educational level, and functional status e.t.c. In line with the above problem, there are many papers that have done about HIV biomarkers by using different regression techniques (Culshaw, 2006**,** Werner, 2009), but they have many drawbacks like accounting within and between subject variability by incorporating random effect in their models and also they did not use the most flexible and power full models to handle non normal data like GLMM and GEE. Besides to this, they do not put a clear explanation about the evolution of $CD4^+$ T cells after the patients initiated to HAART which ultimately indicates how the progression of the disease going on.

Generally, the basic research questions this study addressed were:

- ❖ How the average progressions of $CD4^+$ T-cells for HIV positive patients following HAART changes over time? Which indirectly tells us the disease progression?
- ❖ Which statistical method is appropriate to model the evolution of $CD4^+$ T cells by handling non normal and correlated data?

- What factors predict (influence) the evolution of $CD4^+$ T cells?
- Does the progression differ by patients demographic and clinical characteristics (sex, functional status, e.t.c)?

### 1.2 Significance of the Study

Even though, perfect eradication of Human Immunodeficiency Virus (HIV) from an individual is not currently possible, we can delay the progression of disease by using HAART which provide protection against the development of HIV-related complications by a long-term increase in $CD4^+$ T-cells. Therefore this study helps:-

- To understand the importance of attending HAART program by showing how much $CD4^+$ T-cells increase over time so that the patient's functional status and disease protective ability also increase or simply how much it delay disease progression.
- To provide a confidence for those afraid people to check their blood and save their life by attending HAART properly.
- To compare the different groups of patients how they respond to the drug simultaneously; so that it serve as a base for further study for the question what brings this variation and others.

# CHAPTER-TWO

# 2. LITRETURE

## 2.1 The Human Immunodeficiency Virus (HIV)

### 2.1.1 HIV and Human Body

The Human Immunodeficiency Virus (HIV) is a retrovirus that infects bodily fluids in humans and remains in the immune cells within these fluids. HIV targets these immune cells in order to replicate by damaging them in the process. These immune cells, $CD4^+$ T-cells and macrophages, play an important role in the body's immune system. The $CD4^+$ T-cells are mature T helper cells, a type of white blood cell, which expresses a surface protein. CD4 T-cells cannot kill infected cells or invading pathogens without other immune cells, they cannot fight infection in the human body. Their purpose is to activate and direct other immune cells which play a major role in fighting of disease. Macrophages are another type of white blood cell within tissues and they are also known as `eater cells ´since they remove dead cell material and pathogens. They also stimulate other immune cells to respond to the pathogen and are vital to the regulation of immune responses.

The $CD4^+$ T-cells are the primary entry point for HIV into the host. The virus attaches itself to the CD4 receptor via its own surface protein when exposed to the $CD4^+$ T-cells and makes use of the host cell to replicate itself and destroys it, impairing the functionality of the immune system. Hence that is why medical professionals rely on the $CD4^+$ cells count to decide on the state of the immune system and to make a decision when the patient needs to be initiated on HIV treatment. Within a few weeks of infection; there is a high level of replication in the blood that can exceed ten million viral particles per milliliter of blood. This rapid replication of viral particles is followed by a decline of $CD4^+$ T-cells in the body. However, after a few weeks the body develops its own immune response to the HIV which stops the viral replication and the viral load declines and the number of $CD4^+$ T-cells increase again to levels which are near normal. Thus infected individuals can remain asymptomatic for many years. However, it has been shown that during this time in which the person is feeling well, the body destroys up to a billion HIV particles and produces up to two billion $CD4^+$ T-cells a day (Abdool Karim, 2005). The virus continues to replicate, causing a gradual decline in CD4$^+$ T-cells, which in turn makes the individual susceptible to various opportunistic diseases such as TB and pneumonia.

### 2.1.2 Treatment of HIV

With the CD4$^+$ T-cells at dangerously low levels, an infected person's immunity is compromised and this person becomes prone to get opportunistic infections, such as TB, Cryptococcus Meningitis, Kaposi's sarcoma, Peripheral Neuropathy, and the like. The best way to prevent these opportunistic infections is to improve the level of immune function through highly active antiretroviral therapy (HAART), a combination of three or four different antiretroviral (ARV) drugs. When ARVs were first developed and used to treat HIV, only one drug was prescribed as treatment. Later, as different ARVs were developed and the medical community realized that patients were developing resistance to these ARVs, they started prescribing three or four concurrent ARVs as treatment and found this to be more effective in controlling HIV.

HAART has now become standard treatment and its main aim is to delay or prevent the progression to AIDS and death of those infected with HIV by suppressing and slowing down the replication of the virus. HAART maintains the reproductive number (Anderson and May, 1991) of the viral population below a threshold that cannot allow the viral population to increase and dominate. The World Health Organization (WHO) has recommended guidelines as when to start antiretroviral therapy (Panel on Antiretroviral Guidelines for Adults and Adolescents, 2008). Unfortunately, if a person is at an advanced stage of HIV/AIDS, when the CD4$^+$ T-cells is less than 50cells per cubic milliliter, then starting therapy would not be always successful. Once an HIV infected person initiates to HAART, he or she has to take it for the rest of his or her life in order to control the virus. There are different opinions on when HAART therapy should be initiated. Since the therapy will have to be continued for the rest of the infected person's life and thus many years, it is not advisable to start HAART immediately after testing HIV positive. Another reason for postponing treatment until it is absolutely necessary is that most of the ARV drugs have side-effects. There are some official guidelines regarding the initiation of HAART therapy. CD4$^+$ T-cells is used to make a decision as to when to begin HAART. In particular, the US Department of Health and Human Services (Panel on Antiretroviral Guidelines for Adults and Adolescents, 2008) recommend that HAART should be started if someone has an AIDS-defining illness or if their CD4$^+$ T-cells falls below 350cell per cubic milliliter. They also state that certain groups of people should be initiated on HAART regardless of their CD4$^+$ T-cells, for example pregnant women, patients with HIV-related nephropathy (kidney disease) or patients co-infected with hepatitis B virus.

Guidelines for initiating therapy can differ between the developed and developing countries. In South Africa, the criteria for initiating HAART, according to the National Department of Health, is that a person must either have a CD4$^+$ T-cells under 200 cells per cubic milliliter or a WHO stage IV, regardless of the CD4$^+$ T-cells (National Department of Health South Africa, 2004). In 1990 the World Health Organization (WHO) developed a staging system for people infected with HIV. This system uses conditions and infections to classify someone with HIV into a particular stage, ranging from stage I to IV. The staging increases as the severity of the diseases increases with stage IV corresponding to full-blown AIDS (WHO, 1990). Since the introduction of HAART, there has been dramatic decrease in rates of mortality due to HIV/AIDS. It has changed the perceptions of the HIV/AIDS epidemic from it being viewed as a death sentence to be seen as just a manageable chronic illness. Although HAART is able to control viral replication, it cannot completely eradicate HIV which persists in the host cells. This storage of infected cells allows the virus to replicate when HAART is discontinued or when the therapy can no longer suppress the virus.

### 2.1.3 HIV INFECTION

HIV is among the highly infectious and pathogenic diseases with a high mortality rate. The spread of HIV is influenced by several individual based epidemiological factors such as age, gender, mobility, sexual partner profile and the presence of sexually transmitted infections (STI).There is a continuing, rising trend nationally in HIV infection levels among pregnant women attending public antenatal clinics. In order to analyze the prognosis of patients infected with HIV, it is possible to use the CD4$^+$ T-cells. These cells are a sub-group of lymphocytes (a type of white blood cell or leukocyte) that play an important role in establishing and maximizing the capabilities of the immune system. CD4$^+$ T-cells provided the first reliable marker of disease progression (AbdoolKarim, 2005) as compared to other possible markers and it is one of the markers most closely correlated with the stage of HIV infection (Prins *et al*., 1999). The CD4$^+$ T-cells, a main disease marker is repeatedly measured among those individuals who test HIV positive to monitor the progression of the disease since it is known that HIV/AIDS is a long wave event. It follows from this that longitudinal studies are needed to establish the effects on individuals or on group over time.

Diggle *et al*. (2002) reported that an uninfected individual has around 1100 cells per cubic milliliter of blood. The factors affecting the rate of change are critical. These factors can be grouped broadly into socio-demographic and biomedical factors. The purpose of this study is to describe the evolution of this HIV infection using $CD4^+$ T-cells in a cohort of acutely infected patients who were following up longitudinally over time and understand factors that predict the change in $CD4^+$ T-cells.  Strategies to prevent HIV/AIDS should include education to promote delayed onset of sexual activity since the HIV/AIDS is quite prevalent amongst women as compared to men in the age group 15-24 years. For social, cultural and economic reasons men are usually in a stronger position in their relationships with women and this gives them more control in deciding when to have sex as well as whether or not to use the condom.

Someone who is on HAART or on other immune boosting medication we expect their $CD4^+$ T-cells increase and their viral load to decrease. However, individual responses are quite variable and the correlation between $CD4^+$ T-cells responses and viral load in some individual is very weak (Abdoolkarim, 2005). Higher $CD4^+$ T-cells imply a strong immune system while low $CD4^+$ T-cells imply a weak immune system. However $CD4^+$ T-cells do not always reflect how someone with HIV feels and functions; there could be other latent factors which influence the dynamics of the disease.

## 2.2 Models for HIV infection

Before the start of HAART, modeling the progression of HIV disease using the principal bio markers was not common, this is because knowing this biomarker is helpful only to provide a good care to patients by adjusting their functional status with the appropriate drug. However, after HAART was introduced, modeling of this bio markers became usual as the model helps to know how the disease progress through time so that one can know easily when the patient should start the medication, which type of the drug will appropriate and what is the rate of the disease progression as a result any one can predict what faces to the next time. In relation to this, by using the mixed model a cohort of HIV-1 infected patients being treated with protease inhibitors showed that the relation between changes in $CD4^+$ T-cells and risk of opportunistic infection varied over time (Binquet *et al*. 2001). It also recommended that statistically significant interaction between CD4$^+$ change and time is an additional factor in favor of a delayed but at least partial immune recovery. On the other hand fully Bayesian analysis of the progression of HIV infection using longitudinal CD4 T-cell numbers with a high-dimensional hierarchical mode investigated that these CD4 T-cell numbers simply do not carry all that much information about the progression of the disease (Lange *et al*. 1992).

Different characteristics of the individuals like age, base line clinical stage, plasma HIV RNA and time since HIV diagnosis are recognized predictors of progression of HIV infections (Ville's et al. 2007). In addition Bayeh et al. (2009) identified age, sex, educational levels and income levels of HIV patients as predictors of CD4 count progression after initiated to ART. The study found that at base line female's average CD4 count is greater than males and the average CD4 count through time increase.

### 2.2.1 Generalized linear mixed models (GLMMs) approach

Generalized linear mixed models (GLMMs) (Breslow and Clayton, 1993) are obtained from generalized linear models (GLMs) (McCullagh and Nelder.1986) by incorporating random effects in to the linear predictors, and include the well known linear mixed models (LMMs) for normal responses (Laird and Ware ,1982) as a special case. These models are useful for modeling the dependence among response variables inherent in longitudinal or repeated measures studies, for accommodating over dispersion among binomial or Poisson responses, and for producing shrinkage estimators in multi-parameter problems. Due to the wide range of applications of GLMMs, these models have received substantial attention during the last decade and are available in the major software packages. The computational burden associated with high dimensional numerical integration has limited past studies of GLMMs to the case of simplified models (e.g., random intercept models), to tractable random effects distributions (e.g., the Gaussian and conjugate distributions such as the beta-binomial and negative binomial models), or to conditional inference for the regression coefficients, conditioning on the random effects (Zeger and Karim, 1991).

A variety of novel approaches have been proposed to overcome the computational difficulties, with the goal to improve inference and estimation procedures for the fixed effects in GLMMs.

These include Gibbs sampling (Zeger and Karim, 1991), penalized quasi-likelihood and marginal quasi-likelihood (Breslow and Clayton, 1993), pseudo-likelihood based on approximate marginal models (Wolfinger and O'Connell, 1993), and maximum likelihood with Monte Carlo versions of EM, Newton-Raphson and simulated maximum likelihood algorithms (McCulloch, 1997), among many others (Jiang, 1998). These approaches typically require Gaussian distribution assumptions for the random effects. Methods for non-normal random effects are less common and limited to specialized cases (Magder and Zeger, 1996; Lee and Nelder, 1996, 2001; Gamerman, 1997).

One approach to account for the within subject association is via the introduction of random effects in generalized linear models. This leads to a class of models known as generalized linear mixed models (GLMMs). GLMMs are an extension to GLMs that includes random effects in the linear predictor, giving an explicit probability model that explains the origin of the correlations. The resulting subject-specific parameter estimates are suitable when the focus is on estimating the effect of changing one or more components of the predictor on a given individual. In statistics, a generalized linear mixed model (GLMM) is a particular type of mixed model. Fitting such models by maximum likelihood involves integrating over these random effects.

### 2.2.2 Generalized Estimating Equation (GEE) approach

When interest is in the first-order marginal parameters, McCullough and Nelder (1989) have shown that a full likelihood procedure can be replaced by quasi-likelihood based methods. Wedderburn (1974) shows the likelihood and quasi-likelihood theories coincide for exponential families and that the quasi-likelihood estimating equations provide consistent estimates of regression parameter. In any generalized linear model, even for choices of link and variance functions that do not correspond to exponential families. Consequently, Liang and Zeger (1986) proposed the method of generalized estimating equations (GEE) as an extension of GLM to accommodate correlated data using quasi-likelihood approach. Rather than assuming a particular distribution for the response, GEE method requires a correct specification of the mean as well as how the variance depends on the mean. One of the desirable properties of the GEE method is that it yields consistent and asymptotically normal solutions even with the misspecification of the covariance structure (Liang and Zeger, 1986; Davis, 2002).

Over the past 20 years, the GEE approach has proven to be an exceedingly useful method for the analysis of longitudinal data, especially when the response variable is discrete (e.g., binary, ordinal, or a count). Correlated data are modeled using the same link function and linear predictor setup (systematic component) as the independence case. The random component is described by the same variance functions as in the independence case, but the covariance structure of the correlated measurements must also be modeled. The focus is on estimating the average response over the population (“population-averaged effects) rather than the regression parameters that would enable prediction of the effect of changing one or more components of X on a given individual

# CHAPTER THREE

## 3 Objectives of the study

### 3.1 General objective:

The general objective of this paper is to model and study the progression of HIV infection using longitudinally measured CD4$^+$ T- cells for HIV positive patients following HAART in Jimma University Specialized Hospital.

### 3.2 Specific objectives

- To explore the general average progression of CD4$^+$ T cells over time
- To model the progression of CD4$^+$ T-cells using GLMM and GEE and to compare them
- To identify factors which predict the progression of CD4$^+$ T- cells (disease progression)
- To compare the progression of CD4$^+$ T -cells or the progression of HIV infection between groups (i.e sex, educational level, who stage e.t.c).

# CHAPTER FOUR

# 4. Data and Methodology

## 4.1 Data source and its description

This study used the latest data from medical charts of HIV positive patients initiated to HAART from 2007-2011 in Jimma University Specialized Hospital located in south west of Ethiopia in Jimma town. All patients who are 18 and above years of age and measured their $CD4^+$ T cells at least once constitutes the study population. The data consists of 1504 individuals with a minimum of one and maximum of ten measurements per individual. Originally, the data was recorded by ICAP (International Center for AIDS care and treatment Program) incollabration with Jimma University Specialized Hospital from voluntary as well as medically ordered individuals directly to provide health care service for them. It was recorded on medical charts by assigning an identification number per individual which helps to find the patients profile easily during his/her next visit time. By taking a legal permission letter from the statistics department, the data was collected via extracting the required variables from medical charts in a check list format.

## 4.2 Study variables

**Dependent variable:** $CD4^+$ T- cells (CD4 count) for each individual measured in
Every six Month interval

**Independent variables:** Age, Weight, Sex, Functional status, Educational level,
WHO stage, Time (in Month)

*Table4.1: Variable description and coding for ART data taken at Jimma University Specialized Hospital from 2007 –2011*

| Variable | Coding |
|---|---|
| Age | Age of patients at the start of ART in years |
| Weight | Weight of patients at the start of ART in Kg |
| Sex | Male = 1, Female = 0 |
| Educational level | No Education = 0, Primary = 1, secondary = 2,tertiary = 3 |
| WHO stage | Stage I = 1, Stage II = 2, Stage III = 3, Stage Iv = 4 |
| Functional status | Working = 0, Ambulatory = 1, Bed ridden = 2 |
| Time | Observation time of CD4$^+$ T cells (Every six months) |
| CD4$^+$ T cells | number of cells per cubic millimeter measured for each individual in every six month |

## 4.3 Statistical Analysis technique

### 4.3.1 Exploratory data analysis

It is a technique to visualize the patterns of data relative to research interests. Since exploratory data analysis can serve to discover as much of the information regarding raw data as possible, plotting individual curves to carefully examine the data should be performed first before any formal model fitting is carried out. Thus, this study assessed the nature of the data by exploring individual profiles, the average evolution, the variance function, the correlation structure.

**I. Exploring the individual profile:** To explore the individual profile, plot of the response with time is used to show whether there is a noticeable pattern common to most subjects. These individual profiles can also provide some information on within and between subject variability.

**II. Exploring the Mean Structure:**

The major purpose of exploring the mean structure is to choose the fixed effects for the model. To explore the overall mean, we plot the response variable against time including individual and overall mean profiles. In line with the overall mean, the possible differences between the groups were studied by plotting the mean of each group separately with the same figure.

**III. Exploring the Variance structure:** To explore the variance structure of the data three plots were used. The first one shows the average evolution of the variance as a function of time and the second produces the individual profile plots of the data which shows whether there is a considerable within and between subject variability. The third is the interaction plot which is used to plot the variance functions separated for different groups (sex: male, female e.t.c) as a function of time.

**IV. Exploring the Correlation Structure:** It helps to describe how measurements within an individual correlate. Pair-wise scatter plots matrix was employed for exploring the correlation structure.

### 4.3.2 The statistical Models

#### 4.3.2.1 Generalized Estimating Equations (GEE)

GEE were introduced by Liang and zeger (1986) as a method of dealing with correlated data when, except for the correlation among responses, the data can be modeled as a generalized linear model. Unlike general linear mixed models, models using the GEE method are marginal models that only estimate population average regression coefficients. These models are not flexible enough to specify heterogeneity of the covariance structures. However, fitting models using the GEE approach has been shown to give consistent estimators of the regression coefficients and their variances under weak assumptions about the actual correlation among a subject's observations. A GEE is simply the equation you solve to calculate the parameter estimates and it accommodates the correlation structure of the repeated measurements. It needs to specify only the relationships between the response mean and covariates and between the response mean and variance. GEE has a "working" correlation *R* of the repeated measurements. This working correlation matrix is of size n×n because one assumes that there is a fixed number of time-points *n* that subjects are measured. At a given subject does not have to be measured at all *n* time-points. Each individual's correlation matrix $R_i$ is of size $n_i \times n_i$ with appropriate rows and columns removed if $n_i < n$. It is generally recommended that choice of *R* should be consistent with the observed correlations. With GEE the relationships between the variables of the model at different time-points are analyzed simultaneously.

**The marginal mean model:** We assumed that N patients measured repeatedly through time and let $Y_{ij}$ denote the response for $i^{th}$ patient at $j^{th}$ time. $Y_{ij}$ is count response variable with non negative integer values. Each $Y_{ij}$follows a poisson distribution and the mean is related to X by a log link function.

$$g(\mu_{ij}) = \text{Log}(\mu_{ij}) = X_{ij}t\beta$$

Where,

$\mu_{ij}$: The mean of $Y_{ij}$, which is related to the covariates of $X_{ij}t$ by link function

$X_{ij}$: A px1 vector of covariates

β: A px1 vector of unknown regression coefficients of X, and

g(.): Log link function as $Y_{ij}$is count.

**Method of estimation and statistical inference:** The GEE approach is appealing for analysis of non normal data because of its computational simplicity compared to the maximum likelihood based approaches. However, because there is no likelihood function, likelihood based methods are not available for testing fit, comparing models and conducting inference about parameters rather a quasi-likelihood method of estimation is used. Instead inference can only use Wald statistics constructed with asymptotic normality of the estimators together with their estimated covariance matrix. Moreover, even though GEE estimates are consistent with misspecification of the covariance structure, it is important to choose the covariance structure that closely approximates the true underlying one for greater efficiency.

**Working correlation structures:** Because the repeated observations within one subject are not independent of each other, a correction must be made for these within-subject correlations. With GEE, this correction is carried out by assuming a priori certain 'working' correlation structure for the repeated measurements of the outcome variable *Y*. Before carrying out a GEE analysis, the within-subject correlation structure was chosen based on the results of exploring correlation structure of the observed data. Accordingly two propose working correlations were compared.

**I. Independent structure:** This is the correlation that GEE model assumes by default. With this structure the correlations between subsequent measurements are assumed to be zero or measurements are independent to each other within individuals.

II. **Exchangeable correlation structure (compound symmetry)**: it assumes the correlations between subsequent measurements are assumed to be the same, irrespective of the length of the time interval.

Generally, assuming no missing data, the J x J covariance matrix for y is modeled as:

$V_i = \Phi A_i^{1/2} R_i A_i^{1/2}$ ………………..1.2

Where Φ is a glm dispersion parameter which is assumed 1 for count data, A$_i$ is a diagonal matrix of variance functions, and $R_i$ is the working correlation matrix of Y.

Generalized estimating equations (GEEs) can be used to model correlated data with the variance covariance matrix **V** by iteratively solving the quasi- score equations.

The score function of a GEE for $\boldsymbol{\beta}$ has the form

$$\sum_{i=1}^{N}(\frac{\partial \mu_i}{\partial \beta_j})V_i^{-1}(Y_i-\mu_i)= 0 \text{ ……………………..1.3}$$

Where $\mu_i$ is the fitted mean, which is given by g$(\mu_{it})$= $x_{it}\beta$ for covariates x = $x_{i1}$, $x_{i2}$, . . . , $x_{im}$ and regression parameters $\beta = \beta_1, \beta_2, \ldots \beta_p$

Starting $R_i$ as the identity matrix and Φ=1, the parameters $\beta$ are estimated by solving equations as follows.

**i.e in normal case** $\mu_i = x_i\beta$ and $\frac{\partial \mu_i}{\partial \beta_j} = x_i , V_i = \hat{\Phi}\,\mathrm{R}_i$

$\sum_{i=1}^{N}(x_i^{t})\mathrm{R}_i^{-1}(Y_i-\mu_i) = 0$

$\hat{\beta} = [\sum_{i=1}^{N}(x_i^{t})\mathrm{R}_i^{-1}x_i]^{-1}[\sum_{i=1}^{N}(x_i^{t})\mathrm{R}_i^{-1}y_i]$

$$\widehat{\beta^{t+1}} = \widehat{\beta^{t}} - \{\sum_{i=1}^{N}(\frac{\partial \mu_i}{\partial \beta_j})^t V_i^{-1}(\frac{\partial \mu_i}{\partial \beta_j})\}^{-1}\sum_{i=1}^{N}(\frac{\partial \mu_i}{\partial \beta_j})^t V_i^{-1}(Y_i-\mu_i)$$

$let\ D_i = \frac{\partial \mu_i}{\partial \beta_j}\ AND\ S_i = (Y_i-\hat{\mu}_i)(Y_i-\hat{\mu}_i)^t$Then

1. $\beta^{t+1} = \beta^t - \{\sum_{i=1}^{N}(D_i)^t\hat{V}_i^{-1}(D_i)\}^{-1}\sum_{i=1}^{N}(D_i)^t\hat{V}_i^{-1}(D_i)$

2. var$(\widehat{\beta})$=N$\{\sum_{i=1}^{N}(D_i)^t\hat{V}_i^{-1}(D_i)\}^{-1}\{\sum_{i=1}^{N}(D_i)^t\hat{V}_i^{-1}\hat{S}_i\hat{V}_i^{-1}(D_i)\ \{\sum_{i=1}^{N}(D_i)^t\hat{V}_i^{-1}(D_i)\}^{-1}$

More generally, because solution only depends on the mean and variance of y, these are quasi-likelihood estimates. The estimates from a GEE analysis are robust to miss-specification of the covariance matrix (Liang & Zeger, 1986), so, the regression parameter estimates are consistent even for independent covariance matrix.

Upon convergence, in order to perform hypothesis tests and construct confidence intervals, it is of interest to obtain standard errors associated with the estimated regression coefficients. These standard errors are obtained as the square root of the diagonal elements of the matrix $V(\overline{\widehat{\beta}})$ . The GEE provides two versions of these estimates.

1. Naive or "model-based",

$$V(\overline{\widehat{\beta}}) = \{\sum_{i=1}^{N} (D_i)^t \widehat{V}_i^{-1} ( D_i)\}^{-1}$$

2. Robust or “empirical”

$$V(\overline{\widehat{\beta}}) = M_o^{-1} M_1 M_o^{-1}$$

Where,

$$M_0 = \Sigma_{i=1}^{N} (D_i)^t \widehat{V}_i^{-1} ( D_i)$$

$$M_1 = \Sigma (D_i)^t \widehat{V}_i^{-1} ( Y - \hat{\mu}_i) ( Y - \hat{\mu}_i)^t \widehat{V}_i^{-1} ( D_i)$$

Here, $\widehat{V}_i$ denotes $( Y - \hat{\mu}_i)( Y - \hat{\mu}_i)^t$

In the more general case, the robust or “sandwich” estimator, which is due to Royal1 (1986), provides a consistent estimator of $V(\overline{\widehat{\beta}})$ (even if the working correlation structure $R_i$ is not the true correlation of *yi*.

**Variable selection technique:** In both models, to select significant variables, first the main effect and main effect by time interaction will be incorporated to the initial candidate model. After that, avoid non significant variables one by one starting from the most non significant terms and finally the two models are compared using generalized Wald test for GEE and likelihood ratio test for GLMM (Patetta, 2002) or simply backward selection technique.

#### 4.3.2.1.1 Model comparison technique

**Quasi-information criterion:** Although the AIC can be used in association with mixed models, it cannot be used with GEEs to select either the optimal set of explanatory variables or correlation matrix, because GEE estimation is based on the quasi-likelihood rather than the maximum likelihood. The quasi-likelihood counterpart to the AIC is the QIC, or the “quasi-likelihood under the independence model information criterion” (Pan, 2001). The QIC was derived from the AIC and is conceptually similar. The quasi-likelihood function takes the following form (McCullagh and Nelder, 1989)

$\mathrm{Q}(\mu) = \int_{y}^{\mu} \frac{y-t}{\Phi \mathrm{v(t)}} \mathrm{dt}$ , Where $\mu$= E(y) and var(y) =$\Phi$v($\mu$) with $\Phi$ being the dispersion parameter.

An equation for the QIC is

$$\mathrm{QIC} = -2\mathrm{Q}\,(\hat{\mu},\mathrm{I}) + 2\mathrm{trace}[(\Omega_I{}^{-1}\widehat{V_R}]$$

Where **I** represent the independent correlation structure and ***R*** is the specified working correlation structure. The *p*-dimensional matrices $\Omega_I{}^{-1}$and $\widehat{V_R}$ are variance estimators of the regression coefficients under the correlation structure **I** and **R** respectively. The QIC value is computed based on the quasi-likelihood estimate $\hat{\mu}$ and is used to select the appropriate working correlation structure for the model. However, Hin and Wang (2009) proposed using half of the second term in QIC is appropriate for the selection of the working correlation structure in GEE. This statistic is called the **Correlation Information Criterion (CIC).**

$$\mathbf{CIC} = \mathbf{trace}[(\boldsymbol{\Omega}_I{}^{-1}\widehat{\boldsymbol{V}_R}]$$

The first term in QIC, which is based on the quasi-likelihood, is free from both the working correlation structure as well as the true correlation structure, so it would not be informative in the selection of the covariance structure. Moreover, the form of quasi-likelihood is constructed under the assumption of the independent observations, although the parameters are estimated under the hypothesized working correlation structure. On the other hand, the second term in QIC contains information about the hypothesized correlation structure via the sandwich variance estimator. Even though the second term plays a role as a penalty term for mean model variable selection, the QIC is more heavily impacted by the first term. Hence, QIC is not particularly sensitive measure to use for selection of working correlation structure. Due to the above, Hin and Wang (2009) and Wang and Hin (2010) shown the performance of **CIC** to be much better than the **QIC** in selecting the correct correlation structure. Thus, this paper used **CIC** to compare two models with different working correlation structures.

In addition the two models with exchangeable and independence working correlation structures were compared via their naïve and robust standard error estimates and the one with the closest naïve and robust standard error estimates was preferred. The best correlation is usually selected first and is done based on the full model with all the explanatory variables (Hardin and Hilbe, 2003)

**The generalized Wald test:** is used to compare models with different subsets of the regression parameters. That is, one can use the generalized Wald tests to test the joint null hypothesis that a set of regression parameters $\boldsymbol{\beta}$s are equal to zero (Hedeker and Gibbons 2006). In general, for any matrix **L** a test for hypothesis can be written as follows

$HO{:}\,L\boldsymbol{\beta} = \mathbf{0}\ \ versus\, H1:\ L\boldsymbol{\beta} \neq \mathbf{0}$, Where L is a Pxp indicator matrix of ones and zeros. Here, *p* is equal to the number of parameters in the full model (including the intercept) and ***q*** equals the number of parameters in the generalized Wald test (that is, the difference in parametrs between the full and reduced model). The Wald statistic is a quadratic form defined as follows

$$W^2{}_{stat} = \ \bar{\boldsymbol{\beta}}^t\ \mathbf{L}^t(\mathbf{L}\,\mathbf{Var}(\boldsymbol{\beta}\,)\,\mathbf{L}^t)^{-1}\mathbf{L}\bar{\boldsymbol{\beta}}$$

And is distributed as $x^2$with *q* degrees of freedom under the null hypothesis

#### 4.3.2.2 Generalized linear mixed model (GLMM)

An alternative way to fit a longitudinal model to non-normal response data is to fit a generalized linear mixed model. These models are similar to the ones fit in GEE because the normality assumption regarding the error terms is relaxed. Some of the error distributions supported by generalized linear mixed models include the binomial, Poisson, gamma e.t.c. These models also support a large variety of link functions, which include the logit, log, and reciprocal. The type of response variable determines the distribution and link function for the model. Since the response variable for this paper is discrete count data the natural log link function was used. However, unlike the models fit in GEE, generalized linear mixed models have the flexibility to specify random effects and also to generate subject-specific parameter estimates. let $Y_{ij}$ denote the count response of CD4 count for i$^{th}$ patient at j$^{th}$ time, taking positive integers. The link function can be:

$Y_{ij}$ ~ Poisson ($\mu_{ij}$), log ($\mu_{ij}$)= $X_{ij}\beta + Z_{ij}b_i$

Where

$X_{ij}$: Covariates of the i$^{th}$ patient at j$^{th}$ time,

$\beta$ : Regression coefficients of $X_{ij}$

$\mu_{ij}$: The mean of $Y_{ij}$, which is related to the covariates of ${X_{ij}}^{t}$ by link function, and

$Z_{ij}$: The covariates of the random effects of the ith subject at jth time

$\boldsymbol{b_i}$**:** The random effect which are assumed to be multivariate normal distribution having mean vector **0** and covariance matrix G, i.e. $b \sim N(0, \mathrm{G})$

**Assumptions for GLMM** are:-

1. The conditional distribution of $Y_i$ given $b_i$ follows a distribution from the exponential family with density f($Y_i$ |$b_i$, $\beta$)

$$f(Y_i/b_i, \beta) = \exp\{\frac{Y_{ij}(\theta_{ij}) - \psi(\theta_{ij})}{\Phi} + \mathrm{c}(Y_{ij}, \Phi)\}\,)$$

$$\mu_{ij} = \mathrm{E}(Y_{ij}/b_i) = \psi'(\theta_{ij}) = g^{-1}(Y_{ij}\beta + z_{ij}b_i)$$

Where, $\mu_{ij}$ represents conditional means not marginal

$Y_{ij} = \mathrm{var}(Y_{ij}/b_i) = \psi''(\theta_{ij})$ =V($\mu_{ij}$), or $Y_{ij}$ ~N{ $\psi'(\theta_{ij}), \psi''(\theta_{ij})$}

Where g(.) is the link function $\boldsymbol{g^{-1}}$ (.) is the inverse of the link function) and v(.) is the variance function. Vector X and Z have p and q dimension respectively.

2. Given $b_i$, the repeated measurements $y_{i,1}$ . . . $y_{i,ni}$ are independent.

3. The $b_i$are independent and identically distributed with density function f($b_i$;G). Commonly, f($b_i$;G) is a normal distribution with zero mean and variance matrix G(i.e $b_i \sim N(0, G)$).and the error term is normally and independently distributed due to(2) (i.e $\epsilon_i$~N(0,$\sigma^2 I_i$) ). Correlation amongst observations from a unit, $y_{i1}$ , . . . $y_{i,ni}$ arises from their sharing unobservable variables $b_i$. The random effect model is most useful when the objective is to make inference about individuals rather than population averages.

**Method of estimation and statistical inference:** Maximum likelihood (ML) by Laplace approximation technique is used to estimate the parameters. ML estimates standard deviations of the random effects assume the fixed-effect estimates are correct. The following derivations are done with respect to ML. such likelihood may involve high-dimensional integrals that cannot be evaluated analytically so that much software are able to solve such complex manipulation using iteration technique. The likelihood of the data expressed as a function of unknown parameters is

$$L(\beta,\alpha,y) = \prod_{i=1}^{m} \int \prod_{i=1}^{ni} f(Y_{ij}/\beta, b_i) f(b_i, \alpha) db_i$$

It is the integral over the unobserved random effects of the joint distribution of the data and random effects. With Gaussian data, the integral has a closed form solution and relatively simple methods exist for maximizing the likelihood or restricted likelihood. With non-linear models, numerical techniques are needed. We consider the random effects as no missing data so that the 'complete' data for a unit is $(y_{i1}, b_i)$. Denote $L$= log (L) and $\mu_{ij} = g^{-1}(X_{ij}\beta + z_{ij}b_i)$ the score equation for $\beta$ and b are

$$\frac{\partial L}{\partial \beta} = S_\beta(\beta, \alpha/yb) = \sum_{i=1}^{m} \sum_{i=1}^{ni} x_{ij}(y_{ij} - \mu_{ij}) = 0$$

The score equation for G is

$$S_\beta(\beta, \alpha/yb) = \frac{1}{2} G^{-1}\{\mathrm{E}(b_i b_i^{\,t}/y_i\}G^{-1} - \frac{m}{2} G^{-1}$$

Where, G is variance covariance matrix for random effect. Hereby $\alpha$ denotes the unknown parameter in the density. These are solved using the E-M algorithm. In the estimation step, the expectations are evaluated using current parameter values and this may involve multivariable integration of large dimension. This will usually be done by Monte-Carlo integration.

### 4.3.2.2.1 Model comparison technique for GLMM

The primary objective of model comparison is to choose the simplest model that provides the best fit to the data. This study used Information criteria followed by Likelihood ratio test to select the best model.

**Akaike's information criterion (AIC)** is a measure of goodness of fit of an estimated statistical model. 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 is 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 and reduced models respectively and the test statistic is

Defined as ${T^2}_{LR} = -2ln\lambda_N = -2ln\left(\frac{L_{ML}(\hat{\alpha}_{ml,o})}{L_{ML}(\hat{\alpha}_{ml})}\right)$

Where, $\hat{\alpha}_{ml,0}$ 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.

### 4.3.2.2 Model Checking Technique

In GLMM, 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 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. Generally the data was analyzed using R software version 2.15.1

### 4.3.3 Ethical considerations

A permission to undertake the study has been obtained from Jimma University College of natural science through Ethical review board and official letter of co-operation was written by the department of statistics to Jimma University Specialized Hospital.

# CHAPTER FIVE

# 5. RESULT AND DISCUSSION

## 5.1 RESULTS

A total of 1504 HIV positive patients were included to this study. The baseline characteristics of patients are displayed in table2 below.  Among these patients, more than half 962 (64%) of them were females and 542(36%) were males. About 351(23%) patients were WHO stage I, 529(35%) stage II, 516(34%) stage III and 108(8%) stage IV. 534(36%) of patients had an education at primary level, 507(34%) were secondary, 158(10%) were at college or university level and 305(20%) patients had no education (illiterate). Of the patients, 1032(69%) had good functional status (were "working" classification), 409(27%) were ambulatory and only 63(4%) were bedridden.

*Table 5.1: Baseline demographic and clinical characteristics of ART data taken at Jimma University specialized Hospital from 2007 –2011*

| Characteristics | Category | n (%) |
|---|---|---|
| Sex | Male | 542(36%) |
| | Female* | 962(64%) |
| WHO stage | I* | 351(23%) |
| | II | 529(35%) |
| | III | 516(34%) |
| | IV | 108(8%) |
| Educational level | No education* | 305(20%) |
| | primary | 534(36%) |
| | secondary | 507(34%) |
| | tertiary | 158(10%) |
| Functional status | Working* | 1032(69%) |
| | ambulatory | 409(27%) |
| | bedridden | 63(4%) |

** indicates the reference group for each characteristic*

### 5.1.1 Exploratory analysis

#### 5.1.1.1 Exploring the individual profile over time

The individual profile plot for a sample of 20 subjects is given in figure 5.1. From the actual profile plot, variability of measurements between individuals seems lower at base line as compared to the follow up time and almost all of them gain CD4 count over time.

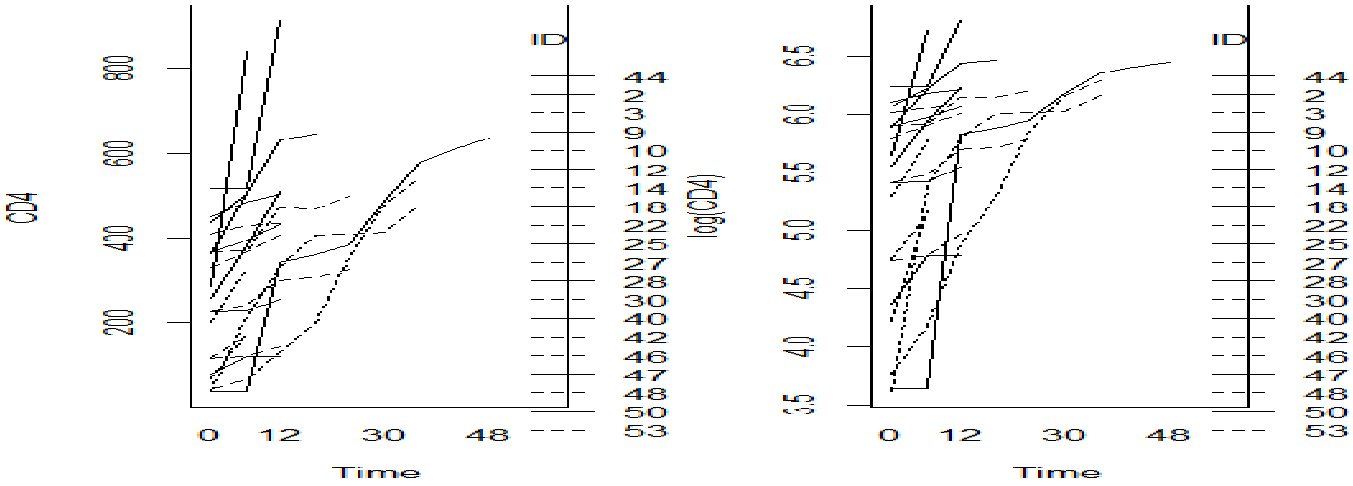

*Figure 5.1: a) Individual profile plot for the actual CD4 count and b) Individual profile plot for transformed CD4 count taken at Jimma University Specialized Hospital from 2007 –2011*

#### 5.1.1.2 Exploring the mean structure

To explore the mean structure, both the actual and transformed CD4 count mean profiles were considered.

*Table 5.2*: *The mean, median of patients CD4 count at each visit time taken at Jimma University Specialized Hospital from 2007 –2011*

| *Time(month)* | *0* | *6* | *12* | *18* | *24* | *30* | *36* | *42* | *48* | *54* |
|---|---|---|---|---|---|---|---|---|---|---|
| *Mean(CD4)* | *207.0* | *291.7* | *342.5* | *378.9* | *403.1* | *423.7* | *437.6* | *503.4* | *418* | *465.8* |
| *Mean(log(CD4))* | *4.96* | *5.47* | *5.67* | *5.80* | *5.85* | *5.90* | *5.94* | *6.07* | *5.93* | *6.06* |
| *median(CD4)* | *162* | *262* | *312* | *337* | *360* | *389* | *413* | *476.5* | *363* | *413* |

As shown in table 5.2, the actual and log transformed mean and median of the observed CD4 count increases over time. This means after patients initiated to ART, their CD4 count increases due to the positive effect of therapy.

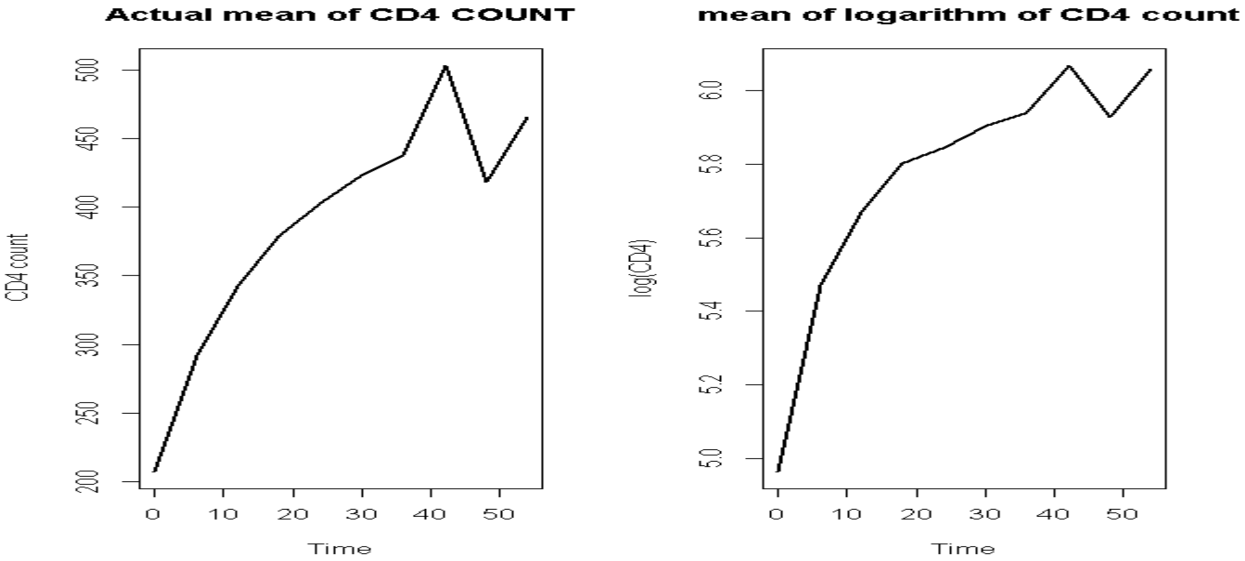

*Figure 5.2: a) The average progression of actual CD4 count for and b) The average progression of transformed CD4 count taken at Jimma University Specialized Hospital from 2007 –2011*

In figure 5.2, it seems that the mean CD4 count evolution shows a quadratic pattern for both actual and transformed CD4 count over time. It indicates the patient's immune system increase or the progression of the disease declines over time (i.e because CD4 count and HIV infection are negatively correlated).

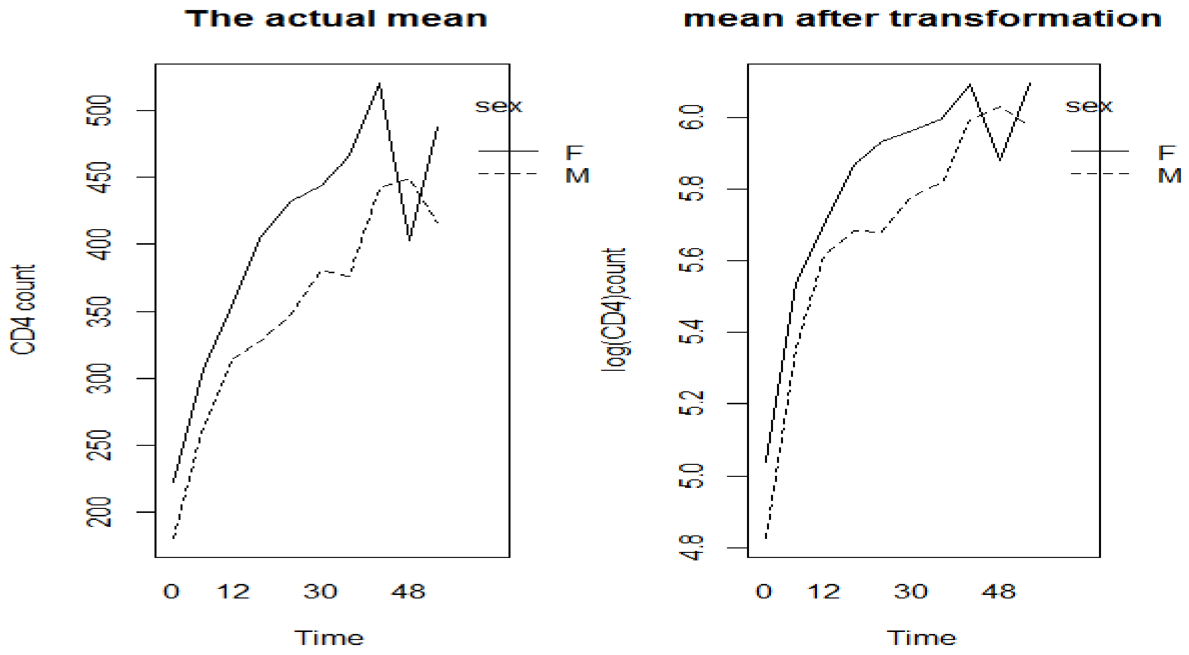

*Figure 5.3: a) the mean profile of the actual CD4 count by sex and b) the mean profile of transformed CD4 count by sex taken at Jimma University Specialized Hospital from 2007 –2011*

The average evolution for the actual and log transformed CD4 count by sex is presented in figure 5.3, where females appear to be higher than males until around 42 months. It also shows both males and females have increasing CD4 count in a quadratic pattern over time which supports the results of the general average evolution given in figure2.

### 5.1.1.3 Exploring the variance structure

Table 5.3: *The actual and transformed Variance of CD4 count at each visit time taken at Jimma University Specialized Hospital from 2007 –2011*

| Time(month) | 0 | 6 | 12 | 18 | 24 | 30 | 36 | 42 | 48 | 54 |
|---|---|---|---|---|---|---|---|---|---|---|
| Variance(CD4) | 37752.5 | 31755 | 36100 | 37597.2 | 43764.6 | 43388.9 | 51076 | 76176 | 75625 | 35532 |
| Variance(log(CD4)) | 0.923 | 0.513 | 0.416 | 0.301 | 0.396 | 0.400 | 0.356 | 0.337 | 0.241 | 0.187 |

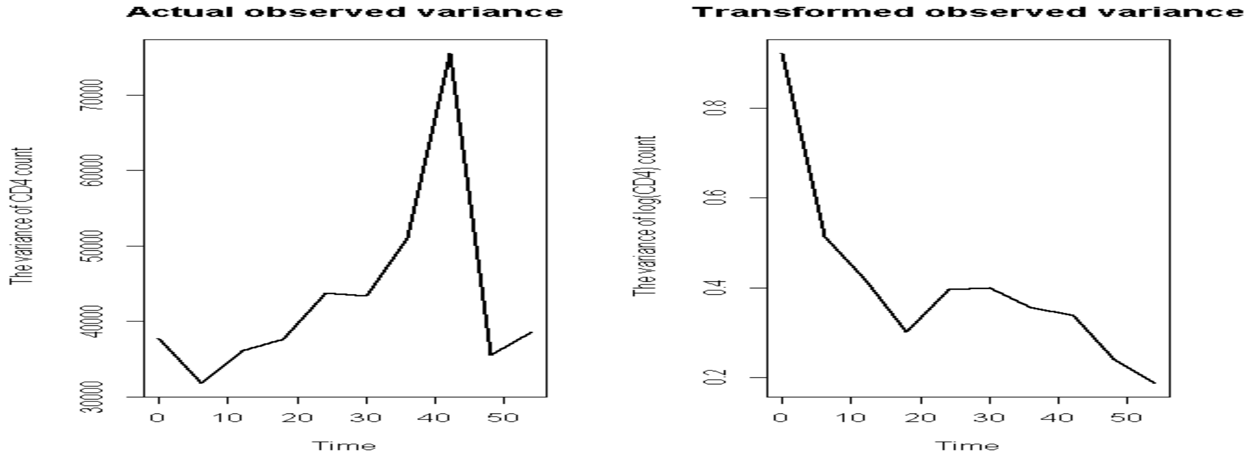

*Figure 5.4: a) The Average variance progression of actual CD4 count and b) The Average variance progression of transformed CD4 count taken at Jimma University Specialized Hospital from 2007 –2011*

The actual and transformed variance of CD4 count is given in table4 and figure 5.4; it seems high variability of CD4 count measurements over time, the actual variance plot shows the variability is small at base line and increases until around 42 months followed by rapid decreasing.

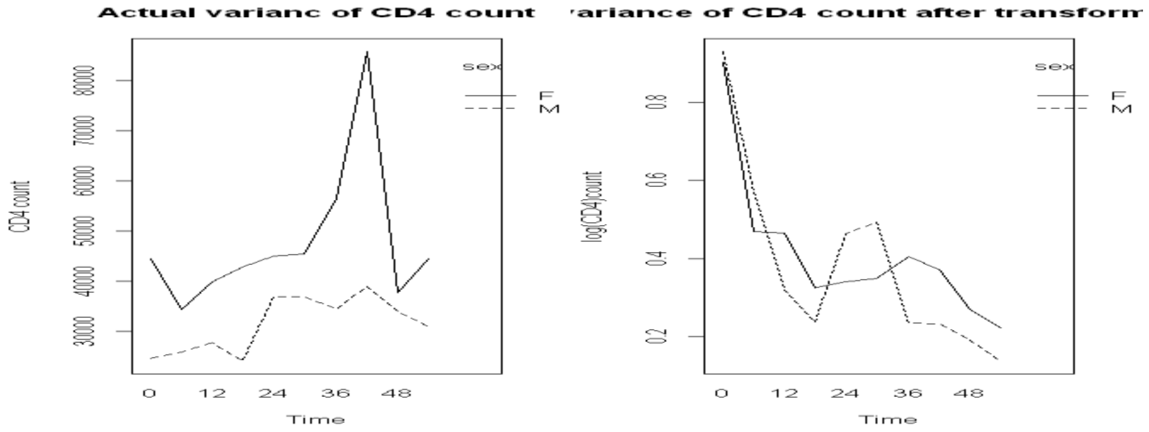

*Figure 5.5: a) The variance profile of the actual CD4 count by sex and b) The variance profile of the transformed CD4 count by sex taken at Jimma University Specialized Hospital from 2007 –2011*

The actual and transformed variance profile of CD4 count by sex is given in figure 5.5; it appears that the actual variance of female is higher than male over time though after transformation it seems existence of some interaction over time.

### 5.1.1.4 Exploring the correlation structure

From the pair wise scatter plot matrix given in figure 5.6, it looks like the correlation is nearly constant over time thus exchangeable correlation might be appropriate.

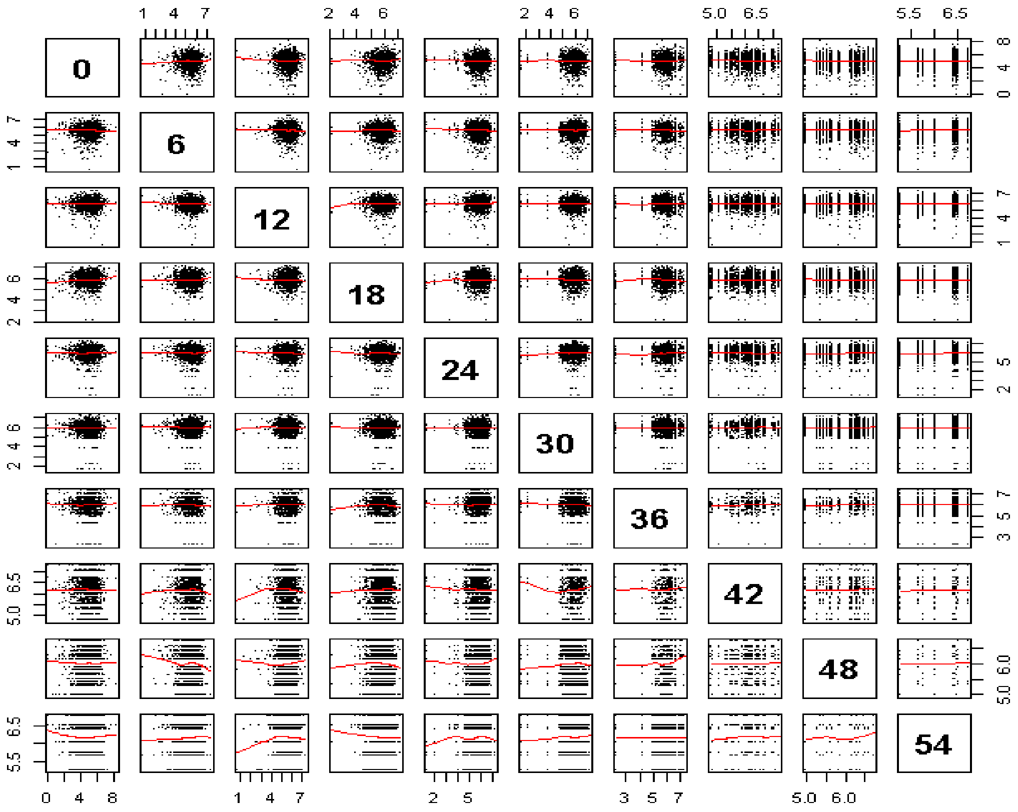

*Figure 5.6: pair-wise scatter plots matrix of CD4 count taken at Jimma University Specialized Hospital from 2007 –2011*

## 5.1.2 Modeling the Progression of CD4 count

### 5.1.2.1 Generalized Estimating Equations (GEE)

In this section the ART data is analyzed using the generalized estimating equation. For this purpose an exchangeable correlation structure is assumed based on the exploratory analysis result in figure 5.6. Though, it is not appropriate for data of this type, GEE with independence correlation assumption will be also considered for the sake of comparison.

To build the GEE model, first we assumed the model with all main effect and time by selected main effect interaction.

$$\log(\mu_{ij}) = \beta_0 + \beta_1 T_{ij} + \beta_2 S_i + \beta_3 A_i + \beta_4 Wt_i + \beta_5 w_{2i} + \beta_6 w_{3i} + \beta_7 w_{4i} + \beta_8 E_{1i} + \beta_9 E_{2i} + \beta_{10} E_{3i} + \beta_{11} F_{1i}$$
$$+ \beta_{12} F_{2i}\mathrm{i} + \beta_{13} \mathrm{T}^2{}_{ij} + \beta_{14} T_{ij} * S_i + \beta_{15} T_{ij} * W_{2i} + \beta_{16} T_{ij} * W_{3i} + \beta_{17} T_{ij} * W_{4i}$$
$$+ \beta_{18} T_{ij} * E_{1i} + \beta_{19} T_{ij} * E_{2i} + \beta_{20} T_{ij} * E_{3i} + \beta_{21} T_{ij} * F_{1i} + \beta_{22} T_{ij} * F_{2i}$$

Where E represents educational level, F= functional status and W= WHO stage, T=time, Wt= weight, S= sex and A= age.

To compare the two propose working correlation structures, first we considered naïve and robust standard error estimates for both correlation to see how it close to each other. As shown in table5, naïve and robust standard error estimates for exchangeable correlation is close to each other as compared to independence suggested that it is a good working correlation structure for the ART data. Secondly, we compared the two correlation structures using the Correlation Information Criterion (**CIC**) and we got 43.86 and 45.4 value for exchangeable and independence working correlation structures respectively. Thus, from both considerations, we got a model with exchangeable working correlation structure is preferred.

*Table 5.4: Comparison of exchangeable and independence working correlation structures using naïve and robust standard error estimates for the ART data taken at Jimma University Specialized Hospital from 2007 –2011*

| *Exchangeable working structure* | | | | *Independence working structure* | | |
|---|---|---|---|---|---|---|
| *Coeff.* | *Estimate* | *Naïve s.e* | *Robust s.e* | *estimate* | *Naïve s.e* | *Robust s.e* |
| $\beta_0$ | 5.486 | 0.092 | 0.098 | 5.494 | 0.069 | 0.099 |
| $\beta_1$ | 0.032 | 0.003 | 0.003 | 0.036 | 0.003 | 0.004 |
| $\beta_2$ | -0.167 | 0.036 | 0.037 | -0.163 | 0.031 | 0.037 |
| $\beta_3$ | -0.001 | 0.002 | 0.002 | -0.002 | 0.001 | 0.002 |
| $\beta_4$ | 0.003 | 0.001 | 0.002 | 0.003 | 0.001 | 0.002 |
| $\beta_5$ | -0.125 | 0.041 | 0.046 | -0.103 | 0.035 | 0.047 |
| $\beta_6$ | -0.156 | 0.045 | 0.051 | -0.141 | 0.039 | 0.051 |
| $\beta_7$ | -0.294 | 0.081 | 0.089 | -0.273 | 0.072 | 0.092 |
| $\beta_8$ | 0.064 | 0.045 | 0.048 | 0.080 | 0.038 | 0.048 |
| $\beta_9$ | -0.043 | 0.046 | 0.048 | -0.036 | 0.039 | 0.049 |
| $\beta_{10}$ | -0.185 | 0.065 | 0.064 | -0.127 | 0.056 | 0.063 |
| $\beta_{11}$ | -0.163 | 0.041 | 0.043 | -0.144 | 0.035 | 0.043 |
| $\beta_{12}$ | -0.090 | 0.088 | 0.090 | -0.103 | 0.076 | 0.093 |
| $\beta_{13}$ | -0.001 | 0.000 | 0.000 | -0.001 | 0.000 | 0.000 |
| $\beta_{14}$ | -0.001 | 0.002 | 0.002 | -0.001 | 0.002 | 0.002 |
| $\beta_{15}$ | 0.009 | 0.002 | 0.003 | 0.010 | 0.002 | 0.003 |
| $\beta_{16}$ | 0.007 | 0.002 | 0.003 | 0.009 | 0.002 | 0.003 |
| $\beta_{17}$ | 0.015 | 0.005 | 0.005 | 0.013 | 0.005 | 0.006 |
| $\beta_{18}$ | -0.003 | 0.002 | 0.004 | -0.006 | 0.002 | 0.003 |
| $\beta_{19}$ | 0.003 | 0.002 | 0.005 | -0.001 | 0.002 | 0.003 |
| $\beta_{20}$ | 0.012 | 0.003 | 0.006 | 0.003 | 0.003 | 0.004 |
| $\beta_{21}$ | 0.008 | 0.002 | 0.002 | 0.005 | 0.002 | 0.003 |
| $\beta_{22}$ | 0.003 | 0.004 | 0.008 | 0.000 | 0.004 | 0.005 |

Using the exchangeable working correlation structure, significant variables were selected using the generalized Wald test.

Table 5.5: Wald *test analysis table for all covariates of ART data taken at Jimma University Specialized Hospital from 2007 –2011*

| Coefficient | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | $\beta_5$ | $\beta_6$ | $\beta_7$ | $\beta_8$ | $\beta_9$ | $\beta_{10}$ | $\beta_{11}$ | $\beta_{12}$ | $\beta_{13}$ | $\beta_{14}$ | $\beta_{15}$ | $\beta_{16}$ | $\beta_{17}$ | $\beta_{18}$ | $\beta_{19}$ | $\beta_{20}$ | $\beta_{21}$ | $\beta_{22}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Estimate*($\hat{\beta}$) | *5.49* | *0.03* | *-0.2* | *-0.001* | *0.003* | *-0.13* | *-0.12* | *-0.3* | *0.1* | *-0.04* | *-0.2* | *-0.2* | *-0.1* | *-0.001* | *-0.001* | *0.01* | *0.01* | *0.02* | *-0.003* | *0.003* | *0.012* | *0.01* | *0.003* |
| *s.e*($\hat{\beta}$) | *0.098* | *0.007* | *0.037* | *0.002* | *0.002* | *0.046* | *0.051* | *0.089* | *0.048* | *0.048* | *0.064* | *0.043* | *0.09* | *0.0001* | *0.002* | *0.003* | *0.003* | *0.005* | *0.004* | *0.005* | *0.006* | *0.002* | *0.008* |
| *p-value* | *0.0* | *0.0* | *0.0* | *0.64* | *0.07* | *0.07* | *0.002* | *0.001* | *0.18* | *0.38* | *0.004* | *0.0* | *0.32* | *0.0* | *0.52* | *0.002* | *0.028* | *0.002* | *0.31* | *0.31* | *0.0* | *0.001* | *0.58* |

As shown in table 5.5 age, weight and time by sex interaction are not significant at 5%. Initially we avoiding non significant variables one by one starting from the most non significant variable then compared the two nested models using generalized Wald statistic. First remove age and refit the model finally compare the two models we got the p value 0.64. This largest p value implies there is no significant difference between the reduced and the full model. Hence the reduced model with small number of parameter is preferred. By following the same procedure we arrived at the final model given below.

$$\text{Log}\,(\mu_{ij}) = \beta_0+\beta_1 T_{ij}+\beta_2 S_i+\beta_5 w_{2i}+\beta_6 w_{3i}+\beta_7 w_{4i}+\beta_8 E_{1i}+\beta_9 E_{2i}+\beta_{10} E_{3i}+\beta_{11} F_{1i}+\beta_{12} F_{2i}$$

$$+\ \beta_{13} \mathrm{T}^2{}_{ij}+\beta_{15} T_{ij} * W_{2i}+\beta_{16} T_{ij} * W_{3i}+\beta_{17} T_{ij} * W_{4i}+\beta_{18} T_{ij} * E_{1i}$$

$$+\beta_{19} T_{ij} * E_{2i}+\beta_{20} T_{ij} * E_{3i}+\beta_{21} T_{ij} * F_{1i}+\beta_{22} T_{ij} * F_{2i}$$

*Table 5.6: Wald test estimates of the final GEE model for ART data taken at Jimma University specialized Hospital from January 2007 –2011*

| *Parameters* | *Estimate(95% conf.int)* | *Standard error estimates [Model based, Empirical]* | *p value* |
|---|---|---|---|
| $\beta_0$ | *5.598 (5.504,5.692)* | *[0.046, 0.048]* | *0.000* |
| $\beta_1$ | *0.031 (0.025,0.038)* | *[0.003, 0.007]* | *0.000* |
| $\beta_2$ | *-0.173 (-0.229,-0.118)* | *[0.028, 0.028]* | *0.000* |
| $\beta_5$ | *-0.129 (-0.219,-0.039)* | *[0.041, 0.046]* | *0.005* |
| $\beta_6$ | *-0.162 (-0.262,-0.062)* | *[0.045, 0.051]* | *0.001* |
| $\beta_7$ | *-0.304 (-0.478,-0.129)* | *[0.081, 0.089]* | *0.001* |
| $\beta_8$ | *0.075 (-0.018,0.168)* | *[0.044, 0.047]* | *0.113* |
| $\beta_9$ | *-0.029 (-0.119,0.062)* | *[0.045, 0.046]* | *0.534* |
| $\beta_{10}$ | *-0.166 (-0.288,-0.043)* | *[0.065, 0.062]* | *0.008* |
| $\beta_{11}$ | *-0.175 (-0.260,-0.089)* | *[0.041, 0.044]* | *0.000* |
| $\beta_{12}$ | *-0.099 (-0.277,0.078)* | *[0.088, 0.091]* | *0.273* |
| $\beta_{13}$ | *-0.001 (-0.001,0.000)* | *[0.000, 0.0004]* | *0.000* |
| $\beta_{15}$ | *0.009 (0.003,0.015)* | *[0.002, 0.003]* | *0.002* |
| $\beta_{16}$ | *0.007 (0.001,0.013)* | *[0.002, 0.003]* | *0.028* |
| $\beta_{17}$ | *0.015 (0.006,0.024)* | *[0.005, 0.0053]* | *0.002* |
| $\beta_{18}$ | *-0.003 (-0.008,0.003)* | *[0.002, 0.005]* | *0.323* |
| $\beta_{19}$ | *0.003 (-0.003,0.008)* | *[0.002, 0.003]* | *0.311* |
| $\beta_{20}$ | *0.011 (0.005,0.018)* | *[0.003, 0.0031]* | *0.000* |
| $\beta_{21}$ | *0.008 (0.004,0.013)* | *[0.002, 0.0021]* | *0.001* |
| $\beta_{22}$ | *0.002 (-0.007,0.011)* | *[0.004, 0.008]* | *0.595* |

*Exchangeable correlation=0.37*

In table7, the intercept ($e^{\beta 0} = e^{5.598}$=269.9) is an estimate of the mean CD4 count at base line (Time=0) for females which is significantly different from zero (p <0.000) given that they are working, illiterate and WHO stage I categories. Time ($e^{\beta 1}$= 1.03), implies the mean CD4 count increases 1.03 times per month among subjects in female group when the remaining variables kept constant. Similarly, the coefficient for sex $-0.173$ indicates, at sero-conversion the average CD4 count for males is $0.84$ times lower than females and their difference is highly significant(p<0.000) at 5%. In addition the negative coefficients for all WHO stages refers to at base line, their mean CD4 counts are significantly lower than the reference group(i.e, p=0.005,0.001,0.001for stage II,III,IV respectively). However, over time all the three groups have significantly better average CD4 count as compared to the reference group. For example, time by WHO stage II interaction ($\beta_{15} = 0.009$) entails the rate of increase in the CD4 count for subjects in WHO stage II category is estimated to be 1.01($e^{0.009}$) times per month higher than the rate of increase among stage I patients. Thus, the rate of change in mean CD4 count is estimated to be 0.89($e^{\beta_5+\beta_{15}} = e^{-0.129+0.009}$) counts per month among stage II patients. in the same way, the rate of increase for subjects in stage III and IV categories are 1.01 and 1.02 times per month higher than the rate of increase in stage I patients respectively or the rates of change in the average CD4 counts are 1.04 and 1.05 counts per month among stage III and IV patients correspondingly. Furthermore, at base line the mean CD4 count among ambulatory and bedridden patients are 0.84 and 0.91 times lower than the mean CD4 count among working patients (reference group) respectively; although over time the rate of increase for subjects among ambulatory and bedridden patients are 1.01 and 1.00 counts per month higher than the rate of increase among subjects in the working category but, there is no significant difference between bedridden and working category patients at base line and over time (p=0.273, 0.595). So that the rates of change in mean CD4 counts are estimated to be 0.84 and 0.91 counts per month among ambulatory and bedridden patients respectively.

#### 5.1.2.2 Generalized Linear Mixed Model (GLMM)

Based on the exploratory analysis result of the mean structure, first we assume the model with the main effect and Time by selected main effect interaction and incorporating intercept, time and time$^2$ as random effects.

Log ($\mu_{ij}$) = $X_{ij}\beta$+$Z_{ij}b_i$, $\mu_{ij} = e^{X_{ij}\beta+Z_{ij}b_i}$, Where X and Z are matrix of fixed and random effect covariates as well as $\beta$ and b are corresponding coefficients respectively.

$$\eta_{ij} = \beta_0+\beta_1 T_{ij}+\beta_2 S_i+\beta_3 A_i+\beta_4 Wt_i+\beta_5 w_{2i}+\beta_6 w_{3i}+\beta_7 w_{4i}+\beta_8 E_{1i}+\beta_9 E_{2i}+\beta_{10} E_{3i}+\beta_{11} F_{1i}$$
$$+\beta_{12} F_{2i}\mathrm{i}+\beta_{13}\mathrm{T}^2{}_{ij}+\beta_{14} T_{ij}*S_i+\beta_{15} T_{ij}*W_{2i}+\beta_{16} T_{ij}*W_{3i}+\beta_{17} T_{ij}*W_{4i}+\beta_{18} T_{ij}*E_{1i}$$
$$+\beta_{19} T_{ij}*E_{2i}+\beta_{20} T_{ij}*E_{3i}+\beta_{21} T_{ij}*F_{1i}+\beta_{22} T_{ij}*F_{2i}+b_{0i}+b_{1i} T_{ij}+b_{2i}\mathrm{T}^2{}_{ij}$$

*Table 5.7: A model with all covariate with the corresponding estimates for the ART data taken at Jimma University specialized Hospital from 2007 –2011*

| Coefficients | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | $\beta_5$ | $\beta_6$ | $\beta_7$ | $\beta_8$ | $\beta_9$ | $\beta_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Estimate*($\hat{\beta}$) | *4.98* | *0.09* | *-0.09* | *-0.01* | *0.003* | *0.08* | *-0.5* | *-0.6* | *0.4* | *0.5* | *0.2* |
| *s.e*($\hat{\beta}$) | *0.06* | *0.006* | *0.023* | *0.001* | *0.001* | *0.029* | *0.032* | *0.053* | *0.037* | *0.037* | *0.044* |
| *p-value* | *0.0* | *0.0* | *0.0* | *0.0* | *0.002* | *0.007* | *0.0* | *0.0* | *0.0* | *0.0* | *0.0* |

| $\beta_{11}$ | $\beta_{12}$ | $\beta_{13}$ | $\beta_{14}$ | $\beta_{15}$ | $\beta_{16}$ | $\beta_{17}$ | $\beta_{18}$ | $\beta_{19}$ | $\beta_{20}$ | $\beta_{21}$ | $\beta_{22}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *-0.13* | *0.8* | *-0.003* | *-0.001* | *-0.02* | *0.02* | *0.04* | *-0.02* | *-0.01* | *-0.001* | *0.01* | *0.004* |
| *0.026* | *0.072* | *0.0001* | *0.002* | *0.003* | *0.003* | *0.005* | *0.004* | *0.004* | *0.004* | *0.002* | *0.007* |
| *0.0* | *0.0* | *0.0* | *0.656* | *0.0* | *0.0* | *0.0* | *0.0* | *0.06* | *0.84* | *0.0* | *0.559* |

In longitudinal data analysis, what random effect should be included to the model in order to account between individual variability is a critical issue. The above model considered all intercept, linear and quadratic time effect as random part. Now let us compared the models by removing each random effect one by one using AIC followed by likelihood ratio test to choose the best random effects that enables to account between individual variability or to fit the ART data well.

*Table 5.8: Comparison of models with different random effects using AIC and likelihood ratio test for the ART data taken at Jimma University specialized Hospital from 2007 –2011*

| | *intercept* | *Intercept+time* | *Intercept+time+time*$^2$ |
|---|---|---|---|
| *AIC* | *322845* | *236384* | *87033* |
| *BIC* | *323002* | *236541* | *87222* |
| *LokLik* | *-163002* | *-118168* | *-43487* |
| $p_{1,2} < 0.000$ *and* $p_{2,3} < 0.000$ | | | |

$p_{1,2}$*: is the p value of likelihood ratio test for comparison of a model with intercept and Intercept+time.* $p_{2,3}$*: is the p value of likelihood ratio test for comparison of a model with Intercept+time and Intercept+time+time*$^2$

In table 5.8, the AIC value is decreasing from 322845 to 87033 which show the model with intercept, time and quadratic time effect is highly improved as compared to the others. This result is confirmed by the likelihood ratio test (p <0.000). We also considered a model without random effects (i.e simply the generalized linear model) and got AIC value 571150 which is too large as compared to the above three models. Thus a model with Intercept+time+time$^2$ as random effect is best for ART data.

After selecting the appropriate random effects, we assessed the significance of the fixed effects. As given in the previous table (table 5.7), time by sex interaction is not significant (p=*0.656*) at 5%. Therefore we removed it from the model and compared the reduced and full model using AIC followed by likelihood ratio test and we obtained AIC value decreases from 87033 to 87024 due to the removal of time by sex interaction (p=0.99). Thus, the reduced model with less number of parameter is preferred

*Table 5.9: Comparison of models with and without sex by time interaction using AIC and likelihood ratio test for the ART data taken at Jimma University specialized Hospital from 2007 –2011*

| | Model with time:sex | Model without time:sex |
|---|---|---|
| *AIC* | *87033* | *87024* |
| *BIC* | *87222* | *87207* |
| *LogLik* | *-43487* | *-43484* |
| *P =0.99* | | |

*P: value of likelihood ratio test for a model comparison with and without sex by time interaction*

The final model for generalized linear mixed model is given below.

$$\eta_{ij} = \beta_0+\beta_1 T_{ij}+\beta_2 S_i+\beta_3 A_i+\beta_4 Wt_i+\beta_5 w_{2i}+\beta_6 w_{3i}+\beta_7 w_{4i}+\beta_8 E_{1i}+\beta_9 E_{2i}+\beta_{10} E_{3i}+\beta_{11} F_{1i}$$
$$+\beta_{12} F_{2i}\mathrm{i}+\beta_{13} \mathrm{T}^2{}_{ij}+\beta_{15} T_{ij} * W_{2i}+\beta_{16} T_{ij} * W_{3i}+\beta_{17} T_{ij} * W_{4i}+\beta_{18} T_{ij} * E_{1i}$$
$$+\beta_{19} T_{ij} * E_{2i}+\beta_{20} T_{ij} * E_{3i}+\beta_{21} T_{ij} * F_{1i}+\beta_{22} T_{ij} * F_{2i}+b_{0i}+b_{1i} T_{ij}+b_{2i} \mathrm{T}^2{}_{ij}$$

*Table 5.10: Wald test estimates of the final GLMM model for ART data taken at Jimma University Specialized Hospital from 2007 –2011*

| *Covariates* | *Estimates($\hat{\beta}$)* | *Std. error* | *P values* | *95%coenf.int* | | *Random effects variance estimates* | |
|---|---|---|---|---|---|---|---|
| | | | | *Lower* | *upper* | | |
| $\beta_0$ | *4.961* | *0.040* | *0.000* | *4.844* | *5.079* | *G00* | *0.82* |
| $\beta_1$ | *0.096* | *0.006* | *0.000* | *0.085* | *0.107* | *G11* | *0.029* |
| $\beta_2$ | *-0.088* | *0.017* | *0.000* | *-0.122* | *-0.054* | *G22* | *0.0001* |
| $\beta_3$ | *-0.006* | *0.001* | *0.000* | *-0.009* | *-0.004* | *G01* | *-0.93* |
| $\beta_4$ | *0.003* | *0.001* | *0.007* | *0.001* | *0.005* | *GO2* | *0.003* |
| $\beta_5$ | *0.083* | *0.029* | *0.004* | *0.027* | *0.140* | *G12* | *-0.002* |
| $\beta_6$ | *-0.508* | *0.032* | *0.000* | *-0.570* | *-0.445* | *Correlation* | |
| $\beta_7$ | *-0.663* | *0.054* | *0.000* | *-0.768* | *-0.558* | $\rho_{01}$ | *-0.6* |
| $\beta_8$ | *0.415* | *0.037* | *0.000* | *0.343* | *0.487* | $\rho_{02}$ | *0.038* |
| $\beta_9$ | *0.511* | *0.036* | *0.000* | *0.440* | *0.583* | $\rho_{12}$ | *-0.92* |
| $\beta_{10}$ | *0.198* | *0.044* | *0.000* | *0.112* | *0.283* | | |
| $\beta_{11}$ | *-0.127* | *0.025* | *0.000* | *-0.177* | *-0.077* | $\delta^2$ | *1.432* |
| $\beta_{12}$ | *0.858* | *0.072* | *0.000* | *0.717* | *0.999* | | |
| $\beta_{13}$ | *-0.003* | *0.0003* | *0.000* | *-0.003* | *-0.002* | | |
| $\beta_{15}$ | *-0.021* | *0.003* | *0.000* | *-0.026* | *-0.015* | | |
| $\beta_{16}$ | *0.020* | *0.003* | *0.000* | *0.014* | *0.026* | | |
| $\beta_{17}$ | *0.045* | *0.005* | *0.000* | *0.036* | *0.054* | | |
| $\beta_{18}$ | *-0.023* | *0.004* | *0.000* | *-0.031* | *-0.016* | | |
| $\beta_{19}$ | *-0.012* | *0.003* | *0.000* | *-0.018* | *-0.005* | | |
| $\beta_{2o}$ | *-0.005* | *0.003* | *0.122* | *-0.012* | *0.001* | | |
| $\beta_{21}$ | *0.013* | *0.002* | *0.000* | *0.009* | *0.017* | | |
| $\beta_{22}$ | *0.002* | *0.007* | *0.774* | *-0.012* | *0.016* | | |

From table 5.10, there are two estimated variance components; these are the random effects variances and the residual variance. The residual variance is var($\varepsilon_{ij}$)= $\delta^2$=1.432 and for the random effects, var(bo)= Goo=0.82,var(b1)=G11=0.029 and var(b2)=G22=0.0001, similarly, cov(b0,b1)=G01= -0.93, cov(bo,b2)=0.003, cov(b1,b2)=-0.002. Under the assumption of normally distributed random effects $\sqrt{Goo}$ =0.91, implies that 95% of female individuals have a mean CD4 count at seroconversion between $e^{4.961-0.91*1.96}$ = 24 and $e^{4.961+0.91*1.96}$ =849.5. We emphasize that this interval is for each female individual values of the CD4 count at baseline rather than for individual measurements at baseline. The interval (24, 849.5) does not include the measurement variation attributable to the residuals. Similarly, there is a discernible heterogeneity in the patient to patient changes in the CD4 count rates. The total variability between individuals is estimated as Goo+ G11+G22= 0.8491 whereas the total variability within individual is 1.432. However, the total variation in CD4 count is estimated to be1.432+0.8491= 2.2811. The proportion of total variability that is attributed to within person variation is given by 1.432/2.2811 is 62.8% while the proportion of total variability attributed to between individual variations in their general level of CD4 count is 0.8491/2.2811is 37.2%. Therefore more than half of the variation is explained by the residuals.

The correlation $\rho_{01}$= -0.6 indicates, there is a negative correlation between intercept and slope of linear time effect for the random part (i.e when patient's intercept increase by one unit of standard deviation, their slope would decrease by 0.6 standard deviations) whilst $\rho_{02}$= 0.38 implies there is a weak positive correlation between the intercept and slopes for quadratic time effect (i.e when patient's intercept increase by one unit of standard deviation, their quadratic time effects slope would increase by 0.38 standard deviations). In the same way, $\rho_{12}$= *-0.*92 which shows there is a strong negative correlation between linear and quadratic time effect.

All the fixed effects parameters in GLMM have subject specific interpretation unlike marginal model. Thus, given the random effects($b_i$); the intercept ( $e^{\beta 0} = e^{4.961}$=142.7) in GLMM is an estimate of the "i$^{th}$ " female subject average CD4 count provided that she is working, illiterate and WHO stage I categories. Similarly, Time ($e^{\beta 1}$= 1.06), implies the mean CD4 count increases 1.06 times per month for the "i$^{th}$" female individual when the remaining variables kept constant and it is significantly different from zero (p<0.000) at 5%. In addition the coefficient for sex ($\beta_2$=-*0.088*) verifies that the mean CD4 count for "i$^{th}$" male individual is 0.92($e^{-0.088}$) times lower than female individual with the same random effects ($b_i$) at base line and their difference is highly significant (p-value<0.000) at 5%. For the sake of comparing the results of GLMM to GEE, let us consider, time by WHO stage II interaction ($\beta_{15} = -0.021$) indicates, there is a significant difference(p<0.000) between stage I and stage II patients in terms of subject-specific changes in the expected rates of CD4 count. In particular, rate of increase in the mean CD4 count for "$i^{th}$ " subject in WHO stage II category is estimated to be 0.98 times per month lower than the rate of increase for any stage I category patient with the same random effects($b_i$). Thus, the rate of change in mean CD4 count for the "$i^{th}$ " individual in stage II category is estimated to be 1.06($e^{\beta_5+\beta_{15}} = e^{0.083-0.021}$) counts per month. Other parameters are interpreted in the same way. Generally, except the interaction of time by secondary educational level and time by bedridden functional status, all other coefficients are highly significant at 5%. It means there is a significant difference between the two comparable groups.

#### 5.1.2.2.1 Model diagnostic for GLMM

Residual versus fitted value plot for final GLMM model is presented in figure1 **(**Annex I**)**, it does not show any systematic pattern this point out the model fits the data well and the Q-Q plot (Annex I) also verifies the residual are normally distributed and symmetric around zero. Thus, it meets the assumption of error term. Besides to the above, the non linearity of the Q-Q plot confirms the model is not linear. Furthermore, the residual versus each categorical predictor recommended that there is a uniformity of residuals across each level of covariates specifies that homogeneity of error variances. Residuals versus observation ID number plot (figure4, Annex I) also suggested the residuals are symmetric around zero (i.e. positive and negative residuals are almost equal) and there is no outlier. Plots of observed versus fitted value of CD4 count is given in figure5 (Annex I), it verifies that there is a close agreement between observed and fitted values suggested that this model is good in predicting CD4 count. Q-Q plots for normality of random effects are also given in figure6 and figure7 of the same annex; which illustrates the random effects are normally distributed with mean zero and variance covariance matrix G. Thus, the fitted GLMM model is fine for ART data.

### 5.1.2.3 Comparison of GLMM and GEE

Even though the parameter estimates in GLMM and GEE have different interpretations (i.e GLMM provides subject-specific parameter estimates where as GEE only estimate population average regression coefficients), we can compare the two models using their respective standard error estimates (ratio). For the sake of comparison, the study did not use the outputs of respective final models directly. This is because non-significant covariates were removed from GEE final model so that it is impossible to compare two models having different number of covariates. Thus, we considered all covariates for both models and the result for both models is presented in *Table 5.11*. The ratio of the standard error estimates of GLMM to GEE is between 0.33 and 1; it means the standard error estimates of GLMM are smaller than that of GEE, except five covariates having approximately equal estimates. In other words the GLMM fits the data with small disturbance than GEE. Therefore, GLMM model is better than GEE provided that GEE has marginal interpretation and GLMM has conditional interpretation, conditionally upon level of random effects. In addition, from the same table we saw that some of the fixed effect coefficients ($\beta_5, \beta_9, \beta_{10}, \beta_{12}, \beta_{15}$, $\beta_{19}$, $\beta_{20}$) have opposite signs for the two models, this result is due to subject specific and population average interpretation for the two models. Moreover, age, weight and sex by time interaction are not significant (p=0.64, 0.07, 0.52 respectively) for GEE model whereas in GLMM, only sex by time interaction is not significant (p=0.66) at 5%. Similarly most covariates p values are higher in GEE as compared to GLMM.

*Table 5.11: Comparison of GEE and GLMM using their standard error estimate ratio for ART data taken at Jimma University Specialized Hospital from 2007 –2011*

| | GEE | | | GLMM | | | Ratio |
|---|---|---|---|---|---|---|---|
| | Estimate | Std.err | P value | Estimate | Std.err | P value | GLMM/gee |
| $\beta_0$ | 5.486 | 0.098 | 0.000 | 4.981 | 0.060 | 0.000 | 0.60 |
| $\beta_1$ | 0.032 | 0.007 | 0.000 | 0.092 | 0.006 | 0.000 | 0.86 |
| $\beta_2$ | -0.167 | 0.037 | 0.000 | -0.089 | 0.023 | 0.000 | 0.62 |
| $\beta_3$ | -0.001 | 0.002 | 0.639 | -0.006 | 0.001 | 0.000 | 0.5 |
| $\beta_4$ | 0.003 | 0.002 | 0.074 | 0.003 | 0.001 | 0.002 | 0.5 |
| $\beta_5$ | -0.126 | 0.046 | 0.007 | 0.078 | 0.029 | 0.007 | 0.63 |
| $\beta_6$ | -0.156 | 0.051 | 0.002 | -0.49 | 0.032 | 0.000 | 0.63 |
| $\beta_7$ | -0.294 | 0.089 | 0.001 | -0.626 | 0.053 | 0.000 | 0.60 |
| $\beta_8$ | 0.064 | 0.048 | 0.179 | 0.375 | 0.037 | 0.000 | 0.77 |
| $\beta_9$ | -0.043 | 0.048 | 0.375 | 0.461 | 0.037 | 0.000 | 0.77 |
| $\beta_{10}$ | -0.185 | 0.064 | 0.004 | 0.156 | 0.044 | 0.000 | 0.69 |
| $\beta_{11}$ | -0.163 | 0.043 | 0.000 | -0.132 | 0.026 | 0.000 | 0.60 |
| $\beta_{12}$ | -0.090 | 0.090 | 0.321 | 0.777 | 0.072 | 0.000 | 0.80 |
| $\beta_{13}$ | -0.001 | 0.000 | 0.000 | -0.003 | 0.000 | 0.000 | 0.33 |
| $\beta_{14}$ | -0.001 | 0.002 | 0.522 | -0.001 | 0.002 | 0.656 | 1 |
| $\beta_{15}$ | 0.009 | 0.003 | 0.002 | -0.019 | 0.003 | 0.000 | 1 |
| $\beta_{16}$ | 0.007 | 0.003 | 0.028 | 0.020 | 0.003 | 0.000 | 1 |
| $\beta_{17}$ | 0.015 | 0.005 | 0.002 | 0.043 | 0.005 | 0.000 | 1 |
| $\beta_{18}$ | -0.003 | 0.004 | 0.314 | -0.020 | 0.004 | 0.000 | 0.97 |
| $\beta_{19}$ | 0.003 | 0.005 | 0.305 | -0.007 | 0.004 | 0.055 | 0.80 |
| $\beta_{20}$ | 0.012 | 0.006 | 0.000 | -0.001 | 0.004 | 0.843 | 0.67 |
| $\beta_{21}$ | 0.008 | 0.002 | 0.001 | 0.013 | 0.002 | 0.000 | 1 |
| $\beta_{22}$ | 0.003 | 0.008 | 0.582 | 0.004 | 0.007 | 0.559 | 0.88 |

## 5.2 Discussion

This study was aiming at modeling the progression of HIV infection using longitudinally measured CD4 count for HIV positive patients initiated to ART. Since the data is correlated and discrete (count), for such types of data two generalized models (GEE and GLMM) were applied. The ART data was analyzed using different plots (exploratory analysis) followed by model based outputs. From individuals profile plot, we observed the existence of variability in CD4 count within and between individuals. The exploratory analysis result for mean structure also suggested that on average, CD4 count increases in a quadratic pattern over time. This supports the results of Moing *et al.* (2002) who found that after the patients initiated to the ART program their CD4 count increases due to the therapy. It is indirectly associated with the decrease in the progression of HIV infection as the immune system and HIV infection are negatively correlated or simply it delays disease progression(time to AIDS). Additionally, the mean CD4 count for female is higher than male up to around 42 months however it is not significant over time. In general, the exploratory analysis of the mean structure supported the findings of both Bayeh *et al*. (2010) and Shiras. R.P (2006) who said the progression of CD4 count increase at high rate after patients initiated to ART.

Two propose working correlation structures, where exchangeable correlation is selected based on exploratory analysis result and independence correlation simply taken for the sake of comparison, for GEE model were compared in this paper and found that exchangeable working correlation structure fits the ART data better than independence. On the other hand, in GLMM a model with only intercept, intercept + time and intercept+ time + time$^2$ were compared for the purpose of selecting the best random effect that enable to account the variability between individuals. The three models were compared using the AIC value followed by likelihood ratio test and we got a model with intercept + time + time$^2$ as random effect is the best.

From the final model results of GEE; age, weight and time by sex interactions (p=0.64, 0.07, 0.52 respectively) are not significant predictors of CD4 count progression at 5% whereas in GLMM, only time by sex interaction is not significant (p=0.66). In this regard, more or less the findings from GLMM supports the previous findings of Ville's et al. (2007) who found that different characteristics of the individuals like age, base line clinical stage, plasma HIV RNA and time since HIV diagnosis are recognized predictors of progression of HIV infections. In addition, Bayeh et al. (2009) identified age, sex, educational levels and income levels of HIV patients as predictors of CD4 count progression after initiated to ART.

According to both models result, the average CD4 count is significantly lower in males than females at base line as suggested by exploratory analysis though there is no significant difference over time. In the same way, Moing *et al.* (2002) got males and females have no significant difference in their CD4 count progression over time. In addition, the evolution of CD4 count is differing by patients' base line educational levels, functional status and clinical stages. From, final GLMM output, the proportion of total variability that is attributed to the within person variation is 68.2% whereas the remaining 37.8% is attributed to the between person variation.
This study also compared the two models using their standard error estimates ratio (Molenberghs, 2008) and obtained GLMM fits the ART data better than GEE with a small disturbance provided that parameters in GLMM have subject specific interpretation while GEE have population average interpretation. Due to the difference in their parameter interpretation, some of the fixed effect coefficients ($\beta_5, \beta_9, \beta_{10}, \beta_{12}, \beta_{15}$, $\beta_{19}$, $\beta_{20}$) have opposite signs for the two models which support the findings of Renard (2002) and Fu(2010). Furthermore, when we compare the corresponding p values, most covariates are highly significant for GLMM than GEE at 5%.

# CHAPER SIX

# 6. CONCLUSSIONS AND RECOMMENDATIONS

## 6.1 CONCLUSSIONS

This study evaluated the association between the progressions of HIV infection using longitudinally measured CD4 count and its possible predictors via longitudinal analysis methodologies. Statistically two modeling approaches (GEE and GLMM) have been compared for the analysis of ART data and we obtained GLMM exhibited the best fit for this data with small disturbance than GEE.

The study also found that except time by sex interaction, all other candidate covariates were identified as significant predictors for the progression of CD4 count (HIV infection). Moreover, on average CD4 count increases in a quadratic pattern over time after patients initiated to ART program (i.e the immune system increases where as the progression of the disease turn down due to the therapy).

We also concluded that the mean CD4 count for male is significantly lower than female at base line. Finally, we end up with a conclusion that the evolution of CD4 count (HIV infection) is differing by patient's baseline demographic and clinical characteristics like sex, WHO stages, educational levels and functional status.

## 6.2 RECOMMENDATIONS

Being HIV infection is the most serious disease in the world, modeling the progression of this disease helps to identify the factors that affect the success of the therapy which helps to discover new vaccine or drug by considering the identified factors. Thus further studies should be done in the area using these newly developed and most flexible methodologies by including additional covariates like regimen and viral load of patients which predict the evolution of HIV infection in a better way.

Furthermore, though the choice between GEE and GLMM for longitudinal data can only be made on subject matter grounds, using generalized linear mixed model is much emphasized than generalized estimating equations for correlated data as GEE can only handle the within subject variations through the assigned working correlation structure where as GLMM in addition to within measurement variation, between individual variations can be accounted by incorporating the random effects. Due to that, GLMM fits a given data with a small disturbance than GEE.

## References

AbdoolKarim, S. & AbdoolKarim, Q.(2005).HIV/AIDS in South Africa, Cambridge University Press.

Anderson, R. & May, R.(1991).Infectious diseases of human dynamics and control, Oxford University Press.

Andrew, R., Judith,A., Perry, S., Edwin, D., Guzman, D., David, R. Bangsberg Reviewed work (2004).Adherence to Highly Active Antiretroviral Therapy in the Homeless Population in San Francisco. A Prospective Study, *Clinical Infectious Diseases, Vol. 39, No. 8, pp. 1190-1198*

BAI, Y.(2011).Longitudinal Analysis to Assess the Impact of Method of Delivery on Postpartum Outcomes. *The Ontario Mother and Infant Study (TOMIS) III*

Binquet, G., Chêne, V., Journot, M., Savès, D. Lacoste, F. and the Grouped Epidemiologic Clinique du SIDA en Aquitaine, Modeling Changes in CD4-positive T-Lymphocyte Counts after the Start of Highly Active Antiretroviral Therapy and the Relation with Risk of Opportunistic Infections. *American Journal of Epidemiology, Vol. 153, No.4*

Breslow, N. and Clayton, D.(1993).Approximate inference in generalized linear mixed models. *J. Am. Statist.Ass*., *Vol.*88, pp. 9-25.

Brookmeyer, R., and Goedert, J.(1989).Censoring in an Epidemic with an Application to Hemophilia-Associated AIDS. *Biometrics, Vol. 45,pp. 325-335*.

Carlo, B. and Claudia, A. Reviewed work (2004). Effectiveness of Potent Antiretroviral Therapy on Progression of Human Immunodeficiency Virus: Bayesian Modeling and Model Checking via Counter factual. *Journal of the Royal Statistical Society Series C (Applied Statistics), Vol. 53, No. 4 pp. 633-650*

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

Eyster, E., Gail, M. and Ballard, J.(1987).Natural History of Human Immunodeficiency Virus Infection in Hemophiliacs. Effects of T-Cell Subsets, Platelet Counts, and Age, Annals of Internal Medicine *Vol.*107,pp. 1-6.

Fu, S.(2010). Regression Approaches to Estimation of Relative Risk: Application to Multiple Sclerosis Studies. University of British Columbia

Goldman, I. (1996).Response of CD4 Lymphocytes and Clinical Consequence of treatment using ddI or ddC In Patients with Advanced HIV infection. *Journal of Acquired Immune Deficiency Syndromes and Human Retro virology Voll.11, No.2*

Hin, L. and Wang, Y. (2009).Working-correlation-structure identification in generalized estimating equations. *Statistics in Medicine Vol.*28, *No. 4, pp.* 642-658.

Jose, C. and Douglas, M.(2000).Mixed Effect Models in S and S-PLUS New York. Springer Verlag.pp.100-200

Joy Rogers (2006). Change point Analysis of HIV Marker Responses.*pp.1-35*

Lee, Y. and Nelder, J.(1996). Hierarchical generalized linear models. *J. R: Statist. Soc. B, Vol. 58, pp. 619-678.*

Littell, R., Henry, P. & Ammerman, C.(1998).Statistical analysis of repeated measures data using SAS procedures. *J Anim Sci, Vol.76, pp. 1216-31.*

Lyle, C., *Katrina*, J. and Martin, L.(2005).Tutorial in biostatistics.Spline smoothing with linear mixed model. *Statist: Med. Vol. 24, pp. 3361–3381*

Magder, L. and Zeger, S.(1996).Smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians. *J. Am: Statist. Ass, Vol. 91, pp. 1141-1151.*

McCullagh, P. and Nelder, J.(1989).Generalized Linear Models. 2ndedition. London: Chapman and Hall. Pp.150-300

Mcculloch, C., Searle, S. & Neuhaus, J.(2008). Generalized, Linear, and Mixed Models. Hoboken

Mem , I. , Rio,  D.( 1997) Vol. 92, No.4, pp. 300-350

Nan, M., James, H.(1982). Random-Effects Models for Longitudinal data. *Biometrics, Vol.38, No. 4, pp. 963-974.*

Nicholas, L., Bradley, P. and Alan, E. Reviewed work (1992).Hierarchical Baye's Models for the Progression of HIV Infection Using Longitudinal CD4 T-Cell Numbers.  *Journal of the American Statistical Association, Vol. 87, No. 419, pp. 615-626.*

Palella, J., Delaney, K., Moorman, A., Loveless, M., Fuhrer, J., Satten, G., Aschman, D. J and Holmberg, S. (1998).Declining morbidity and mortality among patients with advanced Human Immunodeficiency Virus infection.HIV Outpatient Study Investigators. New Engl. J Med., *Vol.338, pp.* 853-860.

Rebecca, V., Culshaw, S.(2006).Mathematical Modeling of AIDS Progression, Limitations, Expectations, and Future Directions.*Journal of American Physicians and Surgeons Volume 11, Number 4.*

Renard, D. (2002). Topics in Modelling Multilevel and Longitudinal Data, Diepenbeek

Ve, V. Moing,L., Thiébaut,T., Chêne,G., Leport,C., Cailleton,V. Michelet,C., Fleury, H., Herson,S., Raffi,F.-Reviewed work(2002). Predictors of Long-Term Increase in CD$^+$ Cell Counts in Human Immunodeficiency Virus-Infected Patients Receiving a Protease Inhibitor Containing Antiretroviral Regimen. *The Journal of Infectious Diseases, Vol. 185, No. 4, pp. 471-480*

Verbeke, G. & Molenberghs, G.(2000).Linear Mixed Models for Longitudinal Data.New York, Springer-Verlag. No.2, pp.31-40

Verbeke, V. and Lesaffre, L.(1996). Linear Mixed-Effects Model with Heterogeneity in the Random-Effects Population. *Journal of the American Statistical Association, Vol. 91, No. 433 pp. 217-221*

Victor, D., Nicholas, L., Urania, D. Reviewed work (1991).Modeling the Progression of HIV Infection. *Journal of the American Statistical Association, Vol. 86, No. 415, pp. 569-577.*

Villes, V., Bruno, S. chareletelewden, C.(2007).The effects of depressive symptoms ART initiation on HIV clinical progression and mortality implications in clinical practice. *Antiviral Therapy, Vol.12, pp.1067-1074*

Wang, Y. and Hin, L.(2010). Modeling strategies in longitudinal data analysis: Covariate, variance function and correlation structure selection. *Computational Statistics & Data Analysis, Vol. **54**, pp. 3359-3370.*

Wang, Y. and Jeremy M. Taylor Reviewed work (2001).Jointly Modeling Longitudinal and Event Time Data with Application to Acquired Immunodeficiency Syndrome. *Journal of the American Statistical Association, Vol. 96, No. 455, pp. 895-905*

Wolfinger, R., and O'Connell, M.(1993).Generalized linear mixed models: Pseudo likelihood approach. J. Statist.*Comput. Simul, Vol.48, pp. 233-243.*

World Health Organization.(1990).Interim proposal for WHO Staging System for HIV infection and disease. *Wkly Epidemiol Rec, Vol. 65, pp. 221-4*.

Zeger, S. and Karim, M.(1991).Generalized linear models with random effects: Gibbs sampling approach. *J. Am. Statist.Ass., Vol.82, pp. 231-238.*

Zeger, S. & Liang K.(1986).Longitudinal data analysis using generalized linear models. *Biometrics, Vol.73, pp. 13-22.*

Zhang, D. and Lin, X.(2003). Hypothesis testing in semi-parametric additive mixed models. *Biostatistics, Vol.4, pp. 57–74.*

# ANNEX I

## Model diagnostics for GLMM model

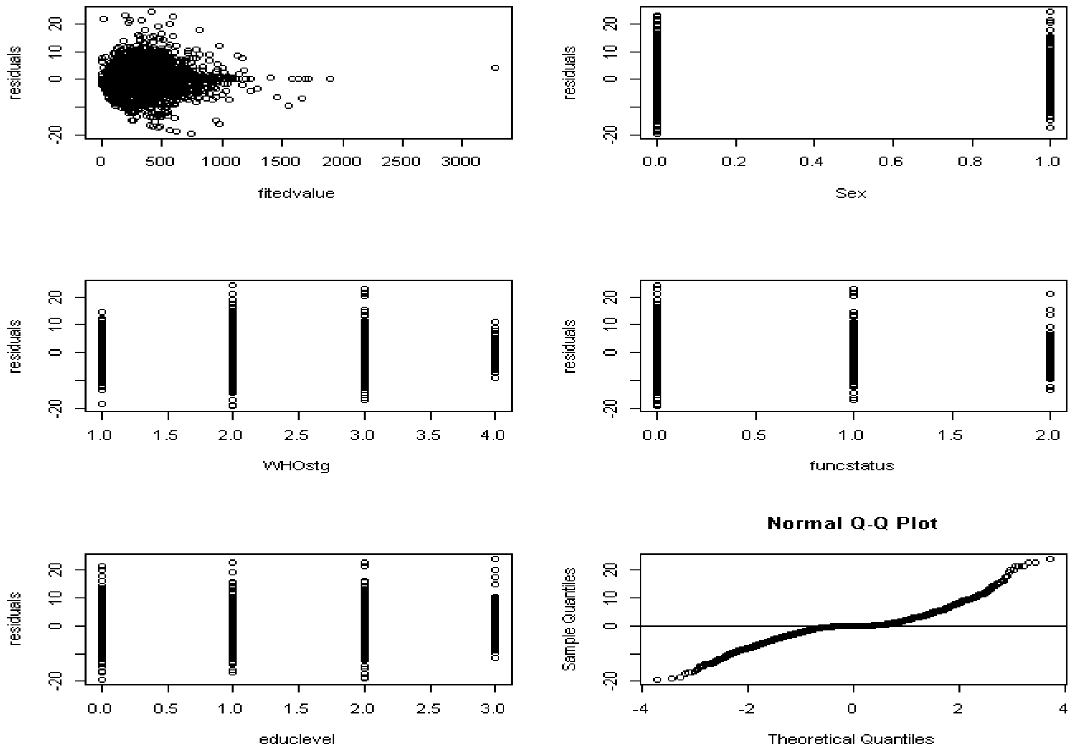

*Figure1: different plots of residuals*

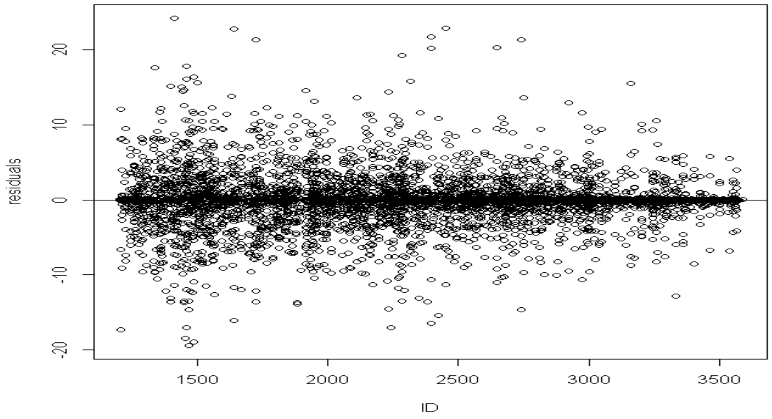

*Figure 2: plots of residuals versus observation id number*

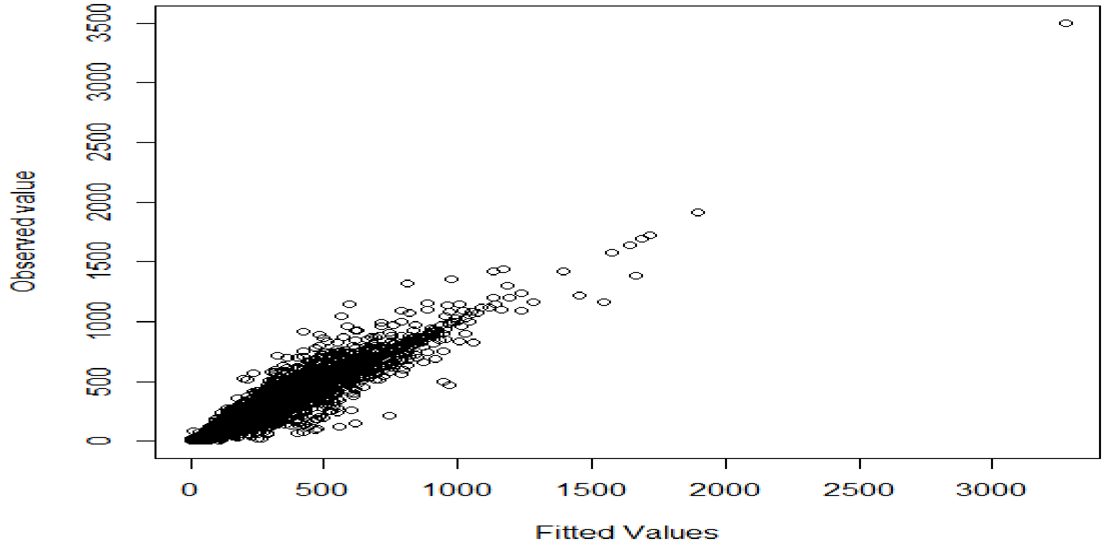

*Figure3: plots of observed versus fitted value for the ART data*

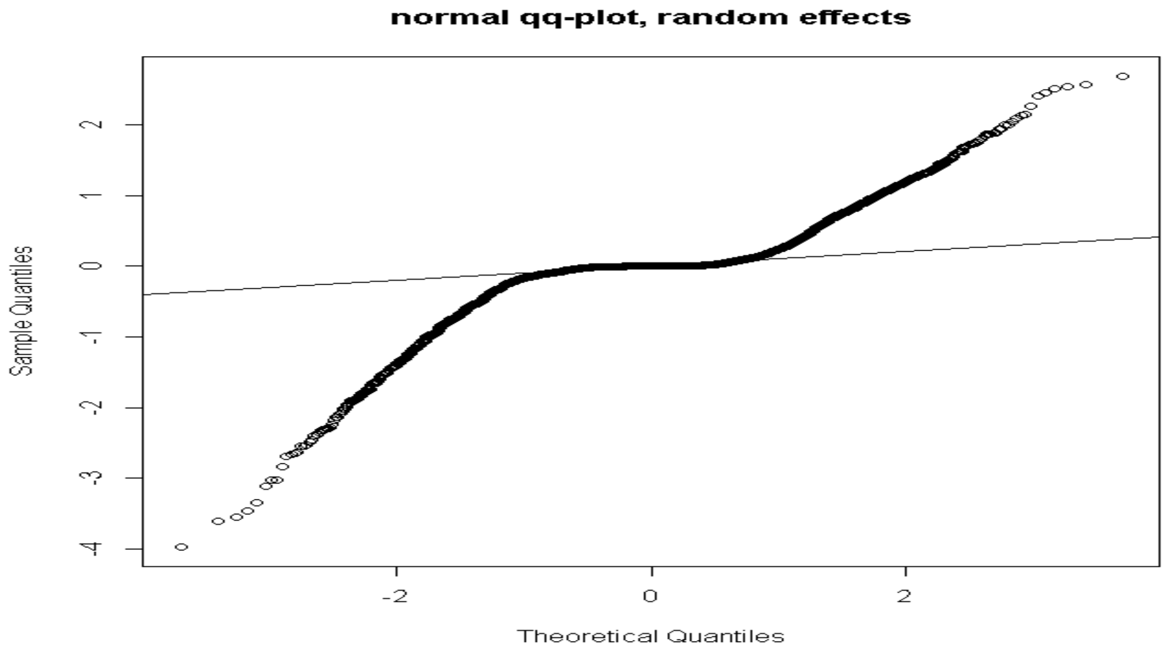

*Figure4: normal qq plots of the random effects*

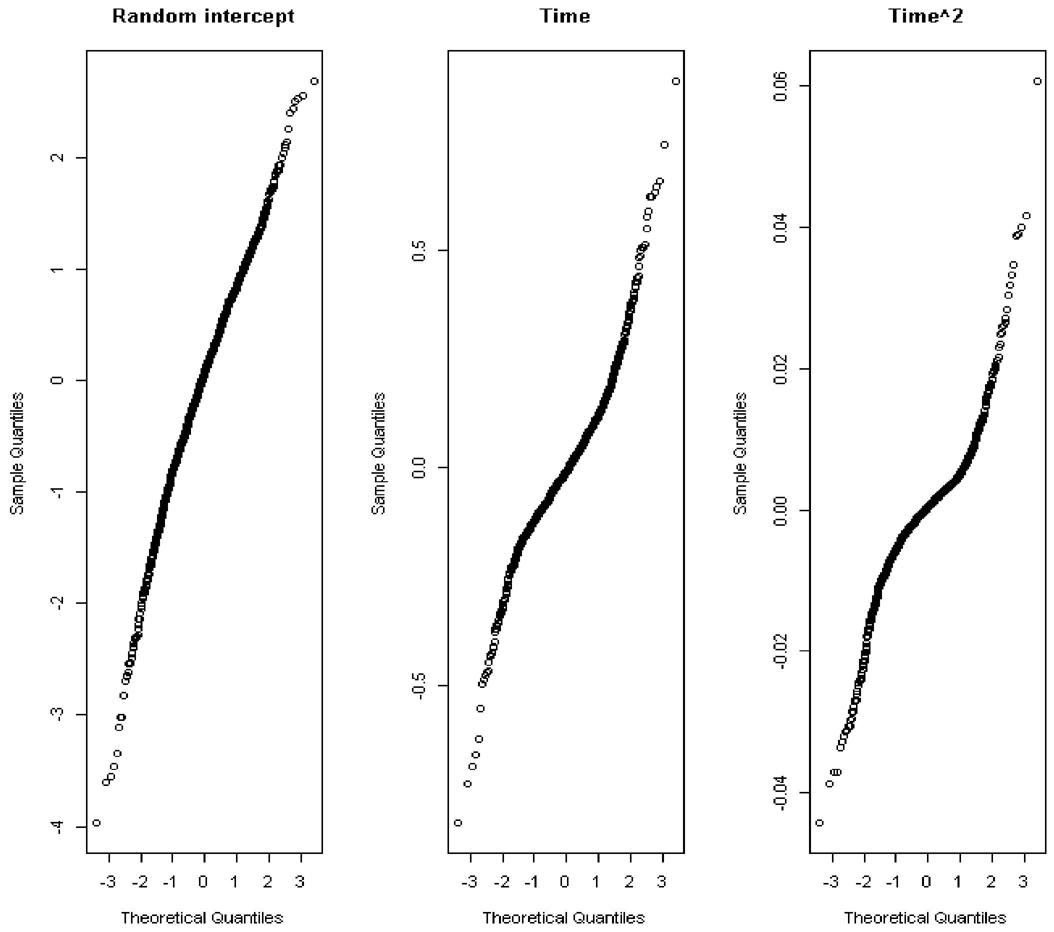

Figure5: *Q-Q plots for the random intercept, slope and quadratic time effects*

## ANNEXII

**Table 1: Observed correlation matrix for CD4 count**

| | 0 | 6 | 12 | 18 | 24 | 30 | 36 | 42 | 48 | 54 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1.000 | -0.013 | 0.007 | 0.048 | -0.012 | -0.040 | 0.043 | -0.021 | -0.010 | 0.031 |
| 6 | -0.013 | 1.000 | 0.006 | 0.053 | 0.016 | 0.016 | -0.013 | 0.032 | -0.016 | 0.000 |
| 12 | 0.007 | 0.006 | 1.000 | -0.012 | -0.031 | 0.011 | -0.001 | -0.049 | -0.001 | -0.049 |
| 18 | 0.048 | 0.053 | -0.012 | 1.000 | -0.006 | -0.015 | -0.020 | -0.016 | 0.026 | -0.016 |
| 24 | -0.012 | 0.016 | -0.031 | -0.006 | 1.000 | 0.025 | -0.013 | 0.002 | -0.010 | 0.015 |
| 30 | -0.040 | 0.016 | 0.011 | -0.015 | 0.025 | 1.000 | 0.000 | -0.023 | 0.025 | -0.008 |
| 36 | 0.043 | -0.013 | -0.001 | -0.020 | -0.013 | 0.000 | 1.000 | -0.005 | 0.014 | -0.001 |
| 42 | -0.021 | 0.032 | -0.049 | -0.016 | 0.002 | -0.023 | -0.005 | 1.000 | 0.013 | -0.002 |
| 48 | -0.010 | -0.016 | -0.001 | 0.026 | -0.010 | 0.025 | 0.014 | 0.013 | 1.000 | 0.006 |
| 54 | 0.031 | 0.000 | -0.049 | -0.016 | 0.015 | -0.008 | -0.001 | -0.002 | 0.006 | 1.000 |

**Table 2: Observed variance covariance matrix for CD4 count**

| | 0 | 6 | 12 | 18 | 24 | 30 | 36 | 42 | 48 | 54 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.922 | -0.009 | 0.004 | 0.025 | -0.007 | -0.024 | 0.025 | -0.012 | -0.004 | 0.012 |
| 6 | -0.009 | 0.522 | 0.003 | 0.021 | 0.007 | 0.007 | -0.005 | 0.013 | -0.005 | 0.000 |
| 12 | 0.004 | 0.003 | 0.396 | -0.004 | -0.012 | 0.004 | 0.000 | -0.018 | 0.000 | -0.013 |
| 18 | 0.025 | 0.021 | -0.004 | 0.296 | -0.002 | -0.005 | -0.007 | -0.005 | 0.007 | -0.004 |
| 24 | -0.007 | 0.007 | -0.012 | -0.002 | 0.399 | 0.010 | -0.005 | 0.001 | -0.003 | 0.004 |
| 30 | -0.024 | 0.007 | 0.004 | -0.005 | 0.010 | 0.388 | 0.000 | -0.008 | 0.007 | -0.002 |
| 36 | 0.025 | -0.005 | 0.000 | -0.007 | -0.005 | 0.000 | 0.353 | -0.002 | 0.004 | 0.000 |
| 42 | -0.012 | 0.013 | -0.018 | -0.005 | 0.001 | -0.008 | -0.002 | 0.332 | 0.004 | -0.001 |
| 48 | -0.004 | -0.005 | 0.000 | 0.007 | -0.003 | 0.007 | 0.004 | 0.004 | 0.232 | 0.001 |
| 54 | 0.012 | 0.000 | -0.013 | -0.004 | 0.004 | -0.002 | 0.000 | -0.001 | 0.001 | 0.173 |