SCHOOL OF GRADUATE STUDIES

DEPARTMENT OF STATISTICS

A Joint Modeling of Longitudinal and Survival data with Application to HIV-Infected Patients under HAART Follow-up: A case of Mekelle General Hospital, Ethiopia

---

By: ***Getu Boja***

Advisor: ***Birhanu Teshome (PhD)***

A Thesis Submitted to Statistics Department in Partial Fulfillment of the Requirements for the Degree of Master of Science in Statistics (Biostatistics)

June 2017

Addis Ababa, Ethiopia

Contents

Declaration

I, ***Getu Boja***, do hereby declare that this thesis entitled: ***"A Joint Modeling of Longitudinal and Survival Data with Application to HIV-infected Patients under HAART Follow-up: A case of Mekelle General Hospital"*** is entirely my own original work and has not been presented for higher degree at any other University or Institute anywhere for that award of any academic degree, diploma or certificate. All references made to works of other persons have been duly acknowledged.

Name: Getu Boja Gari

Signature: – – – – – – – – – – – – – – – – –

Place: Faculty of Science, Addis Ababa University

Date: June, 2017

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

– – – – – – – – – – – – – – – – –

Birhanu Teshome (PhD)

Approval

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
DEPARTMENT OF STATISTICS

This is to certify that, the thesis work prepared by ***Getu Boja***, entitled: ***"A Joint Modeling of Longitudinal and Survival Data with Application to HIV-infected Patients under HAART Follow-up: A case of Mekelle General Hospital"*** was carried out under strict supervision and has been approved for submission to the School of Graduate Studies of Addis Ababa University in partial fulfillment of the requirements for the award of the Degree of Master of Science in Statistics (Biostatistics) assembles with the regulations of the University and meets the accepted standards with respect to originality and quality.

Approved by the board of examiners:

Birhanu Teshome (PhD)	_____	_____
Advisor	Signature	Date
Butte Gotu (PhD)	_____	_____
Examiner	Signature	Date
Emmanuel G/Yohannes (PhD)	_____	_____
Examiner	Signature	Date
Mekonnen Tadesse (Associate Prof.)	_____	_____
Chair of Department	Signature	Date

The signature of the department head or an authorized signatory is an assertion of the authenticity of the committee's signature and the acceptability of the thesis to the department; therefore, the sign of the signatory must be original.

Acknowledgment

First of all, my unreserved gratitude goes to Almighty God, who generously gave me the strength, health and being with me in every step, with whom I attained this stage and grow to be successful. My special heartedly thanks goes to my advisor and instructor ***Dr. Birhanu Teshome****, for his motivations, inputs and contributions inspired me to vigorously pursue the execution of this research with my utmost interest which made this work a reality from inception to the final. I also wish to acknowledge the impulses I got from my former instructor* ***Mr. Said Musa*** *who helped me with resources to successfully accomplish this research amid other competing demands. Lastly, I would also like to express the great debt I owed to Addis Ababa University Department of Statistics and Wolkite University for giving me the opportunity to grasp a profound knowledge and financial support.*

Thanks again to all who helped me!

Getu Boja

June 2017

Addis Ababa, Ethiopia

Glossary

AIDS: Commonly refers to the advanced stage of HIV illness, when the CD4 cells count falls under 200.

Ambulatory: An individual able to perform activities for daily living.

ART: A class of drugs which inhibit the activity of retroviruses such as HIV. ART involves lifelong treatment.

ARV: Refers to drugs used against retroviruses, commonly anti-HIV drugs.

Bedridden: An individual unable to perform activities of daily living.

CD4: A receptor on the surface of cells that HIV attaches to. The cells involved in cell-mediated immunity known as T-lymphocytes have the CD4 marker. Other cells, including some in the brain have the same marker and are the targets of HIV. Higher is healthier.

CD4 cell: A type of white blood cell that fights infection. Also known as T-helper cells.

CD4 count: Represents the count of the cells with CD4 receptor in circulation. That is, the number of CD4 cells per microliter (μL) of blood.

HAART: Treatment with a combination of at least three different ARVs, such as different association of protease inhibitors (PI), non-nucleoside reverse transcriptase inhibitors (NNRTI) and nucleoside reverse transcriptase inhibitors (NRTI).

HIV: Refers to the human immunodeficiency virus, the virus that causes AIDS. There are two different types HIV-1 and HIV-2. HIV-1 is responsible for the vast majority of HIV infections globally.

WHO clinical stages of AIDS: Classification of the stages of HIV-associated clinical disease where *stage I* indicates asymptomatic disease, *stage II* indicates mild disease, *stage III* indicates advanced disease and *stage IV* indicates severe disease.

Working: An individual able to perform usual work in and out of the house, harvest, go to school for children, normal activities or playing.

List of Tables

List of Figures

List of Acronyms

3TC	Lamivudine
AIC	Akaike's Information Criterion
AIDS	Acquired Immunodeficiency Syndrome
ART	Antiretroviral Therapy
ARV	Antiretroviral
AZT	Azidothymidine/Zidovudine
BIC	Bayesian Information Criterion
CD4	Cluster of Differentiation Four
CI	Confidence Interval
d4T	Stavudine
EDA	Exploratory Data Analysis
EFV	Efavirenz
HAART	Highly Active Antiretroviral Therapy
HIV	Human Immunodeficiency Virus
I-TECH	International Training and Education Center for Health
JM	Joint Modeling
LMM	Linear Mixed Model
LRT	Likelihood Ratio Test
ML	Maximum Likelihood
NEV	Nevirapine
NRTI	Nucleoside Reverse Transcriptase Inhibitor
NNRTI	Non-nucleoside Reverse Transcriptase Inhibitor
RML	Restricted Maximum Likelihood
UNAIDS	Joint United Nations Programme on HIV/AIDS
USAID	United States Agency for International Development
WHO	World Health Organization

Abstract

A Joint Modeling of Longitudinal and Survival data with Application to HIV-Infected Patients under HAART Follow-up

Despite tremendous progress in the control of the global HIV epidemic, the burden of HIV is still severe in Sub-Saharan Africa. Longitudinal and survival data frequently observed together in practice and useful for analysis of HIV related data. The separate analyses of longitudinal and survival endpoints may not be adequate and could lead to inefficient estimation or biased results. Joint modeling approaches correct for this bias by accounting for the association between the two responses. The main purpose of this study was to jointly model and analyze longitudinal and survival endpoints with application to retrospective cohort data of 469 *HIV-infected patients under HAART follow-up in Mekelle General Hospital, Tigray, Ethiopia. The analysis consists of exploratory data analysis and fitting three different models namely; a linear mixed effects model for the longitudinal data, a semi-parametric survival model for the time-to-event data and a joint modeling of the two responses via shared random-effects approach. The results of both the separate and joint analyses are consistent. However, the use of a joint analysis compared to independent models shows a reduction in the standard errors which indicates that more adequate and efficient inferences can be made by using joint model estimates. The estimated association parameter (α) in the joint model is -0.138 (with 95% CI: -0.196 − -0.079) and statistically significant ($p - value < 0.0001$). This indicates that there is strong evidence of association between the effect of the longitudinal biomarker to the risk of death. The results indicates that higher initial values of CD4 cells is associated with a better survival. Furthermore, patients with lower initial weight, being male, late WHO clinical stage, being ambulatory and bedridden were associated with higher risk of death. Future extension of this research could possibly be to account for missing data and attempt should be given to health workers and data clerks working with patients under HAART to improve the quality of the data records of patients.*

Keywords: *HAART, HIV/AIDS Data, Joint Modeling, Longitudinal Data Analysis, Survival Data Analysis*

M.Sc Thesis

Getu Boja

July 4, 2017

Chapter 1

Introduction

1.1 Background of the Study

Human immunodeficiency virus (HIV) is the virus that causes HIV infection. HIV attacks and destroys the infection-fighting CD4+ T lymphocyte cells (hereafter referred to as CD4 cells) of the immune system. Acquired immune deficiency syndrome (AIDS) is the final stage of HIV infection. Since the first cases of what is now known as AIDS were reported back in 1981, an entire generation has grown up under the constant cloud of this modern day plague and the virus has infected people of all ages, sexes, races and income status, leading to poor health and socio-economic outcomes across the world (Moore, 2011). According to the 2016 UNAIDS report, there were 36.7 million people living with the HIV across the globe at the end of 2015.

Africa, Asia and Latin America were the major continents affected by the disease. Africa, and particularly Sub-Saharan Africa, has the most serious HIV/AIDS epidemic in the world, and is home for 76% of the global morbidity and 75% of the global mortality (Wang *et al.*, 2016 as cited in Gesesew *et al.*, 2017). In 2015, an estimated 19 million people were living with HIV, of whom women accounts for more than half the total number of people living with HIV. In the same year, there were an estimated 960,000 new HIV infections and 470,000 AIDS-related deaths. South Africa has the biggest and most highest profile of HIV epidemic in the world, with an estimated seven million people living with HIV in 2015. In the same year, there were 380,000 new infections while 180,000 South Africans died from AIDS-related illnesses (UNAIDS, 2016).

Ethiopia is one of the few countries with the highest number of people living with HIV/AIDS. Based on a single point estimate, there were nearly 1.2 million people living with HIV/AIDS in Ethiopia. Recent evidences show that HIV infection has significantly decreased over the years in the country. In 2015, Ethiopia had 39,140 new HIV infections, 786,040 people living with HIV, and 28,650 HIV/AIDS deaths (Wang *et al.*, 2016 as cited in Gesesew *et al.*, 2017). However, the prevalence and incidence rates significantly vary between geographical areas and gender. Across all the regions, urban areas are more affected than rural ones, while females are more affected than male population by the HIV epidemic.

In Tigray, the region where our study was conducted, the prevalence of the disease was estimated at about 3.1% in 2010 (urban areas up to 15% in females and 11.6% in males) and 1.3% rural (National Factsheet, 2010). In 2012, the Federal HIV/AIDS Prevention and Control Office (HAPCO) estimated that there were about $56,900$ HIV positive individuals in the region. HIV prevalence in Tigray varies widely across zones from 0.4% (Central zone) to 2.2% (Western) HAPCO and USAID, Tigray Health Bureau (2012) as cited in (Melaku and Zeleke, 2014).

The goal of antiretroviral (ARV) treatment is to decrease the morbidity and mortality that is generally associated with HIV infection. Antiretroviral therapy (ART) is a treatment for people infected with HIV using anti-HIV drugs. The standard treatment for patients infected with HIV is referred to as highly active antiretroviral therapy (HAART).

HAART generally consists of three or more different medications usually from at least two different classes, such as two nucleoside reverse transcriptase inhibitors (NRTIs) and a protease inhibitor (PI), a non-nucleoside reverse transcriptase inhibitor (NNRTI) or other such combinations. The U.S. Department of Health and Human Services (HHS) provides guidelines on the use of HIV medicines and also recommend starting ART with a regimen that includes three HIV medicines from at least two different drug classes (HHS, 2011). This treatment has led to a substantial reduction in mortality and disease progression to AIDS by increasing CD4 cells.

Currently available treatment can't cure HIV infection, but it can help people infected with HIV live longer, healthier lives. ARV treatment is the best option for long lasting viral suppression and, subsequently, for reduction of morbidity and mortality. Moreover, HAART stops

viral replication, allowing for CD4 cells reconstitution and delay in the onset of AIDS and the otherwise fatal course of HIV/AIDS (Palella *et al.*, 2006). But, the critical issue to the success of HAART is retention to the treatment regimen as HAART is a lifelong commitment that requires patients to diligently adhere to daily medication, dosing schedules protocol that often involves coordination of dietary intake and make regular clinic visits for care (Ickovics and Meade, 2002).

With the advent of the HAART, individuals living with HIV/AIDS are expected to live ever longer and slower progression to AIDS had been observed in many studies . As of June 2016, 18.2 million people living with HIV were accessing ART globally, up from 15.8 million in June 2015, 7.5 million in 2010, and less than one million in 2000 (UNAIDS, 2016). Although more and more people now have access to HIV treatment, there is still a long way to go. Only 60% of all people living with HIV know their HIV status whereas the remaining 40% (over 14 million people living with HIV) deserves to get *antiretroviral* treatment. Furthermore, despite tremendous progress in the control of the global HIV epidemic during the past decade, still there is a need to address critical gaps in prevention, testing and treatment services.

One of the main interest in HIV clinical research is the CD4 cells progression and survival of patients who receive HAART. Thus, statistical analyses and modeling have greatly contributed to understand the relationship between trends in a CD4 lymphocyte cells as a biomarker of disease progression and time to death of a patient under HAART follow-up. They also provide guidance for the treatment of AIDS patients and evaluation of HAART that can yield important insight into the mechanisms of disease progression of HIV-infection by modeling disease evolution.

Therefore, this research work was undertaken against the above background and explores the factors that have strong association with the longitudinal measures and the survival experience of HIV-infected patients who started HAART in Mekelle General Hospital, Tigray, Ethiopia.

1.2 Statement of the Problem

Although many advances have been made in the way HIV/AIDS is identified and treated, the burden of HIV is particularly severe in Sub-Saharan Africa. In many clinical studies, longitudinal data and survival data are frequently observed together in practice. A typical example of this

setting is HIV clinical trials, from which CD4 cells count as a biomarker of disease progression are regularly measured repeatedly at different time points and time-to-event of a patient (e.g. *death*) is recorded under HAART follow-up. For most HIV/AIDS clinical trials, longitudinal and survival data have been usually analysed considering time-to-event data (survival outcome) or repeated measurements (longitudinal outcome) separately (Fitzmaurice *et al.*, 2004).

These include linear mixed effects models for longitudinal data and parametric survival models or semi-parametric (Cox) proportional hazards models for survival data on the spread of HIV/AIDS in a given population. However, the uses of separate analysis of longitudinal variable which is correlated with patient health status and survival endpoint (either with the subject's status as well as the possibility of study dropout) may not be adequate and can lead to inefficient estimation or biased results because they fail to take into account the association between the two components of the data (Lim *et al.*, 2013).

In a situation, where both outcomes are observed in one subject, separate modeling does not take into account the dependence between the two types of responses. In order to overcome this problem, a powerful method is a joint modeling (JM) of longitudinal and survival data. The JM approach was introduced to address statistical issues that cannot be handled in separate analysis of longitudinal and survival data. It is generally behaved that when association between the two processes exists, less biased and more efficient inferences will be obtained by using joint model (Guo and Carlin, 2004).

Therefore, it is usually recommended to jointly model repeated measurements and time-to-event data altogether via shared random effects to account for the dependence between longitudinal and survival components on the same subject and any available covariates. This approach enables researchers to make the most efficient use of all data and identify effects of variables after correctly controlling the interplay among these processes.

1.3 Objectives of the Study

1.3.1 General Objective

The main purpose of this study was to jointly model and analyze longitudinal and survival endpoints with application to HIV-infected patients under HAART follow-up based on a retrospective cohort data records of Mekelle General Hospital, Tigray, Ethiopia.

1.3.2 Specific Objectives

In the light of this major objective, the specific objectives of the study were:

✓ To estimate effects of baseline covariates on longitudinal and survival endpoints.

✓ To examine the association between longitudinal biomarkers and survival event of interest.

✓ To demonstrate the advantage of joint model analysis techniques to the data.

1.4 Significance of the Study

Nowadays, there is increased medical interest in personalized medicine. As a result, joint models have been utilized to provide individualized predictions (Dimitris *et al.*, 2014). Within this context, our study was conducted based on a retrospective cohort data from Hospital records at Mekelle General Hospital consisting of patients under HAART follow-up. In many medical studies, longitudinal biomarkers and the event time of interest are collected simultaneously in order to explore their association. One main interest of these studies is to detect any impact of longitudinal CD4 cells count and treatments on the time to death.

The study was aimed at applied aspects of the joint modeling framework, and specifically, to examine whether different features of the longitudinal processes would change significantly the prediction for the events of interest by considering different types of association structures. That is, this study is important to understand how the repeated biomarker (CD4+ lymphocyte cells count) and the risk of event (survival time-to-death) outcomes are linked.

The outcome of this study would also provide information for public health practitioners and stakeholders who are working in the areas of giving care, support and treatment for HIV/AIDS patients by modeling disease evolution to understand HIV/AIDS prognosis and treatment effects

in HIV/AIDS.

Therefore, this study was designed to identify additional factors that affect CD4 cells progression and the survival time of HIV-infected patients after initiation of HAART as a case study in a governmental Hospital based on a retrospective cohort data records at Mekelle General Hospital, Tigray, Ethiopia.

1.5 Scope of the Study

The study was conducted in Mekelle General Hospital, Tigray regional state, which is 783kms away from the capital city of Ethiopia, Addis Ababa, in North Ethiopia. The area is located at 13^o $32'$ North latitude and 39^o $28'$ East longitude, with an elevation of 2084 meters above sea level. Tigray region has an estimated total population of $4,664,071$, of which $2,367,032$ are females. More than 80% of the population is estimated to be rural inhabitants (CSA, 2007).

Mekelle General Hospital is a governmental hospital in Mekelle which is one of the 14 hospitals in Tigray Regional state. The hospital serves as a referral hospital for nearby lower level hospitals and health centers. The hospital provides clinical care for patients infected with HIV/AIDS free of charge since March 2005. The hospital is under the Tigray Administration Health Bureau and gets technical and financial support from government and nongovernmental organizations such as International Training and Education Center for Health (I-TECH), the United States Agency for International Development (USAID), and others.

Chapter 2

Literature Review

2.1 Overview of HAART

HIV infection remains a global health problem of unprecedented dimensions, although the development of ART and mainly HAART in 1995 has changed the epidemiology of opportunistic infection and has significantly modified the course of HIV disease into a manageable chronic disease with longer survival and improved quality of life by decreasing the mortality and morbidity of HIV-infected subjects (Mataftsi *et al.*, 2011). Without ART, most HIV-infected individuals will eventually develop progressive immunodeficiency marked by CD4 T lymphocyte (CD4) cells depletion and leading to AIDS-defining illnesses and premature death.

The therapeutic benefits of ART are often limited by long-term toxicities and evolution of drug-resistant virus. In resource-rich countries, HIV treatment is monitored routinely with laboratory measures such as blood chemistry, HIV viral load, and CD4 cells count for early detection of side effects of medications and drug-resistant virus. Due to the lack of accessible and affordable laboratory services, routine laboratory monitoring is not feasible in most resource-limited countries like Ethiopia (Barry *et al.*, 2013). Without laboratory monitoring, many patients may experience prolonged virologic failure and develop drug resistance mutations, which could ultimately limit second-line treatment options, increase morbidity, mortality and increase transmission of resistant viruses in the population (Sawe and McIntyre, 2009).

The World Health Organization (WHO) recommends CD4 cells count monitoring every six months and viral load testing only when the capacity exists. However, the quality and access to

CD4 cells count tests and viral load measurements vary in resource-limited settings, even where they are recommended in local treatment guidelines, because of inadequate resources. This is going to become even more challenging as treatment programs are rolled out from big hospitals in urban centers to primary health care facilities in the rural areas that are closer to the patients for initiation and continued care of stable patients on treatment (Geretti *et al.*, 2008).

2.2 Empirical Literature

Joint models for longitudinal and time-to-event data are increasingly used to assess relationships between serial measurements of one or several markers and time-to-event of interest. Joint models were introduced during the 90s (Faucett and Thomas, 1996; Wulfsohn and Tsiatis, 1997 as cited in Neuhaus *et al.*, 2009) and since then have been applied to a great variety of studies in epidemiological and biomedical areas. In turn, these studies have fed a wide methodological research on the subject, with models focused on event times, longitudinal patterns, or both.

The early development of joint models for longitudinal and survival data was largely motivated by data from HIV/AIDS clinical trials that were designed to evaluate the therapeutic effects of treatments on the development of AIDS or death, where CD4+ lymphocyte cells count and viral loads were used as markers for disease escalation. These articles include (Wang and Taylor, 2001; Brown *et al.*, 2005; Chi and Ibrahim, 2007) among others.

The two different approaches for joint models of longitudinal and time-to-event data are: the *shared random-effect models* and the *joint latent class models*. The difference between them depends on the parametrization of the joint likelihood of the longitudinal and survival processes; and the research interest. However, the popular approach in joint modeling of longitudinal and survival data is the one based on shared random effects, where the longitudinal model and survival model share common random effects and these random effects then induce correlation between the longitudinal and survival components of the model (Zhang *et al.*, 2016). In this case, longitudinal data and survival data are considered to be independent conditional on the random effects and observed covariates.

An excellent overview of the development of joint models is made by (Tsiatis and Davidian,

2004). The author's focus on models for the longitudinal process and the hazard for the time-to-event that depend jointly on shared, underlying random effects. As demonstrated in this article, the joint model leads to correction of potential biases for enhanced efficiency. Andrinopoulou, (2014) following on the previous discussion, under the joint modelling of longitudinal and survival data to optimally utilize the relationship between repeated aortic valve function measurements and time-to-death or time-to-re operation and found a model is more realistic from a biological point of view compared to the time-dependent Cox models due to the fact that they explicitly assume that biomarkers evolve smoothly over time and do not remain constant between visits.

As Henderson *et al.*, (2000) noted, joint modeling is a flexible methodology for handling combined longitudinal and event history data. But when the association parameter between the longitudinal and survival data is not significant, the joint model analysis should have the same results as would be obtained from separate analyses for each component. Joint modelling is a valuable technique not only in its efficient use of all available data and its ability to obtain accurate inference, but it is highly recommended when survival time and longitudinal measurements have the same clinical meaning. It is especially applicable to problems involving biomarkers where the focus is on using longitudinal measurements to improve prediction of survival prognosis. If survival time and longitudinal measurements have a different clinical meaning or are not comparable, joint modeling is not appropriate because the result may lead to the conclusion that a covariate effect with worse survival is superior (Finkelstein and Schoenfeld, 1999 as cited in Lim *et al.*, 2013).

While not yet considered a standard modeling technique in most areas of application, joint models for longitudinal and time-to-event data have been the topic of numerous research publications and several excellent review papers. Accordingly, Ibrahim *et al.*, (2010) provided a non-technical description of an application of joint modeling techniques for a cancer clinical trial with a quality of life outcome and noted the following advantages of joint modeling in clinical trials:

1. They provide more efficient estimates of the treatment effects on the time-to-event,
2. They provide more efficient estimates of the treatment effects of the longitudinal marker, and

3. They reduce bias in the estimates of the overall treatment effect, that is, the treatment effect on survival and the longitudinal marker. Therefore, a less biased estimate leads to a more accurate estimate of the treatment effect.

Martins *et al.*, (2010) aimed to study the relationship of CD4+ lymphocyte cells count (longitudinal outcome) with time to death (survival outcome) in predicting the median survival time of HIV/AIDS patients in Brazil, and they showed that the Bayesian joint model presents considerable improvements in the median survival time distributions when compared with those obtained through longitudinal and survival models separately.

Rizopoulos, (2010) proposed a joint model where the time-to-event process is of main interest and influenced by a longitudinal time-dependent covariate measured with error and made a great contribution facilitating the use of the joint modeling methodology by developing the ***JM R*** packages for the shared-effects modeling approaches.

Philipson *et al.*, (2012) developed a shared random effects joint models where the focus is on both survival and longitudinal processes, with normality assumptions on the random effects. Vonesh *et al.*, 2006 addressed the need of jointly modeling analysis for the longitudinal repeated biomarker measurements, usually a linear mixed effects model, jointly with the survival submodel of a time-to-event process with informative censoring time.

A study that included 1,259 adult (>18 years age) HIV/AIDS patients who were undertaking Antiretroviral Therapy in the ART centre of Dr. Ram Manohar Lohia Hospital, New Delhi, India, gave a proper platform to study situations where the association of longitudinal data with the time-to-event has utmost importance. So, a fitted joint model to simultaneously study the longitudinal repeated measures on CD4+ cells count and the time-to-event (event being defined as loss to follow-up) process of HIV/AIDS patients under ART treatment should be preferred over separate models for longitudinal and survival data analysis (Gurprit *et al.*, 2015).

Recently, (Aboma and Teshome, 2016) used data from Jimma University Specialized Hospital, South West of Ethiopia and jointly modelled the longitudinal CD4 cells count and weight measurements of HIV/TB co-infected patients, and found that sex, educational level and functional status were the factors contributing to the prediction of HIV/TB co-infected patients

weight at baseline among other variables included in the study from the joint model.

Seid *et al.*, (2014) also, fitted a joint model of the longitudinal CD4 cells count and the default time processes and linked them using unobserved random effects through the use of a shared parameter model. They concluded that the results of both the separate and joint analyses are consistent. However, they suggested that the joint model is the simplest model compared to the separate model as its effective number of parameters is smaller.

Similarly, based on secondary data collected from 354 HIV/AIDS patients with ages 16 years and older (Gemeda *et al.*, 2015) employed separate and joint statistical models in the Bayesian framework for longitudinal measurements and time to death event data of HIV/AIDS patients at Shashemene Referral Hospital, Ethiopia. The results of both the separate and joint analyses were consistent. However, the final joint model was found to be simpler (less complex) model than the separate models.

In the circumstances discussed above, a joint model is desirable over separate models or even over the time-dependent Cox model, to understand the association between the longitudinal and the time-to-event processes. Joint modeling can be perceived to be a sophisticated and complex approach in terms of estimation, however, its superiority comes from its ability to model the longitudinal repeated biomarkers measurements and the survival processes together while also taking into account the association between them (Gurprit *et al.*, 2015).

Therefore, joint models of longitudinal and survival data (Rizopoulos, 2012) represent a powerful statistical tool capable of capturing the association between longitudinal and survival time data. Specifically, they incorporate all information simultaneously and provide valid and efficient inferences. That is, joint models for longitudinal and time-to-event data are models that bring these two data types together (simultaneously) into a single model so that one can infer the dependence and association between the longitudinal biomarker and time-to-event to better assess the effect of a treatment or confounding variables.

Chapter 3

Data and Methodology

3.1 Description of Data

The data used for this study were obtained from a retrospective cohort study based on HAART electronic data base and from the review of patient charts at Mekelle General Hospital, Tigray, Ethiopia. The study population consists of all HIV+ patients who were 16 years old and older, and started the HAART treatment between 1^{st} January 2009 to 31^{st} December 2011. People aged under 15 years at seroconversion were excluded from all analyses as the definition of AIDS differs in children. Both the longitudinal and survival data were extracted from the patient's register which contains socio-demographic characteristics, baseline clinical, and laboratory measurement information of all patients under follow-up from the selected hospital records.

3.2 Variables in the Study

Variables considered in this study were selected based on related studies conducted at Shashemene Referral Hospital, Ethiopia (Gemeda *et al.*, 2015), Jimma University Specialized Hospital, South West of Ethiopia (Seid *et al.*, 2014) and other related literatures.

3.2.1 Response Variables

The two response (*outcome*) variables considered for this study are the longitudinal CD4 cells count and the survival outcome. The number of CD4 cells count per mm^3 of blood, which is considered as a biomarker, was measured approximately every six months irrespective of their visit to ART centres whereas the survival outcome is time in months to death from HIV/AIDS

and *death* was defined as confirmed deaths from medical records.

3.2.2 Predictor Variables

The predictor (*independent*) covariates comprise baseline demographic and socio-economic variables, and were presented in the Table 3.1 together with their descriptions.

Table 3.1: Predictors used in the Separate and Joint Analysis of the HIV/AIDS Data

No.	Variable	Description
1	Observation time	The time points at which the CD4 cells count was recorded
2	Baseline Age in years	Indicating patients age at enrollment
3	Sex of patients	Female, Male
4	Functional Status	Ambulatory, Bedridden, Working
5	WHO clinical stage	Stage I, Stage II, Stage III, Stage IV
6	Baseline Weight	Patient's weight in kilograms at enrollment
7	Baseline Regimen Type	AZT+3TC+NVP or EFV, d4T+3TC+EFV or NVP, Others

3.3 Methodology

The analysis consists of exploratory data analysis, and three different models namely; a linear mixed effects model for the longitudinal data, the Cox proportional-hazards model for the time-to-event data and a joint modeling of them altogether. Analyses were conducted using ***SAS version 9.4*** and ***R version 3.3.3*** statistical software packages. Statistical decision was made at ***5%*** level of significance.

3.3.1 Data Exploration

Exploratory data analysis (EDA) was conducted in order to investigate various associations, structures and patterns exhibited in the data set. This consists of obtaining the summary statistics such as frequencies and percentages in a particular group. In addition, the individual profile plots, mean structure, correlation structure and variance structure plots were obtained in order to gain some insights of the data (Verbeke and Molenberghs, 2000).

The individual profile plots and the variance structure were used to gain insight of the variability in the data and to determine whether random effects (random intercepts and slopes) were to be considered in the analysis. The mean structure was used to gain intuition on the

time function that can be used to model the data. Furthermore, the Kaplan-Meier Estimator was used to estimate and graph survival probabilities as a function of time. Also, it was used to obtain univariate descriptive statistics for the survival data, including the median survival time, and compare the survival experience for two or more different groups of patients.

3.4 Statistical Models

3.4.1 Longitudinal Sub-model

Longitudinal responses may arise in two common situations. One is when the measurements are taken on the same subject at different times (*i.e.*, when multiple observations are made on the same subject or unit of analysis over time) and the other is when the measurements are taken on related subjects. In both cases, the responses are likely to be correlated (Laird and Ware, 1982). For longitudinal data, two sources of variations are considered. These are the within-subject which arises during the measurements within each subject, and between subject variation which arises during the measurement between different subjects. Modeling within subject variations help us to study changes overtime while modeling between subject variation help us to understand differences between subjects.

A standard modeling framework for the analysis of longitudinal data is the mixed-effects model. A mixed model is one that contains both fixed and random effects. The fixed effects part of the model represents the mean response, while the random effects part represents the individual level responses. Linear mixed models (LMM) may be expressed in different but equivalent forms. For the continuous case, the LMM provide a general and flexible modeling framework where subject-specific random effects, assumed to follow a normal distribution, are included to account for the correlation (Laird and Ware, 1982; Verbeke and Molenberghs, 2000).

Let $\boldsymbol{\beta}$ denote a $p \times 1$ vector of unknown population coefficients for the fixed effects and $\mathbf{X}_i$ be a known $n_i \times p$ design matrix values of the fixed predictors linking $\boldsymbol{\beta}$ to set of longitudinal measurements $\mathbf{y}_i$. Let $\mathbf{b}_i$ denote a $k \times 1$ vector of unobservable individual random effects and $\mathbf{Z}_i$ be a known $n_i \times k$ design matrix values of the random factors linking $\mathbf{b}_i$ to $\mathbf{y}_i$, and $\boldsymbol{\epsilon_i}$ is the $n_i \times 1$ vector of unknown random errors. Then against this background, the general LMM for

the longitudinal endpoint has the form:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon_i}, \tag{3.4.1}$$

$$\begin{cases} \mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D}) \\ \boldsymbol{\epsilon_i} \sim N(\mathbf{0}, \sum_i) \\ \mathbf{b}_1, ..., \mathbf{b}_n \text{ , and } \boldsymbol{\epsilon}_1, ..., \boldsymbol{\epsilon}_n \text{ are independent} \end{cases}$$

where $\mathbf{y}_i$ is the $n_i \times 1$ response vector for observations in the i^{th} subjects and $\boldsymbol{\epsilon_i}$ is distributed as N($\mathbf{0}$, $\sum_i$) is a vector of residuals components, combining measurement error and serial correlation. The $\mathbf{b}_i$ are distributed as N($\mathbf{0}$, $\mathbf{D}$), independently of each other and of the within-subjects residuals $\boldsymbol{\epsilon_i}$. That is, $cov(b_i, \epsilon_i) = 0$. Furthermore, $\sum_i$=$\delta^2 I_{ni}$ is the $n_i \times n_i$ positive-definite variance-covariance matrix for the errors in subject i, where I_{ni} denotes the $n_i \times n_i$ identity matrix.

Marginally, the vector $\mathbf{y}_i$ is normally distributed with mean $\mathbf{X}_i\boldsymbol{\beta}$ and variance-covariance matrix of $\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^{\mathbf{T}} + \boldsymbol{\delta}^2\mathbf{I}_{ni}$. Here $\mathbf{D}$ is a $k \times k$ positive-definite covariance matrix for random effects. Conditional on $\mathbf{b}_i, \mathbf{y}_i$ is normally distributed with mean $\mathbf{X}_i\boldsymbol{\beta}+\mathbf{Z}_i\mathbf{b}_i$ and with variance-covariance matrix $\sum_i$. It can also be rewritten as: $\mathbf{y}_i|\mathbf{b}_i \sim N(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \sum_i)$. That is, given the random effects $\mathbf{b}_i$, the dependent variable $\mathbf{y}_i$ is normally distributed with variance-covariance structure.

Covariance Structures

A model for the covariance must be chosen on the basis of some assumed model for the mean response. In order to reduce the number of parameters in the variance-covariance structure $\sum$, we can fit models with more parsimonious structures. The following are commonly used $\sum$'s among others: Independent (IND), Compound symmetry (CS), Heterogeneous compound symmetry (CSH), First-order autoregressive (AR(1)), and Unstructured (UN). These often lead to more efficient inferences for the mean parameters, and particularly useful when many repeated measurements are taken per subject (Fares, 2016).

Independent (IND)

The simplest covariance structure is the IND structure, where the within-subject error correlation is zero.

$$\sum = \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ 0 & 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

Compound symmetry (CS)

The covariance structure with the simplest correlation model is the CS structure. It assumes that the correlation is constant regardless of the distance between the time points. The corresponding correlation model is

$$\sum = \begin{bmatrix} \sigma^2 & \sigma^2\rho & \sigma^2\rho & \dots & \sigma^2\rho \\ \sigma^2\rho & \sigma^2 & \sigma^2\rho & \dots & \sigma^2\rho \\ \sigma^2\rho & \sigma^2\rho & \sigma^2 & \dots & \sigma^2\rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma^2\rho & \sigma^2\rho & \sigma^2\rho & \dots & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \rho & \rho & 1 & \dots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \dots & 1 \end{bmatrix}$$

Heterogenous compound symmetry (CSH)

This structure has non-constant variance and constant correlation. The general form of this covariance structure for each subject is as follows:

$$\sum = \begin{bmatrix} \sigma_1^2 & \sigma_2\sigma_1\rho & \sigma_3\sigma_1\rho & \dots & \sigma_n\sigma_1\rho \\ \sigma_2\sigma_1\rho & \sigma_2^2 & \sigma_3\sigma_2\rho & \dots & \sigma_n\sigma_2\rho \\ \sigma_3\sigma_1\rho & \sigma_3\sigma_2\rho & \sigma_3^2 & \dots & \sigma_n\sigma_3\rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_n\sigma_1\rho & \sigma_n\sigma_2\rho & \sigma_n\sigma_3\rho & \dots & \sigma_n^2 \end{bmatrix}$$

First Order Autoregressive (AR(1))

The AR(1) structure is often used to fit models to data sets with equally spaced longitudinal observations on the same units of analysis. This structure implies that observations closer to

each other in time exhibit higher correlation than observations farther apart in time (Brady *et al.*, 2007). The general form of the $\sum_i$ matrix for this covariance structure is as follows:

$$\sum = \begin{bmatrix} \sigma^2 & \sigma^2\rho & \sigma^2\rho^2 & \dots & \sigma^2\rho^{n-1} \\ \sigma^2\rho & \sigma^2 & \sigma^2\rho & \dots & \sigma^2\rho^{n-2} \\ \sigma^2\rho^2 & \sigma^2\rho & \sigma^2 & \dots & \sigma^2\rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma^2\rho^{n-1} & \sigma^2\rho^{n-2} & \sigma^2\rho^{n-3} & \dots & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{bmatrix}$$

The AR(1) is a special case of the Toeplitz covariance structure and is useful for modeling first order temporal autocorrelation structure.

Unstructured (UN)

The most complex covariance structure is UN covariance because it is estimating unique correlations within-subject errors for each pair of time interval. It adds a significant number of free parameters to the fitting process since a $p \times p$ covariance matrix has $\frac{p(p+1)}{2}$ non-redundant elements.

$$\sum = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_p^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & \rho_{12} & \dots & \rho_{1p} \\ \rho_{21} & 1 & \dots & \rho_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \dots & 1 \end{bmatrix}$$

Each of these can be described in a fairly intuitive manner. The correlation and/or the covariance structure will be obtained in order to determine the type of correlation structure of the random effects to be considered in the model. Therefore, in any given analysis, we try to determine the structure for the $\sum_i$ matrix that seems most appropriate and parsimonious, given the observed data and knowledge about the relationships between observations on an individual subject. Hence, a common recommendation is to only choose the covariance structures that make sense given the data.

Random Intercept Model

The random effects model or subject-specific model assumes that extra correlation arises among the longitudinal response (Diggle *et al.*, 2002). The random intercepts model allows intercepts to vary across groups. In particular, a basic example of a random intercepts model which is

included in order to illustrate the model fitting is formed by two clearly distinct parts,

$$y_i = \beta_0 + \beta_1 x_{ij} + b_{0i} + \epsilon_i$$

these are, a fixed part (which is the intercept and the coefficient of the explanatory variable times the explanatory variable) and a random part. The random part is composed of two random terms, $\epsilon_i \sim N(0, \sigma^2)$ and $b_i \sim N(0, \sigma_b^2)$. The random effect b_i and within-subject error ϵ_i are independent for different subjects and independent of each other for the same subject. *i.e.*, $Cov(b_i, b_j) = 0$ if $i \neq j$, $Cov(\epsilon_i, \epsilon_j) = 0$ if $i \neq j$, and $Cov(b_i, \epsilon_i) = 0$. In the mixed model formulation in Equation (3.4.1), the design matrices are replaced by:

$$X_i = \begin{bmatrix} 1_1 & x_{i1} \\ \vdots & \vdots \\ 1_{ni} & x_{ini} \end{bmatrix}, \; Z_i = \begin{bmatrix} 1_1 \\ \vdots \\ 1_{ni} \end{bmatrix}, \; \beta = \begin{bmatrix} \beta_0 & \beta_1 \end{bmatrix}^T$$

and the random effects model covariance structure, $b_i \sim N(0, D_i)$, with $D_i = \sigma_b^2$.

Random Intercept and Slope Model

An intuitive extension that also allows a random shift in the subject-specific slopes is known as random intercepts and random slopes model. Consider the simple random intercepts and slopes model,

$$y_i = \beta_0 + \beta_1 x_{ij} + b_{0i} + b_{1i} x_{ij} + \epsilon_i$$

In this model we additionally have b_{1i} which represents the random slope effect of the coefficient x_{ij}, $j = 1, ..., n_i$ denotes the j^{th} response on i^{th} subject. As a result, actually two extra parameters should be estimated: the variance in intercepts between groups σ_{b0}^2 and the variance in slopes between groups σ_{b1}^2. In this case the model matrix Z_i has the form,

$$Z_i = \begin{bmatrix} 1_1 & x_{i1} \\ \vdots & \vdots \\ 1_{ni} & x_{ini} \end{bmatrix},$$

and the random effects model covariance structure,

$$\begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} \sim N(0, D_i), \ with \ D_i = \begin{bmatrix} \sigma^2_{b0} & \sigma_{b0b1} \\ \sigma_{b0b1} & \sigma^2_{b1} \end{bmatrix}$$

where σ_{b0b1} denotes the covariance between the intercepts and slopes.

3.4.2 Survival Sub-model

Survival analysis is an area of statistics that studies the time until a pre-specified event of interest occurs. To determine time-to-event correctly, it is necessary to choose an appropriate time origin which has to be easily identified for all patients. Usually, in a medical context a single time-to-event is the time to recurrence of a health condition, time of response to a treatment or time to death from a certain cause that can be measured in days, weeks, months, years, *etc.* In this study, we refer *failure time* or *time to death* with the same meaning as *time-to-event.*

The most important characteristic that distinguishes the analysis of survival times from other areas in statistics is *Censoring.* Subjects are said to be *censored* if they are lost to follow up, withdrawing from the study, or if the study ends before they die or have an outcome of interest. That is, observations are called *censored* when information about their survival time is incomplete. There are three kinds of censoring: *right censoring*, *left censoring*, and *interval censoring* (Klein and Moeschberger, 2003). By far the most common type of censoring is right censoring, which occurs when observation is terminated before an individual experiences the event of interest. This could happen if a patient survives through the experiment and was still alive when the experiment concludes. An observation is also right-censored if a patient leaves the experiment for some reason not connected with survival. A slightly less common type of censoring is interval censoring, which means that an individual is known to have an event between two points in time, but the exact time is unknown. The least common type of censoring is left censoring, which happens when an event is known to have occurred before the start of the study, but the exact time is unknown.

Let T be a non-negative continuous random variable representing the time until the event of interest. The more optimistic *survival function S(t)* at time *t*, S(t)=P(T > t) is defined to be the probability that a randomly selected individual will survive beyond time *t.* It is a decreasing

function, taking values in [0, 1] which equals 1 at $t = 0$ and 0 at $t = \infty$. We regard T as the failure time for the i^{th} patient and C as the corresponding censoring time. When T is subject to right censoring, X_i is the observed time which is a minimum of (T_i, C_i), *i.e.,* X_i is equal to T_i if the event is observed and is equal to C_i if it is censored. Let $\delta_i = \mathrm{I}(T_i \leq C_i)$, where I(.) is an indicator function and takes the value

$$\delta_i = \begin{cases} 1, & when\ T_i \leq C_i. \\ 0, & otherwise. \end{cases}$$

The Proportional Hazards (PH) regression, also called the Cox PH model (Cox, 1972), is the most widely used semi-parametric survival regression model in which for a set of covariates $\mathbf{x}_i$ for the i^{th} subject, and $\boldsymbol{\beta}$ is the $p \times 1$ parameter vector of coefficients, the hazard at time t is expressed as:

$$\lambda_i(t) = \lambda_0(t) exp\{\boldsymbol{\beta}^T X_i\}, \tag{3.4.2}$$

where $\lambda_i(t)$ represents the hazard of death for a patient i at time t. $\lambda_0(t)$ is a baseline hazard function that describes the risk for individuals with $\mathbf{x}_i = 0$ which serves as a reference cell or pivot. The hazard function is a measure of the potential for the event to occur at a particular time t, given that the event did not yet occur. A larger values of the hazard function indicate greater potential for the event to occur. In this model (3.4.2), $exp\{\beta_i\}$ denotes the ratio of hazards for one unit change (increase or reduction in risk) in the i^{th} covariates at any time t, and the model also assume that covariates have a *multiplicative* effects on the hazard function for an event associated with the set of characteristics $\mathbf{x}_i$. However, taking on the log scale we find that the proportional hazards model is a simple *additive* model. Then alternatively, the model can be expressed as:

$$log\ \lambda_i(t) = log\ \lambda_0(t) + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}, \tag{3.4.3}$$

where $log\ \lambda_0(t)$ is the log of the baseline hazard. As in all additive models, we assume that the effect of the covariates x is the same at all times t. Moreover, in Cox PH model, no distributional assumption is made for the survival data, the only assumption is that the hazards ratio is constant over time $\psi = \frac{\lambda_i(t)}{\lambda_0(t)} = exp\{\beta_i\}$ (*i.e.,* proportional hazards). Owing to its semi-parametric nature, the Cox PH model has become routine in survival analysis in many situations. Because of the model form in Equation (3.4.2), the estimated hazards are always non-negative.

3.4.3 Joint Longitudinal-Survival Models

A novel use of joint models, which gains increasing interest in recent years, refers to the statistical analysis of the resulting data while taking account of any association between the repeated measurement and time-to-event outcomes (Diggle *et al.*, 2008; Jue *et al.*, 2016). Joint longitudinal-survival models can be formed where the association between the two endpoints is due to shared random effects.

This study was mainly focused on the use of a joint model, where the longitudinal and survival processes are assumed to be conditionally independent given unobserved random effects. That is, the key assumption of a joint model is that the random effects underlie both the longitudinal and survival processes. This means that these random effects account for both the association between the longitudinal and event outcomes, and the correlation between the repeated measurements in the longitudinal process (conditional independence assumption). This type of joint model is also called a shared parameter model, as both processes share these random effects (Rizopoulos, 2010; Sousa, 2011).

Therefore, in this study, we have used some of the methodologies, notations and equations used by (Rizopoulos, 2012) and employ the joint models that belong to the random effects shared parameter models framework as both sub-models share the same random effects. Let T_i represent the failure time for the i^{th} individual such that either censoring or the event has occurred. Without loss of generalizability, our aim is to associate the *true* and *unobserved* value of the longitudinal outcome at time t, denoted by $m_i(t)$, with event outcome T_i. The longitudinal and survival components of the joint model are typically linked (joined) through the trajectory function. Specifically, the shared random-effect models at time t can be written as:

$$\lambda_i(t|M_i(t), \boldsymbol{\omega}_i) = \lambda_0(t) exp\left\{\boldsymbol{\gamma}^T \boldsymbol{\omega}_i + \alpha m_i(t)\right\}, \ t > 0, \tag{3.4.4}$$

where $M_i(t)$ represents the history of the true (*unobserved*) longitudinal response, $m_i(t)$, up to time t, $\boldsymbol{\omega}_i$ represents the vector of baseline covariates with corresponding parameter estimates $\boldsymbol{\gamma}$, and α measures (quantifies) the effect of the longitudinal outcome to the risk of an event (*i.e.*, in our case effect of number of CD4 cells to the risk of death). Hence with this formulation, the risk of an event at time t is dependent on the true value of the longitudinal endpoint at that

time.

3.5 Estimation and Inference

Parameter estimates are mainly obtained through the use of maximum likelihood (ML) or restricted maximum likelihood (REML) estimation. ML is a very general approach to statistical estimation that is widely used to handle many difficult estimation problems.

3.5.1 Linear Mixed Effects Estimation

In general terms, we use efficient estimation using likelihood-based models either ML or REML estimation to obtain estimates of the covariance parameters in LMM with the remark that REML is usually better than ML, because it reduces the well-known finite sample bias in the estimation of the covariance (Fitzmaurice *et al.*, 2004).

The distinction between ML and REML is the construction of the likelihood function. However, the two methods are asymptotically equivalent and often give very similar results except the difference becomes important only when the number of fixed effects is relatively large.

Given that the i^{th} subject outcomes have the same random effects they will be marginally correlated, so we assume that

$$f(y_i|b_i;\theta) = \prod_{j=1}^{n_i} f\left\{y_{ij}|b_i;\theta\right\},$$

That is, longitudinal responses of a subject are independent conditionally on its random effect. As random effects have expected values of zero and therefore do not affect the mean, this distribution has a mean vector $X_i\beta$ and a covariance matrix V_i, then

$$f(y_i;\theta) = (2\pi)^{-\frac{n_i}{2}} exp\{\frac{-1}{2}(y_i - X_i\beta)^T V_i^{-1}(y_i - X_i\beta)\},$$

where $\theta^T = (\beta, V)$ with $V_i = Z_i D Z_i^T + \sigma^2 I_{ni}$.

Taking into account that we assume independence across subjects, the likelihood function is

simply the product of the density functions for each subject. The log-likelihood of a linear mixed model is given by:

$$l(\theta) = \sum_{i=1}^{n} log f(y_i; \theta),$$

Given V_i, the estimates of fixed-effects parameters are obtained by maximizing the log-likelihood function, conditionally on the parameters in V_i, and have a closed-form solution:

$$\hat{\beta} = (\sum_{i=1}^{n} X_i^T V_i^{-1} X_i)^{-1} \sum_{i=1}^{n} X_i^T V_i^{-1} X_i,$$

In general, ML and REML both have the same merits of being based on the likelihood principle which leads to useful properties such as consistency, asymptotic normality, and efficiency. But the REML produces less biased estimators for many special cases (Verbeke and Molenberghs, 2000).

3.5.2 Survival Estimation

The Kaplan-Meier estimator or Product Limit Estimator provides a non-parametric maximum likelihood estimate of the survivor function (Kaplan and Meier, 1958). The Kaplan-Meier estimate of $S(t)$ is given as

$$\hat{S(t)} = \prod_{t_i<t} \frac{n_i - d_i}{n_i},$$

where n_i corresponds to the number of observations at risk of failing just prior to time t_i; d_i denotes the number of failures at time t_i.

In order to estimate the survival function, the parameter estimation and their estimated variances in the Cox PH model can be found by maximizing the log-partial likelihood function with respect to the parameters.

Let the sub-index i refer to the subject indicator and consequently, $\{X_i, \delta_i\}; i = 1, ..., n$ denote their survival information. Taking a random sample from a certain distribution, parameterized by θ, the likelihood function is given by,

$$l(\theta) = \prod_{j}^{n} f(X_i; \theta)^{\delta_i} S_i(X_i; \theta)^{1-\delta_i}$$

Note that it takes to account for censoring information, by contributing with $f(T_i;\theta)$ when an event is observed at time T_i and with $S(T_i;\theta)$ when subjects survived up to that point, that is $T_i > X_i = C_i$. This can be rewritten in terms of hazard function as,

$$l(\theta) = \prod_{j}^{n} \lambda(X_i;\theta)^{\delta_i} exp\{-\Lambda(t)\}^{1-\delta_i}, \tag{3.5.1}$$

where $\Lambda(.)$ is the cumulative risk function which describes the probability that the event of interest has occurred up until time t. To address this issue, iterative optimization procedures could be necessary to locate the maximum likelihood estimates $\hat{\theta}$ using iterative numerical analysis techniques often done via the Newton-Raphson algorithm (Lange, 2004), which is based on the following iterative procedure:

$$\hat{\beta}_{New} = \hat{\beta}_{Old} + I^{-1}(\hat{\beta}_{Old})U(\hat{\beta}_{Old}),$$

with $U(\hat{\beta}_{Old})$ is the vector of scores and $I^{-1}(\hat{\beta}_{Old})$ is the inverse of the observed information matrix. Convergence is reached when $\hat{\beta}_{Old}$ and $\hat{\beta}_{New}$ are sufciently close together.

3.5.3 Joint Modeling Estimation

The main estimation method that has been proposed for joint models is ML (Hsieh *et al.*, 2006; Henderson *et al.*, 2000 as cited in Rizopoulos, 2012). The standard ML method involves maximizing the log-likelihood, given in Equation (3.4.4), corresponding to the joint distribution of the time-to-event and longitudinal data processes. Strictly, both processes share the same unobserved random effects, and are conditionally independent given these random effects (Rizopoulos, 2012), thus

$$f(T_i, \delta_i, y_i|b_i;\theta) = f(T_i, \delta_i|b_i;\theta) f(y_i|b_i;\theta) \tag{3.5.2}$$

with

$$f(y_i|b_i;\theta) = \prod_{j} f\left\{y_i(t_{ij})|b_i;\theta\right\}, \tag{3.5.3}$$

Because of the fact that the survival and longitudinal sub-models share the same random effects, joint models of this type are also known as shared random-effects models. Under this conditional independence assumptions between longitudinal outcome and time-to-event given the random

effects, b_i , the joint log-likelihood contribution of the i^{th} subject is expressed as

$$log\ f(T_i, \delta_i, y_i; \theta) = log \int f(T_i, \delta_i, y_i, b_i; \theta) db_i$$

$$= log \int f(T_i, \delta_i | b_i; \theta_t, \beta) \left[\prod_j f\left\{ y_i(t_{ij}) | b_i; \theta_y \right\} \right] f(b_i; \theta_b) db_i, \quad (3.5.4)$$

where θ_t, θ_y and θ_b represent the parameters for the survival process, the longitudinal process and the random-effects respectively, $f\left\{ y_i(t_{ij}) | b_i; \theta_y \right\}$ is the density for the longitudinal process and $f(b_i; \theta_b)$ is the density for the random effects. The likelihood of the survival part $f(T_i, \delta_i | b_i; \theta_t, \beta)$ is written as,

$$f(T_i, \delta_i | b_i; \theta_t, \beta) = \left[\lambda_i(T_i | M_i(T_i); \theta_t, \beta) \right]^{\delta_i} S_i(T_i | M_i(T_i); \theta_t, \beta) \quad (3.5.5)$$

where the hazard $\lambda_i(.)$ is given by Equation (3.4.4), and, the survivor function for the i^{th} individual is given by,

$$S_i(t | M_i(t), \omega_i; \theta_t, \beta) = Pr(T_i > t | M_i(t), \omega_i; \theta_t, \beta)$$

$$= exp\left\{ - \int_0^t \lambda_i(s | M_i(s); \theta_t, \beta) ds \right\} \quad (3.5.6)$$

The log-likelihood for the joint model is approximated using the Expectation-Maximisation (EM) algorithm, because both the integral with respect to the random effects in Equation (3.5.4) and in the survival function given by Equation (3.5.6) typically do not have an analytical solution, except in some special cases. Some authors have employed standard numerical integration techniques that have been developed to deal with the intractable integral, such as Gaussian quadrature and Monte Carlo method have been successfully applied in the joint modelling framework (Song *et al.*, 2002; Henderson *et al.*, 2000) which, however, could be very computationally intensive.

However, in this study joint models of longitudinal and a time-to-event outcome was implemented in ***R*** software environment for statistical computing and graphics using the freely available package ***JM*** written by Rizopoulos, (2010). The package has been developed to fit a variety of joint models for longitudinal response and time-to-event data under ML approach, in addition it contains all the methodologies explained above.

3.6 Variables and Model Selection

In order to select the parsimonious model which appropriately fits the given data, it is necessary to compare different models by using different techniques and methods. Hence, the comparison between different models is an important issue in the statistical inference. In the literature on joint modeling of longitudinal and survival data, existing research on model selection is limited.

However, in this study models are compared with Akaike's Information Criterion (AIC), the Bayesian Information Criterion (BIC) and the Likelihood ratio test (LRT) methods for nested models. Accordingly, (Park and Qiu, 2014) defined the popular model assessment criterion for model selection as,

$$AIC = -2log(L_{max}) + kp, \tag{3.6.1}$$

$$BIC = -2log(L_{max}) + klog(n) \tag{3.6.2}$$

where L_{max} is the maximized value of the likelihood function of the model under consideration, and p is the number of parameters in the model and k is constant (often 2). By this criterion, among all candidate models, the one with the smallest AIC and/or BIC value is selected and indicates preferred models. Therefore, the smaller the information criteria value, the better the fit.

Likelihood Ratio Tests (LRTs)

LRTs are a class of tests that are based on comparing the values of likelihood functions for two models (*i.e.*, the nested (null hypothesis) and reference models) defined as

$$-2log\left[\frac{L_{nested}}{L_{refrence}}\right] = -2log\left[L_{nested}\right] - \left[-2log(L_{reference})\right] \sim \chi^2_{df}, \tag{3.6.3}$$

where L_{nested} and $L_{reference}$ denote the ML or REML estimates under the null and alternative hypothesis, respectively. Likelihood theory states that under mild regularity conditions the LRT statistic asymptotically follows a χ^2 distribution, in which the number of degrees of freedom, df, is obtained by subtracting the number of parameters in the nested model from the number of parameters in the reference model (Brady *et al.*, 2007).

When the number of variables is relatively large, it can be computationally expensive to fit

all possible models. In this situation, automatic routines for variable selection that are available in many software packages might seem an attractive prospect.

These routines are based on *forward selection*, *backward elimination* or the combination of the two known as the *stepwise procedure*. The model selection strategy depends to some extent on the purpose of the study. In a situation where the aim is to identify variables upon which the hazard function depends, instead of using the automatic variable selection procedures, the following procedure is recommended.

1. The first step is fitting a univariable model for each of explanatory variables and identifying the variables that are significant at some level from 20% to 25% is recommended in (Hosmer and Lemeshow, 1999).

2. The variables that appear to be important from step 1 are then fitted together in a multivariable model. In the presence of certain variables, others may cease to be important. Consequently, backward elimination is used to omit non-significant variables from the model. Once a variable has been dropped, the effect of omitting each of the remaining variables in turn should be examined.

3. Variables which were not important on their own, and so were not under consideration in step 2, may become important in the presence of others. These variables are therefore added to the model from step 2, with forward selection method. This process may result in terms in the model determined at step 2 ceasing to be significant.

4. A final check is made to ensure that neither insignificant variable is included in the model nor significant variable is excluded from the model. At this stage the interactions between any of the main effects currently in the model can be considered for inclusion if the inclusion significantly modifies the model. For steps 2, 3 and 4 a level of significance around 10% is recommended.

3.7 Models Diagnostics

Diagnostic checking is particularly important. A standard tool to perform model diagnostics are residual graphical methods, as many model checking procedures are based on quantities known as residuals plots, and formal statistical tests. Residuals are values that can be calculated for

each observation and have the feature that their behavior is known, at least approximately, when the fitted model is satisfactory. The following residuals have been proposed for use (Rizopoulos, 2012) in connection with the types of residuals for joint models.

Standardized marginal and Standardized subject-specific residuals

For the longitudinal part of the joint model, two frequently used types of residuals are the standardized marginal and standardized subject-specific residuals, which are defined as

$$r_i^{(ym)} = \hat{V}_i^{-1/2}(y_i - X_i\hat{\beta}), and$$

$$r_i^{(ys)}(t_{ij}) = \{y_i(t_{ij}) - x_i^T(t_{ij})\hat{\beta} - z_i^T(t_{ij})\hat{b}_i\}/\hat{\sigma},$$

where $\hat{\beta}$, $\hat{\sigma}$, and $\hat{D}$ denote the maximum likelihood estimates under model in Equation (3.4.1), $\hat{b}_i$ are the empirical Bayes estimates for the random effects, and $\hat{V}_i = Z_i\hat{D}Z_i^T + \hat{\sigma}^2 I$, with I denoting the identity matrix of appropriate dimensions. The marginal residual $r_i^{(ym)}$ predict the marginal errors $y_i - X_i\beta = Z_i b_i + \varepsilon_{yi}$, and can be used to investigate miss-specification of the mean structure $X_i\beta$ as well as to validate the assumptions for the within-subjects covariance structure V_i. The subject-specific residuals $r_i^{(ys)}(t_{ij})$ predict the conditional errors $\varepsilon_i(t)$, and can be used for checking the homoscedasticity and normality assumptions.

Martingale and Cox-Snell residuals

For the survival part of the joint model, a standard type of residuals is the martingale residuals defined as

$$r_i^{(tm)} = \delta_i - \int_0^{T_i} h_i(s|\hat{M}(s);\hat{\theta})ds.$$

These are commonly used for a direct assessment of excess events (*i.e.*, to reveal subjects that are poorly fit by the model), and for evaluating whether the appropriate functional form for a covariate is used in the model.

Another type of residuals for survival models, related to the martingale residuals, is the Cox-Snell residuals. These are calculated as the value of cumulative risk function evaluated at the observed event times T_i *i.e.*,

$$r_i^{(tcs)} = \int_0^{T_i} h_i(s|\hat{M}(s);\hat{\theta})ds.$$

If the assumed model fits the data well, we expect $r_i^{(tcs)}$ to have a unit exponential distribution; however, when T_i is censored, $r_i^{(tcs)}$ will be censored as well. To take censoring into account in checking the fit of the model, we can compare graphically the Kaplan-Meier estimate of the survival function of $r_i^{(tcs)}$ with the survival function of the unit exponential distribution.

Nevertheless of it being intensively studied for longitudinal and survival analysis, this topic has not received special attention in the joint modelling literature. We therefore, employe graphical methods for assessing the goodness of fit to the given data, and ***R*** packages ***JM*** (Rizopoulos, 2010) was used for technical details.

3.8 Ethical considerations

We obtained a formal ethical letter of permission to collect data for this study from the Addis Ababa University Department of Statistics administration office (***Ref no. Stat42/9/17***). Any personal information regarding study subjects was replaced by a number and patient evidence was kept confidential without disclosing to others during data collection from clinical charts.

Chapter 4

Results and Discussion

4.1 Descriptive Data Analysis

Table 4.1 displays patients characteristics for the HIV/AIDS data from Mekelle General Hospital. Out of the total 469 patients, 288 (61.41%) were females and the remaining 181 (38.59%)

Table 4.1: Baseline characteristics of HIV-infected patients under HAART

Characteristics	Category	No. of patients	Percent (100%)	p-value[a]
Sex	Female	288	61.41	0.1860
	Male	181	38.59	
Functional Status	Ambulatory	106	22.60	$< 0.0001^{\dagger}$
	Bedridden	33	7.04	
	Working	330	70.36	
WHO Clinical Stage	Stage I	45	9.59	$< 0.0001^{\dagger}$
	Stage II	74	15.78	
	Stage III	225	47.97	
	Stage IV	125	26.65	
Regimen Type	AZT-3TC-EFV	76	16.20	$< 0.0001^{\dagger}$
	AZT-3TC-NVP	243	51.81	
	d4T-Based	39	8.32	
	Others	111	23.67	
Status	Censored	376	80.17	
	Event	93	19.83	
Mean Age (Std.dev)			35.11 (8.71)	
Mean Weight (Std.dev)			49.54 (9.05)	
Mean Baseline CD4 (Std.dev)			129.39 (90.64)	

[a] *Log − rank χ^2 test for equality of the groups.*
[†] *Indicates the significance of covariates at 5% level of significance.*

were males. Majority (70.36%) of the infected patients were with working functional status, (*i.e.*, an individual able to perform usual work in and out of the house), followed by those with

ambulatory type of functional status who accounted for 22.60% of the total, and 7.04% were unable to perform activities of daily living (bedridden) patients. Regarding the clinical stage of patients, 45 (9.59%) were at clinical stage I, 74 (15.78%) at clinical stage II, 225 (47.97%) at clinical stage III and the rest 125 (26.65%) were at clinical stage IV when they started HAART.

With respect to the distribution of ART regimens among patients, the data was unbalanced. From Table 4.1 we can see that 319 (68.02%) patients were given ART which are AZT-Based (AZT+3TC+NVP or EFV), only 39 (8.32%) patients were given ART which are d4T-Based (d4T+3TC+EFV or NVP) and the remaining 111 (23.67%) patients were given others combination of drugs. The mean patients' age at enrollment of HAART was 35.11 years with a standard deviation (Std.dev) of 8.71 and the mean weight (Std.dev) at baseline was 49.54 (9.05) kg.

The longitudinal response variable was the number of CD4 cells count per mm^3 of blood which was measured approximately every six months; at the study entry, and again at the six, 12, and 24 months visits. As can be seen from (Table A.1; in Appendix) common measurements used for all patients at these four time points are 469 (100%), 299 (63.75%), 193 (41.15%), and 42 (8.95%) which show a sharply increasing degree of missing data over time. Due to variety of reasons such as deaths, dropouts, missed clinic visits and transferring to other hospitals, only 42 (8.95%) continued up to the end of the study. The average number of baseline CD4 cells count was 129.39 per mm^3 with a standard deviation of 90.64 per mm^3 of blood implying that patients were at higher risk of getting AIDS related illness.

The survival response variable was the length of time from HAART start date until the date of death or censor (measured in months). 80.17% and 19.83% of patients were alive on HAART (Censored) and died due to HIV/AIDS related death, respectively. In the results depicted in Table 4.1[a], the Log-rank test was used to test the difference between the categories of baseline covariates with the probability of death. This test revealed the presence of a statistically significant ($p-value < 0.0001$) difference among the categories of baseline functional status, and WHO clinical stage.

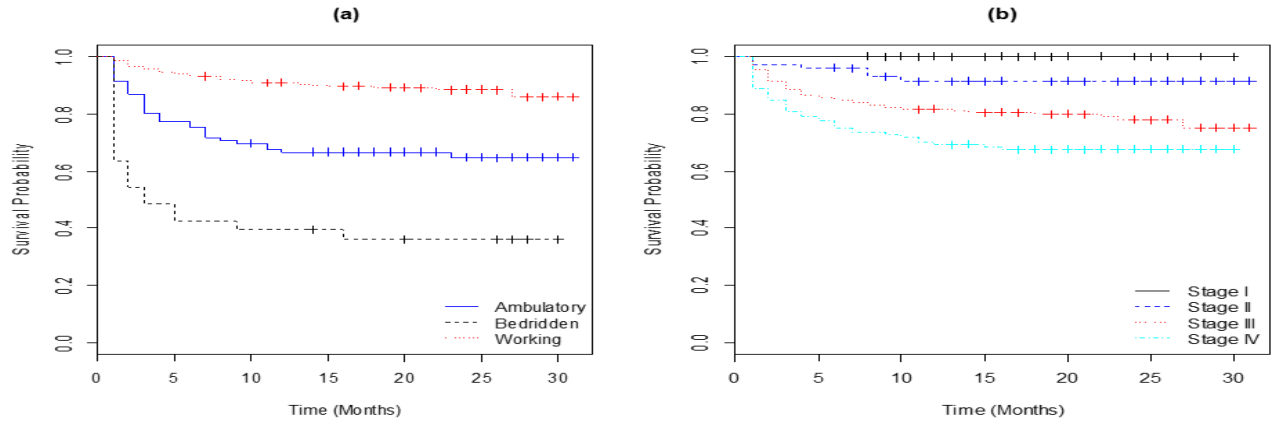

Figure 4.1: Kaplan-Meier Survival plots of Functional Status (a) and WHO clinical stage (b) of HIV-infected patients under HAART

The Kaplan-Meier survival plot (Figure 4.1) also shows a difference between the survival curves. The plots of other baseline covariates are presented in Figure B.6 (Appendix). The plots indicate that female patients had slightly higher survival rate than male patients after six months of follow-up. The survival times are found to be significantly different in regimen type.

4.2 Exploring Individual profile and Mean structure

To check for normality, the basic assumption of linear mixed effects model, histograms, boxplots and normal Q-Q plot of the CD4 cells count with corresponding Shapiro-Wilk and Kolmogorov-Smirnov statistical tests of normality (Figure B.1, B.2, B.3, B.4 and B.5; in Appendix) were conducted. From all these Figures, the CD4 cells count are identified to exhibit right skewed shapes

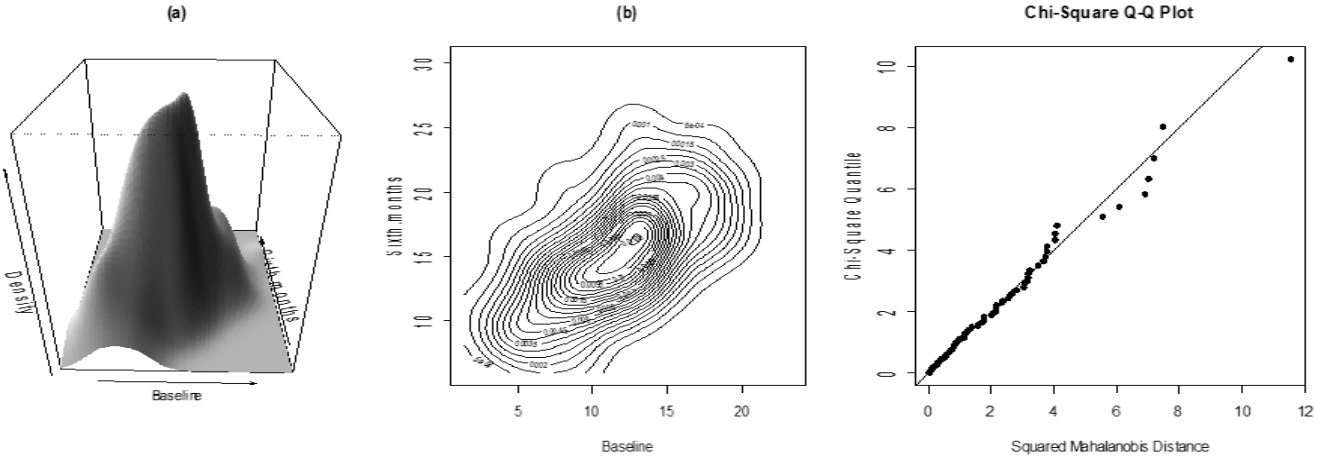

Figure 4.2: Perspective (a) plot, contour (b) plot for randomly selected patients at baseline and six months and Chi-Square Q-Q plot for the square root CD4 cells count

of distribution, the left side (a) of the plot shows a high degree of skewness toward high CD4 cells count plus a significant test means the fit is poor, suggesting some transformation to meet the assumptions. After a square root transformation ($\sqrt{CD4\ cell\ count}$), the right side (b) of the plot attained normality (*i.e.*, the test is not significant $p - value = 0.0795, p - value = 0.8266$) implying that the data set appear to follow a univariate normal distribution.

Korkmaz *et al.*, (2014) demonstrated the three most widely used multivariate normality tests, including *Mardia's*, *Henze-Zirkler's* and *Royston's*, and graphical approaches (*chi-square Q-Q*, *perspective* and *contour* plots). Contour graphs are very useful as they give information about normality and correlation at the same time. Figure 4.2(b) shows the contour plot of the square root CD4 cells count. As can be seen from the graph, this is simply a top view of Figure 4.2(a) the perspective plot where the third dimension is represented with ellipsoid contour lines. From this graph, we can say that there is a positive correlation among the square root CD4 cells count measures of randomly selected patients at baseline and six months since the contour lines lie around the main diagonal. If the correlation were zero, the contour lines would be circular rather than ellipsoid.

Moreover, all three test results (Table A.2; Appendix) as well as their corresponding statistical significance indicate that the data set satisfies approximate multivariate normality assumption at the significance level 0.05, and in agreement with Figure 4.2 the chi-square Q-Q plot, perspective and contour plots as well. Therefore, for the remainder of this analysis we worked with the square root of the CD4 cells count values. Prior to model building, we visualize the pattern

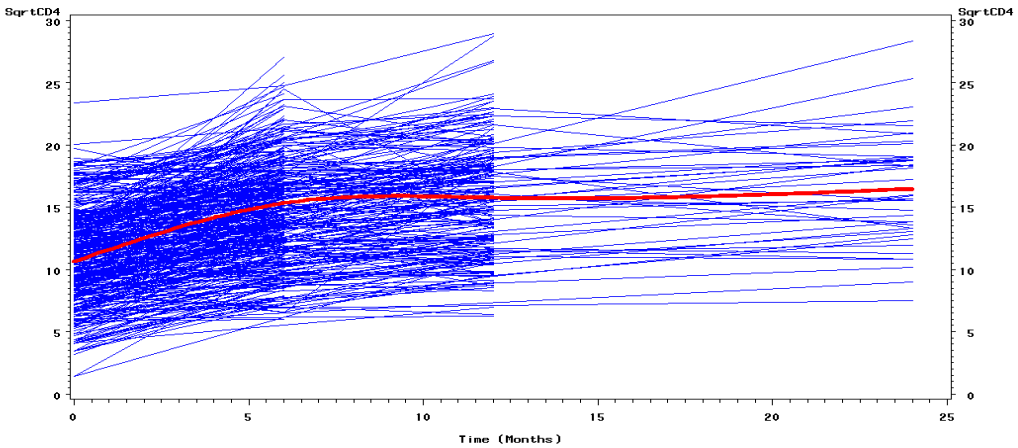

Figure 4.3: Individual Profiles with Average Trend Line.

of the overall individual plots of CD4 cells count measurements of patients overtime. Figure 4.3 demonstrates the variability (within and between patients) in square root of CD4 cells count of HIV-infected patients. Since the measurements were not equally spaced across the different subjects and data is not balanced, loess smoothing technique was used instead. The bold red line loess smoothing technique suggests that the mean structure of the variable is nearly linear (*i.e.*, the relationship between CD4+ cells count and time seems to be linear). The smoothed

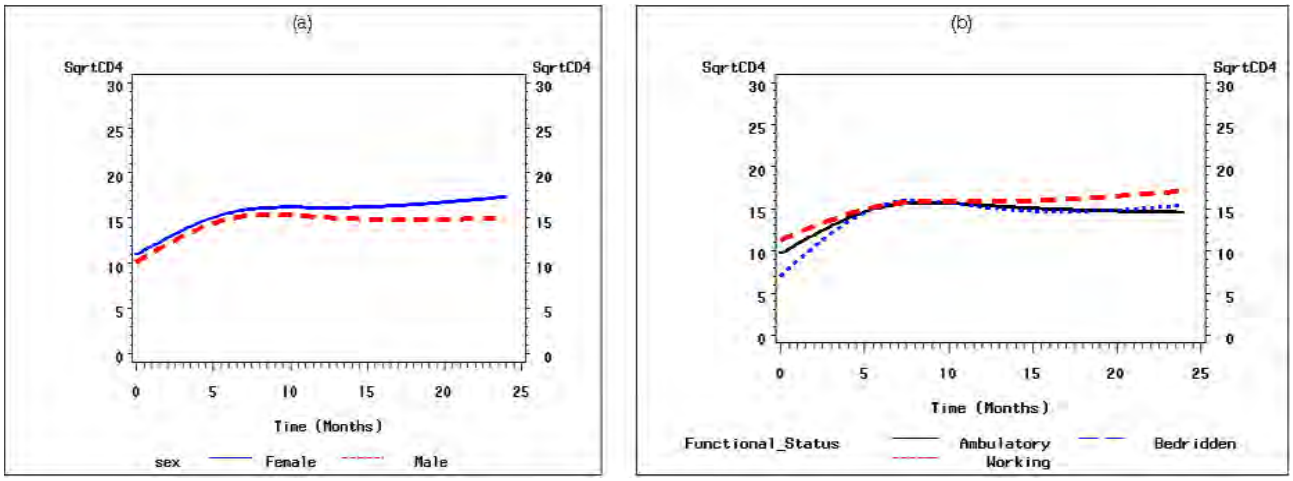

Figure 4.4: The Mean profile plots of Sex (a) and Functional Status (b) for HIV-infected patients under HAART

lines representing the average trend of different subgroups were also displayed. Figure 4.4 depicts the smoothed average evolution by gender (a) and functional status (b) of HIV-infected patients under HAART.

From the Figures, there seems to be a difference between the functional status subgroups at baseline since the smoothed average evaluation of working patients seem to have more CD4 cells count compared to ambulatory and bedridden. The difference in the smoothed lines may indicate a time by functional status interaction. On the other hand, female patients seem to have higher average CD4 cells count as compared to that of male patients at all time points.

The mean profile plots of other baseline covariates are also presented in Figure B.7 (in Appendix). The plots indicate that HIV-infected patients having low CD4 cells count were at higher risk of death and interaction with time as well. Furthermore, the plotted profiles tend to generate a linearly increasing pattern which rationalizes the use of *Linear Mixed Effects model* to analyze the trajectory of CD4 cells count.

4.3 Model Building

4.3.1 The Separate Longitudinal and Survival Analysis

Following from the observations in the exploratory data analysis, good models that best describe the observed average trends and also reflect the observed correlation structures were sought for the data sets.

To identify the appropriate covariance structure three different commonly used covariance structures such as compound symmetry (CS), unstructured (UN) and first order autoregressive (AR(1)) could be considered. In Table 4.2 the smallest values of AIC and BIC for the AR(1)

Table 4.2: Comparison of covariance structure for linear mixed-effects model

	Covariance Structures		
Information Criteria	CS	UN	AR(1)
-2 Res Log Likelihood	-2772.541	-2767.974	-2771.963
AIC	5589.082	5589.948	5587.925
BIC	5696.743	5722.077	5695.586

model suggests that the first order autoregressive (AR(1)) structure best fits our data compared to the remaining covariance structures.

It may be of interest to test whether random intercept and random time effects are both needed after we fixed the correlation structure with AR(1). As such, we implemented different longitudinal sub-models to study the longitudinal outcome by including the subject-specific random effects named as,

1. *Random Intercept Model*:

   In this model, the intercepts are allowed to vary based on patients, and therefore the measurements on the response variable for each individual observation are predicted by the intercept that varies across individuals. The sub-model can be written as:

$$Sqrt(CD4count)_{ij} = \beta_0 + \beta_1 time_i + \beta_2 sex_i + \beta_3 FS_i + \beta_4 WHO_i + \beta_5 RT_i$$

$$+ \ \beta_6 age_i + \beta_7 weight_i + b_{0i} + \epsilon_{ij}, \tag{4.3.1}$$

   where $Sqrt(CD4count)_{ij}$ is the square root of the CD4 cells count of the i^{th} patient at the

j^{th} time, in our case $i = 1 \ldots 469$ patients, $j = 1 \ldots n_i$ observations (max=4) for patient i, time is the *time* that repeated measurements are taken and b_{0i} is the random intercept effect for each patient. However, one of the drawbacks that comes from models with only random intercept is assuming that slopes are fixed. On top of that, random effects for the intercept and linear slope were included to account for the variability between the different subjects. For this reason, the following sub-models were considered.

2. *Random Intercept and Slope Model*:

   Besides considering random intercepts, this model also allows random slopes to vary across subjects. The corresponding sub-models is given by,

$$Sqrt(CD4count)_{ij} = \beta_0 + \beta_1 time_i + \beta_2 sex_i + \beta_3 FS_i + \beta_4 WHO_i + \beta_5 RT_i$$

$$+ \ \beta_6 age_i + \beta_7 weight_i + b_{0i} + b_{1i} t_{ij} + \epsilon_{ij}, \tag{4.3.2}$$

   which additionally incorporate $b_{1i}t_{ij}$ that represents the random slope effect of the different CD4 cells count trajectories of each patient. Furthermore, for the random effects covariance matrix $\boldsymbol{D}$, we set $D_{11} = var(b_{0i})$, $D_{22} = var(b_{1i})$, and $D_{12} = cov(b_{0i}, b_{1i})$.

The summary of the linear mixed-effects models which were modeled by considering different random effects are shown in Table 4.3. Hence, we consider the Model 2 (the model with inter-

Table 4.3: Selection of random effects to be included in the linear mixed-effects model

Model	Random effects	AIC	BIC	logLik	L.Ratio	p-value
1	Only intercept	5617.928	5715.801	-2788.964		
2	Intercept and linear time slope	5587.925	5695.585	-2771.963	34.003	$< 0.0001^{\dagger}$

†*Indicates the significance of covariates at* 5% *level of significance.*

cept and linear time slope) as the most parsimonious model for the separate linear mixed-effects model on the basis of its lower values of AIC, BIC and a significant fit based on the likelihood ratio test as well.

As can be observed from Table 4.4 the fitted linear mixed effects model for preliminary final model containing significant main effects and possible interactions, reveals that sex, baseline functional status, patients' WHO clinical stage at baseline and baseline regimen type were statistically significant ($p - value < 0.05$). Likewise the time by regimen type interaction indicates

that on average the square root CD4 cells count increases with time at 5% level of significance for patients undergoing HAART. In contrast, the mean change in square root of CD4 cells count overtime of baseline age (in years), weight (in kilograms) at enrollment, time by functional status interaction, and time by WHO clinical stage interaction categories were eliminated from the final model based on Hosmer and Lemeshow (1999) recommendation as presented in (Table A.4, and A.5; Appendix). To explore the survival process, we assessed each factor through univariate

Table 4.4: Parameter Estimates, Standard Errors (Std.Err) and 95% CI under the marginal linear-mixed effects analysis with AR(1) covariance structure

Parameter	Estimate	Std.Err	95% CI	p-value
Intercept	9.579	0.526	(8.546 – 10.613)	$< 0.0001^{\dagger}$
Time	0.553	0.053	(0.449 – 0.657)	$< 0.0001^{\dagger}$
Sex				
Female(ref)				
Male	-0.844	0.372	(-1.576 – -0.113)	$0.0237^{\dagger}$
Functional Status				
Ambulatory (ref)				
Bedridden	-2.377	0.783	(-3.916 – -0.838)	$0.0025^{\dagger}$
Working	0.986	0.467	(0.068 – 1.905)	$0.0354^{\dagger}$
WHO Clinical Stage				
Stage IV (ref)				
Stage I	0.809	0.739	(-0.643 – 2.260)	0.2742
Stage II	0.646	0.625	(-0.582 – 1.875)	0.3015
Stage III	0.930	0.468	(0.011 – 1.849)	$0.0474^{\dagger}$
Regimen Type				
Others (ref)				
d4T-Based	0.071	0.723	(-1.349 – 1.491)	0.9219
AZT-3TC-NVP	0.847	0.476	(-0.087 – 1.783)	0.0755
AZT-3TC-EFV	1.348	0.594	(0.179 – 2.516)	$0.0239^{\dagger}$
Time×d4T-Based	-0.075	0.087	(-0.246 – 0.095)	0.3859
Time×AZT-3TC-NVP	-0.151	0.055	(-0.259 – -0.043)	$0.0060^{\dagger}$
Time×AZT-3TC-EFV	-0.133	0.064	(-0.260 – -0.006)	$0.0398^{\dagger}$

†*Indicates the significance of covariates at* 5% *level of significance.*

Cox regression model and found the variables sex, functional status, regimen type and weight at enrollment were statistically significant under separate model analysis (Table A.6, and A.7; in Appendix) and so all were selected to be included in the survival model. In contrast interaction effects, WHO clinical stage and age at enrolment of patients were not found to be significant predictors. Henceforth, the final Cox regression fitted model was:

$$log(time_i) = \beta_1 sex_i + \beta_2 FS_i + \beta_3 RT_i + \beta_4 weight_i \tag{4.3.3}$$

From the results displayed in Table 4.5. It can be seen that sex, functional status and baseline

Table 4.5: Parameter Estimates, Standard Errors (Std.Err) and 95% CI under the survival modeling analysis

Parameter	Estimate	Std.Err	HR (95% CI)	p-value
Sex				
Female(ref)				
Male	0.663	0.225	1.941 (1.248 − 3.018)	0.0032†
Functional Status				
Ambulatory (ref)				
Bedridden	1.098	0.289	2.997 (1.701 − 5.280)	0.0001†
Working	-0.771	0.257	0.462 (0.280 − 0.765)	0.0026†
Regimen Type				
Others (ref)				
d4T-Based	-0.347	0.347	0.707 (0.358 − 1.395)	0.3173
AZT-3TC-NVP	-0.436	0.258	0.647 (0.390 − 1.073)	0.0914
AZT-3TC-EFV	-0.684	0.339	0.505 (0.259 − 0.982)	0.0439†
Weight	-0.058	0.014	0.943 (0.917 − 0.970)	$< 0.0001^{\dagger}$

†*Indicates the significance of covariates at* 5% *level of significance.*

regimen type (AZT-3TC-EFV) and weight are statistically significant at 5% level of significance. Significant lower hazard of death is associated with patients having higher weight after initiation of HAART ($p - value < 0.0001$). That is, higher value of initial weight is associated with a lower mortality.

The estimated risk ratio for sex suggests that the risk of death for male patient is 1.941 times greater than female patient. Likewise, the estimated risks of death for a patient with bedridden functional status compared to ambulatory patient is (HR=2.997, 95% CI: 1.701−5.280) indicating that the hazard rate of death for bedridden patients is around three times higher than ambulatory patients, whereas being working patient reduces the risk for death by about 53.8% compared to ambulatory patient.

4.3.2 The Joint Longitudinal and Survival Analysis

The estimates of the parameters of the separate and joint models are quite similar to each other but not identical. In Table A.3, in the results of the survival process, the parameter labeled `"Assoct"` is in fact parameter α in equation (3.4.4) that measures the effect of $m_i(t)$, where $m_i(t)$ represents the history of the true (*unobserved*) longitudinal response.

The estimate of the association parameter due to the slope (trend) of square root CD4 cells count is negative (-0.134), indicating that CD4 cells count is negatively associated with the risk of death of patients from HAART treatment. This indicates that an increasing trend in the CD4 cells count in patients undergoing HAART treatment significantly reduces the risk of death of those patients. Moreover, the estimate of the association parameter in the joint analysis is

Table 4.6: Parameter Estimates and Standard Errors (Std.Err) under the joint modeling analysis

Survival Process				Longitudinal Process			
Parameter	Value	Std.Err	p-value	Parameter	Value	Std.Err	p-value
				Intercept	9.692	0.523	$< 0.0001^{\dagger}$
				Time	0.493	0.051	$< 0.0001^{\dagger}$
Sex				Sex			
Female(ref)				Female(ref)			
Male	0.490	0.233	$0.0356^{\dagger}$	Male	-0.851	0.371	$0.0219^{\dagger}$
Functional Status				Functional Status			
Ambulatory (ref)				Ambulatory (ref)			
Bedridden	0.735	0.291	$0.0115^{\dagger}$	Bedridden	-2.473	0.780	$0.0015^{\dagger}$
Working	-0.694	0.257	$0.0069^{\dagger}$	Working	1.024	0.466	$0.0281^{\dagger}$
Regimen Type				Regimen Type			
Others (ref)				Others (ref)			
d4T-Based	-0.313	0.351	0.3727	d4t-Based	0.072	0.716	0.9196
AZT-3TC-NVP	-0.365	0.267	0.1713	AZT-3TC-NVP	0.837	0.473	0.0764
AZT-3TC-EFV	-0.560	0.343	0.1023	AZT-3TC-EFV	1.345	0.593	$0.0234^{\dagger}$
Weight	-0.057	0.009	$< 0.0001^{\dagger}$	WHO Clinical Stage			
Assoct (α)	-0.138	0.030	$< 0.0001^{\dagger}$	Stage IV(ref)			
				Stage I	0.885	0.715	0.2160
				Stage II	0.681	0.605	0.2600
				Stage III	0.689	0.456	0.1308
				$T^{\ddagger}\times$d4t-Based	-0.073	0.087	0.4018
				$T^{\ddagger}\times$AZT-3TC-NVP	-0.134	0.053	$0.0123^{\dagger}$
				$T^{\ddagger}\times$AZT-3TC-EFV	-0.122	0.064	$0.0485^{\dagger}$

†*Indicates the significance of covariates at* 5% *level of significance.* ‡*Indicates time.*

significantly different from zero, providing strong evidence of association between the effect of the longitudinal outcome to the risk of an event.

Table 4.7 below shows the covariance parameter estimates for separate and joint modeling analysis. It shows that the highest variability came from the random intercepts in both models. It also shows that the variance of the random intercepts was higher than that of the random slopes. The covariance of random effects is positive, implying that patients with high CD4 cells count tend to have higher estimated coefficient than patients with low CD4 cells count. The

Table 4.7: Covariance parameter Estimates under separate and joint modeling analysis

Separate Analysis		Joint Analysis	
Parameter	Estimate	Parameter	Estimate
$Var(b_{oi})$	7.066	$Var(b_{oi})$	8.793
$Cov(b_{0i}, b_{1i})$	0.245	$Cov(b_{0i}, b_{1i})$	0.202
$Var(b_{1i})$	0.025	$Var(b_{1i})$	0.029
$Var(\varepsilon_i)$	8.570	$Var(\varepsilon_i)$	6.795

residual variability was smaller in joint analysis (6.795) compared to the relative linear mixed effects analysis (8.570) which was probably because the standard errors were adjusted for the correlation between the responses.

4.4 Assessing Models Fit

Once the models are fitted, the next step is to verify if all the necessary model assumptions are valid. In order to check these model assumptions, we often make use of standard types of residuals plots to validate the assumptions behind mixed models and Cox proportional hazard models when these are separately fitted. In order to validate the Cox proportional hazards model assumption of the survival sub-model, a graph of the Schoenfeld residuals was displayed to check the overall goodness-of-fit of our survival sub-models. Figure 4.5 shows that the scaled

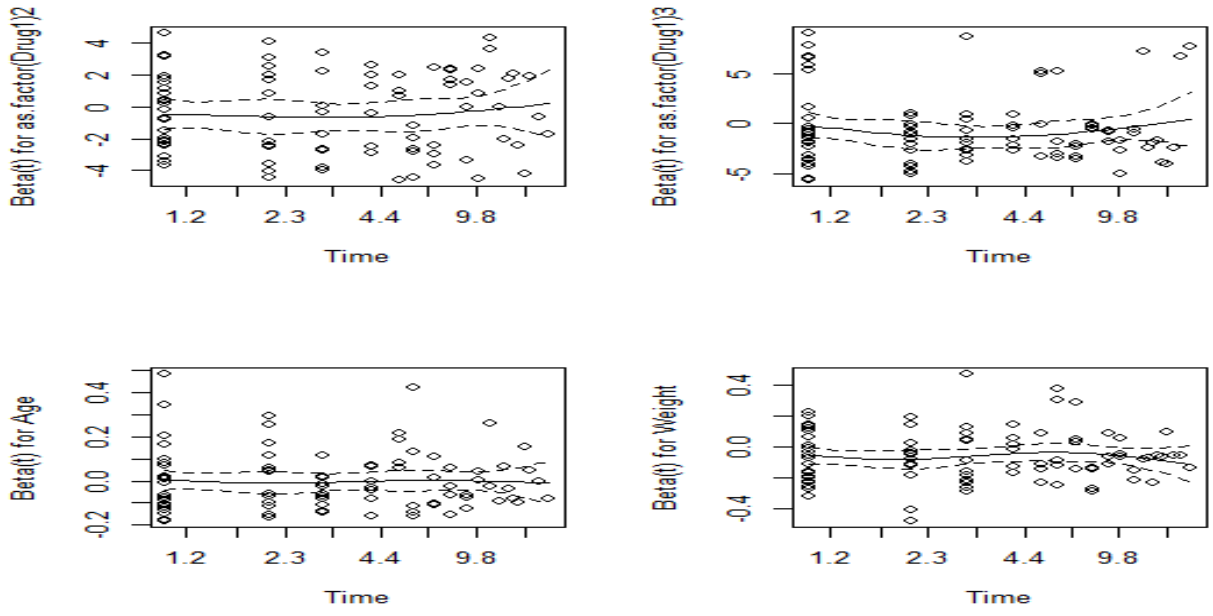

Figure 4.5: Schoenfeld residuals for the survival of patients under HAART

Schoenfeld residuals are randomly distributed and a loess smoothed curve do not exhibit much departure from the horizontal line suggest that the proportional hazards assumption not violated. The proportional hazards assumption was also tested using the interaction of the covariate with the log of survival time (Table A.3; in Appendix). We found that the interaction coefficients are not significant at the 5% level, implying that the assumption of proportionality holds.

Figure 4.6, shows the diagnostic plots for the fitted joint model for HIV-infected patients on HAART follow-up. The top left panel depicts the subject-specific residuals for the longitudinal

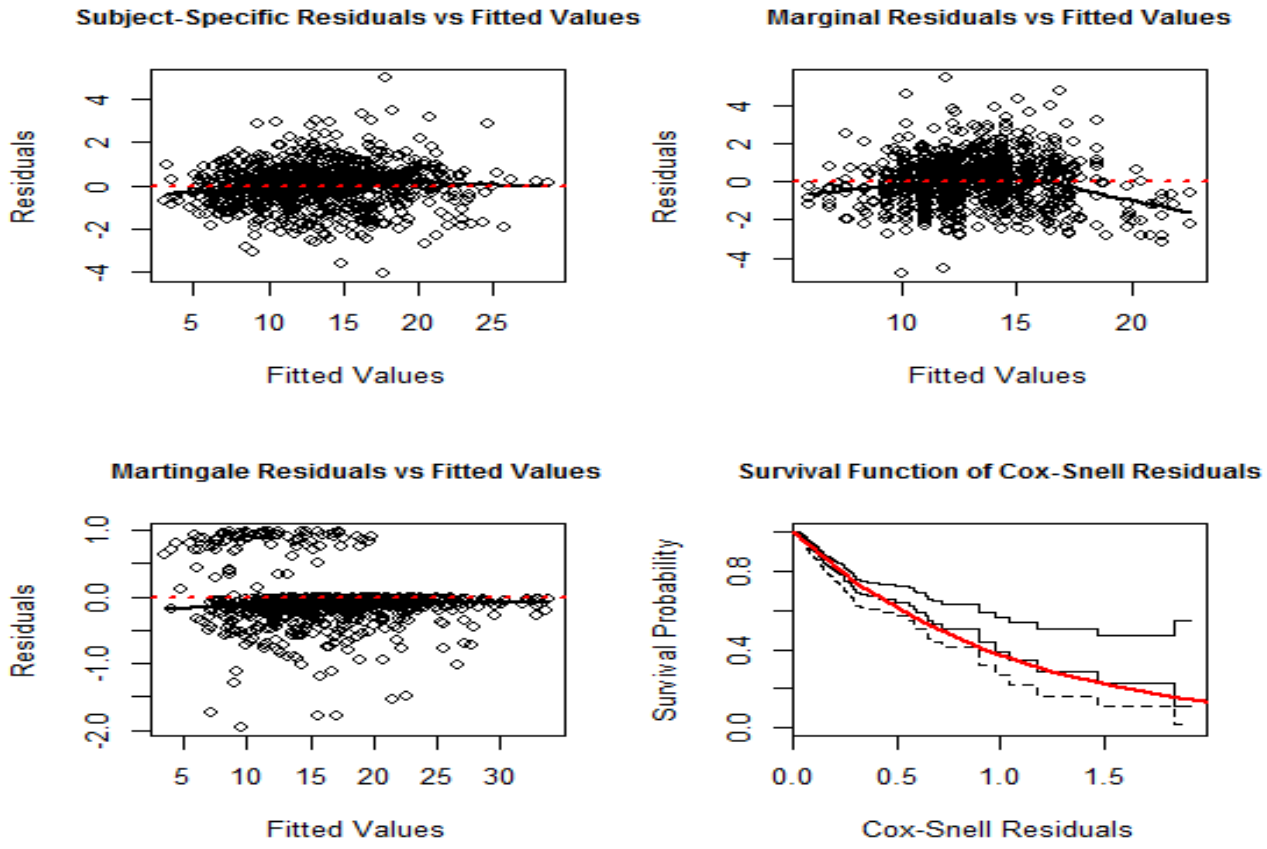

Figure 4.6: Diagnostic plots for the fitted joint model for HIV-infected patients under HAART

measurements on the square root of CD4 cells count plotted against their corresponding fitted values and the residuals can be seen to be trending very close along the fitted line. The top right panel shows a marginal residual versus fitted values plot of the standardized residuals for longitudinal process which is almost coinciding with the reference line passing through the origin and hence, is validating our assumption of normality of the error term in the longitudinal sub-model.

The bottom left plot of the estimated martingale residuals versus the subject-specific fitted values of the survival process shows no much deviations from the null horizontal line.

The diagnostic based on Cox-Snell residuals bottom right panel denote the 95% pointwise CI for the Kaplan-Meier estimate of the Cox-Snell residuals along the red line. The survival function of the unit exponential distribution, indicates that the survival function of the standard exponential distribution lies within the 95% CI of the Kaplan-Meier estimate. This indicate the survival process model fits the data well.

4.5 Interpretation and Discussion of the results

4.5.1 Interpretation of the results

A joint model and the corresponding independent sub-models were built using a retrospective cohort data obtained from HIV+ patients on HAART at Mekelle General Hospital, Tigray, Ethiopia to show the benefits of joint modeling when both the longitudinal and survival processes are associated through shared random-effects. In our study we addressed the relationship between the CD4 cells count over time and the risk of death among HIV patients using joint modeling with a longitudinal linear mixed effects sub-model and a Cox proportional hazards survival sub-model.

The results from the Cox proportional hazards model are in line with the results obtained from the corresponding survival process in joint model. In the Cox model, the relative hazard rate for patients weight is exp(-0.058) = 0.943 as compared to 0.944 under the joint model. However, the joint model shows a reduction in the standard errors when compared to Cox model. This indicates that the results of both the separate and joint analyses are consistent.

The estimated association parameter (α) in the joint model is -0.138 corresponding with the (95% CI: -0.196 − -0.079) and statistically significant ($p - value < 0.0001$). This indicates that there is strong evidence of association between the effect of the longitudinal outcome to the risk of an event, implying higher initial values of the CD4 cells associated with a better survival.

The estimated coefficient for gender is negative and significantly different from zero in both separate and joint models, suggesting that male patients had lower CD4 cells count than females during the follow up. In addition, being male is associated with a risk of death (HR=1.941, 95% CI: $1.248 - 3.018$) that is 1.941 times the risk of death in females holding others covariates in the model constant.

On the other hand, we observed that being working patient reduces the risk for death by about 53.8% compared to ambulatory patient, and a unit increase in weight significantly reduces the risk of death by about 5.7% keeping other variables constant.

4.5.2 Discussion of the results

In this study, three different models were explored, the linear mixed effects model, Cox proportional hazards model for each outcome independently, and joint modeling of the two outcomes together. All approaches ended up with similar results except that the joint analysis added up another information about the association between the two responses. In the separate analysis of the longitudinal data, the square root transformation CD4 cells count measurements were used to meet the normality assumption.

In the first few months (0 to 7 months) after HAART initiation date, an increase in CD4 cells count was observed and then average stable level was noted from eight months till the end of the study period. CD4 cells count was found evolving differently between women and men patients based on the result from the two models, (*i.e.*, separate and joint models). The evolution level was higher for female patients compared to males. This result also conforms to the result obtained by Gurprit *et al.* (2015) whereby female patients had higher CD4 cells count than males during the follow up.

Our study was also in agreement with the studies of (Lim *et al.*, 2013) and (Ibrahim *et al.*, 2010) in showing the significance of the shared parameter that links the two processes, and the reduction in the standard error of the parameter estimates when compared to independent model estimates. This suggesting the need for a joint analysis of this data compared to the use of independent models.

Furthermore, these studies including (Seid *et al.*, 2014) have supported joint modeling of longitudinal data and survival time-to-event process over separate modeling, which has been again emphasized upon by the results of our study which has shown a very significant association between the longitudinal CD4 cells count and the time-to-event.

A limitation of this study is the short duration of follow-up time, which might affect the estimates of the covariates. In the data, the median follow-up time was 21 months and only 19.83% of the study patients died with censoring rate of about 80.17%. When the follow-up duration is not long enough, it has an impact on the number of CD4+ cells measurements, possibly leading to less reliable estimation of the random effect model (Kenward and Rosenkranz, 2011). So one future extension of this work could possibly be to account for the missing data.

Moreover, we have used only one HAART centre retrospective cohort data for analysis that may not be representative for the whole country. Considering the socio-economic and demographic diversity of Ethiopia, our results need to be substantiated by similar studies from other parts of Ethiopia to raise up a comprehensive picture of HIV/AIDS in Ethiopia.

Chapter 5

Conclusion and Recommendation

5.1 Conclusion

In conclusion, when the longitudinal and survival processes are correlated, valid inferences can be made through the use of a joint modeling approach. This has been demonstrated using Mekelle General Hospital retrospective cohort HIV/AIDS data. In the longitudinal sub-model, the predictors: sex, functional status, WHO clinical stage, baseline regimen type, and time by regimen type interaction were statistically significant at 5% level of significance. For the survival sub-model, sex, functional status, regimen type and initial weight were important factors which have significant effect on time to death.

The results of both the separate and joint analyses are consistent. However, the use of a joint analysis compared to independent models adjusted for the correlation between the responses which indicates that more adequate and efficient inferences can be made using joint model estimates. This means that joint modeling can benefit the analyses of both longitudinal biomarker and survival time-to-event data outcomes.

5.2 Recommendation

It is recommended that further studies of this nature include other important covariates that were not included in this study. Such covariates include: viral load results, opportunistic infections, socio-economic status, marital status, level of education and many others.

Having lower CD4 cells count, lower initial weight, late WHO clinical stages, being ambulatory and bedridden are associated with higher risk of death and are indicators of the progression of the disease. Therefore, patients should be informed about the need for early diagnosis of HIV infection and starting treatment early is very important as per the recent WHO 'treat all' recommendation.

Finally, health workers and data clerks working with patients under HAART should be given special and continuous training to improve the quality of the data records of patients. Moreover, future extension of this work could possibly be to account for missing data, and mechanisms should be devised to trace patients who lost to follow up.

References

Aboma Temesgen and Teshome Kebede (2016). Joint Modeling of Longitudinal CD4 Count and Weight Measurements of HIV/Tuberculosis Co-infected Patients at Jimma University Specialized Hospital. *Ann. Data. Sci.*, **3**(3): 321-338.

Andrinopoulou ER (2014). Joint Modeling of Longitudinal and Survival Data with Applications in Heart Valve Data. *Ph.D. Dissertation.*

Barry O., Powell J., *et al.* (2013). Effectiveness of first-line antiretroviral therapy and correlates of longitudinal changes in CD4 and viral load among HIV-infected children in Ghana. *BMC Infectious Diseases*, **13**(476): 1-10.

Brady TW., Kathleen B., Welch AT., and Gal Ecki. (2007). *Linear Mixed Models. A Practical Guide Using Statistical Software.* Boca Raton:Chapman & Hall/CRC.

Brown ER, Ibrahim JG, and DeGruttola V. (2005). A flexible B-spline model for multiple longitudinal biomarkers and survival. *Biometrics*, **61**: 64-73.

Chi Y. and Ibrahim JG. (2007). A new class of joint models for longitudinal and survival data accommodating zero and non-zero cure fractions: A case study of an International Breast Cancer Study Group trial. *Statistica Sinica*, **17**: 445-462.

Cox D.R. Regression models and life-tables (1972). *Journal of the Royal Statistical Society, Series B*, **34**(2): 187-202.

Central Statistical Agency (2007). *Ethiopia Census report 2007.* Addis Ababa, Ethiopia.

Diggle P.J., Heagetry P.J., Liang K.Y and Zeger S.L. (2002). *Analysis of Longitudinal Data.* (2^{nd} Ed.). Oxford Science Publications. Oxford: Clarendon Press.

Diggle P.J., Sousa I. and Chetwynd A.G. (2008). Joint modelling of repeated measurements and time-to-event outcomes: The 4^{th} Armitage lecture. *Stat. Med.*, **27**: 2981-2998.

Dimitris Rizopoulos, Laura A. Hatfield, Bradley P. Carlin and Johanna J. M. Takkenberg (2014). Combining Dynamic Predictions From Joint Models for Longitudinal and Time-to-Event Data Using Bayesian Model Averaging, *Journal of the American Statistical Association*, **109**(508): 1385-1397.

Fares Qeadan (2016). Longitudinal Data Analysis by Example. A seminar in biostatistics for the Mountain West Clinical Translational Research Infrastructure Network. *University of New Mexico Health Sciences Center. Albuquerque, New Mexico.*

Fitzmaurice G., Laird N.M., and Ware J.H. (2004). *Applied Longitudinal Data Analysis.* Wiley, New York.

Gemeda Bedaso, Ayele Taye and Hailemichael M. (2015). Bayesian Joint Modelling of Disease Progression Marker and Time to Death Event of HIV/AIDS Patients under ART Follow-up. *British Journal of Medicine & Medical Research* , **5**(8): 1034-1043.

Geretti AM, Smith C, Haberl A, *et al.* (2008). Determinants of virological failure after successful viral load suppression in first-line highly active antiretroviral therapy, *Antivir Ther*, **137**: 927-36.

Gesesew HA, Ward P, Hajito KW, Feyissa GT, Mohammadi L. and Mwanri L. (2017). Discontinuation from Antiretroviral Therapy: A Continuing Challenge among Adults in HIV Care in Ethiopia: A Systematic Review and Meta-Analysis. *PLoS ONE*, **12**(1): 1-19.

Guo X. and Carlin B.(2004). Separate and joint modeling of longitudinal and event time data using standard computer packages. *The American Statistician*, **58**: 16-24.

Gurprit G., Prafulla K.S., Vishal D. and Manoj K.V. (2015). A Joint Modeling Approach to Assess the Impact of CD4 Cell Count on the Risk of Loss to Follow up in HIV/AIDS Patients on Antiretroviral Therapy. *International Journal of Statistics and Applications*, **5**(3): 99-108.

Henderson R., Diggle P. and Dobson A. (2000). Joint Modelling of Longitudinal Measurements and Event Time Data. *Biostatistics*, **1**: 465-480.

Hosmer D.W and Lemeshow S. (1999). *Applied Survival Analysis Regression Modelling of Time to Event Data*, John Wiley and Sons, Inc. New York.

Hsieh F., Tseng YK. and Wang JL. (2006). Joint Modeling of Survival and Longitudinal Data: Likelihood Approach Revisited. *Biometrics*, **62**: 1037-1043.

Ibrahim J., Chu H., and Chen L. (2010). Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology*, **28**(16): 2796-2801.

Ickovics JR., and Meade CS. (2002). Adherence to HAART among patients with HIV: breakthroughs and barriers. *AIDS Care*, **14**: 309-318.

Jue W., Sheng L. and Liang L. (2016). Dynamic Prediction for Multiple Repeated Measures and Event Time Data: An Application to Parkinsons Disease. *Stat App*, 1-31 available at **URL**: *https://arxiv.org/abs/1603.06476v1*

Kaplan, E. and Meier, P. (1958). Nonparametric estimation for incomplete observations. *Journal of the American Statistical Association*, **93**: 457-481.

Kenward MG and Rosenkranz GK. (2011). Joint modeling of outcome, observation time, and Missingness. *J Biopharm Stat*, **21**: 252-262.

Klein J.P. and Moeschberger M.L. (2003). *Survival Analysis Techniques for Censored and Truncated Data*. Springer.

Korkmaz S., Goksuluk D., and Zararsiz G. (2014). **MVN**: An R Package for Assessing Multivariate Normality. *The R Journal*, 2014, **6**(2): 151-162, available at **URL**: *https://journal.r-project.org/archive/2014-2/RJ-2014-2.pdf*

Laird N.M. and Ware J.H. (1982). Random-Effects Models for Longitudinal Data. *Biometrics*, **38**: 963-974.

Lange Kenneth (2004). *Mathematical Analysis and Optimization*. Springer-Verlag, New York.

Lim HJ, Mondal P. and Skinner S. (2013). Joint modeling of longitudinal and event time data:application to HIV study. *J Med Stat Inform.*, **1**:1-9 available at **URL**: *http://dx.doi.org/10.7243/2053-7662-1-1*

Martins R., Silva G.L. and Andreozzi V. (2010). Joint analysis of longitudinal and survival AIDS data in Brazil. *METMAV International Workshop on Spatio-Temporal Modeling*. Santiago de Compostela.

Mataftsi M., Skoura L. and Sakellari D. (2011). HIV infection and periodontal diseases:an overview of the post-HAART era. *Oral Diseases*, **17**(1): 13-25.

Melaku YA, and Zeleke EG (2014). Contraceptive Utilization and Associated Factors among HIV Positive Women on Chronic Follow Up Care in Tigray Region, Northern Ethiopia: A

Cross Sectional Study. *PLoS ONE*, **9**(4): e94682. available at **URL**: *https://doi.org/10.1371/journal.pone.0094682*

Moore Richard D. (2011). Epidemiology of HIV Infection in the United States: Implications for Linkage to Care. *Clinical Infectious Diseases*, **52**(S2): S208-S213 available at **URL**: *https://doi.org/10.1093/cid/ciq044*

National AIDS resource center. National Factsheet 2010, available at **URL**: *http://www.etharc.org/resources/healthstat/nationalfactsheet/13-nationalfactsheet2010.*

Neuhaus A, Augustin T, Heumann C. and Daume M. (2009). A Review on Joint Models in Biometrical Research. *Journal of Statistical Theory and Practice*, **3**: 855-868.

Palella FJ, Baker Moorman AC *et al.*, (2006). Mortality in the highly active antiretroviral therapy era: changing causes of death and disease in the HIV outpatient study. *Journal Acquired Immune Deciency Syndrome*, **43**: 27-34.

Park KY. and Qiu P. (2014). Model Selection and Diagnostics For Joint Modeling of Survival and Longitudinal Data with Crossing Hazard Rate Functions. *Statistics in Medicine*, **33**(26): 4532-4546.

Philipson P, Sousa I, Diggle P, Williamson P, Kolamunnage-Dona R, and Henderson R (2012). ***joineR****: Joint Modelling of Repeated Measurements and Time-to-Event Data*. R package version 1.0-3, available at **URL**: *https://CRAN.R-project.org/package=joineR.*

Rizopoulos, D. (2010). **JM**: an R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, **35**(9): 1-33.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data with Applications in R*. Boca Raton: Chapman & Hall/CRC Biostatistics Series.

Rizopoulos D (2012). ***JM****: Joint Modeling of Longitudinal and Survival Data*. R package version 1.1-0, available at **URL**: *https://CRAN.R-project.org/package=JM.*

Sawe FK, and McIntyre JA. (2009). Monitoring HIV antiretroviral therapy in resource limited settings: time to avoid costly outcomes. *Clin Infect Dis*, **49**(3): 463465.

Seid A., Getie M., Birlie B. and Getachew Y. (2014). Joint modeling of longitudinal CD4 cell counts and time-to-default from HAART treatment:a comparison of separate and joint models. *Electronic Journal of Applied Statistical Analysis*, **07**(2): 292-314.

Song X, Davidian M, and Tsiatis A. (2002). A Semiparametric Likelihood Approach to Joint Modeling of Longitudinal and Time-to-Event Data. *Biometrics*, **58**: 742-753.

Sousa I. (2011). A Review on Joint Modelling of Longitudinal Measurements and Time-to-event. *REV-STAT*, **9**(1): 57-81.

The U.S. Department of Health and Human Services (HHS), Washington, DC. Panel on Antiretroviral Guidelines for Adults and Adolescents (2011). *Guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents.*

Tsiatis A.A. and Davidian M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, **14**: 809-834.

UNAIDS, Joint United Nations Programme on HIV/AIDS, Global AIDS Update − 2016, June 2016, available at **URL**: *http://www.unaids.org/sites/default/files/media_asset/global-AIDS-update-2016_en.pdf* [accessed on 5 December 2016]

Verbeke G. and Molenberghs G. (2000). *Linear Mixed Models for Longitudinal Data.* Springer Science & Business Media.

Vonesh E., Greene T. and Schluchter M. (2006). Shared parameter models for the joint analysis of longitudinal data and event times. *Statistics in Medicine*, **25**: 143-163.

Wang H, Wolock TM, Carter A, Nguyen G, Kyu HH, *et al.*, (2016). Estimates of global, regional, and national incidence, prevalence, and mortality of HIV, 1980-2015:The Global Burden of Disease Study 2015. *The Lancet HIV*, **3**(8): e361-e387 available at **URL**: *http://dx.doi.org/10.1016/S2352-3018(16)30087-X*

Wang Y. and Taylor JMG. (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of American Statistical Association*, **96**: 895-905.

Zhang D, Chen MH, Ibrahim JG, Boye ME, Wang P. and Shen W. (2016). JMFit: A SAS Macro for Joint Models of Longitudinal and Survival Data. *Journal of Statistical Software*, **71**(3): 1-24 available at **URL**: *http://dx.doi.org/10.18637/jss.v071.i03*

Appendix A

Appendix A: Summary Results for Selected Tables

Table A.1: Summary measures of Square root CD4 cells count at each time points with respective Sample sizes, Mean and Standard deviation

Measurements Time	No. of patients (100%)	Mean	Std.dev
Baseline	469 (100%)	10.626	4.065
6 months	299 (63.75%)	15.374	4.234
12 months	193 (41.15%)	15.856	4.666
24 months	42 (8.95%)	16.481	4.465

Table A.2: Result of Multivariate Normality test for Square root CD4 cells count

Test	Test Statistic	p-value
Mardia		
Skewness	7.113	0.752
Kurtosis	20.436	0.354
Henze-Zirkler	0.568	0.772
Royston	2.124	0.658

Table A.3: Parameter Estimates, Standard Errors (Std.Err) and 95% CI for the Cox proportional hazards model with interaction of the covariate by the log of survival time

Variable	Estimate	Std.Err	HR (95% CI)	p-value
log(Time)×Male	-0.016	0.023	0.009 (0.001 − 2.583)	0.5059
log(Time)×Bedridden	0.186	0.462	1.205 (0.487 − 2.979)	0.6865
log(Time)×Working	-0.309	0.298	7.342 (0.410 − 1.316)	0.2993
log(Time)×d4T-Based	0.173	0.452	1.188 (0.490 − 2.880)	0.7023
log(Time)×AZT-3TC-NVP	-0.048	0.301	0.953 (0.528 − 1.719)	0.8726
log(Time)×AZT-3TC-EFV	0.152	0.360	1.165 (0.575 − 2.358)	0.6717
log(Time)×Weight	-0.022	0.015	0.978 (0.949 − 1.008)	0.1577

Table A.4: Univariable linear-mixed effects model for HIV-infected patients under HAART

Variable	Estimate	Std.Err	95% CI	p-value
Sex				
Female(ref)				
Male	-0.779	0.371	(-1.509 − -0.0488)	$0.0366^{\dagger}$
Functional Status				
Ambulatory (ref)				
Bedridden	-2.523	0.777	(-4.050 − -0.996)	$0.0013^{\dagger}$
Working	1.473	0.428	(0.633 − 2.314)	$0.0006^{\dagger}$
WHO Clinical Stage				
Stage IV (ref)				
Stage I	2.162	0.513	(1.154 − 3.170)	$< 0.0001^{\dagger}$
Stage II	1.943	0.565	(0.832 − 3.053)	$0.0006^{\dagger}$
Stage III	1.475	0.460	(0.570 − 2.380)	$0.0015^{\dagger}$
Regimen Type				
Others (ref)				
d4T-Based	-0.100	0.722	(-1.535 − 1.335)	0.8912
AZT-3TC-NVP	1.326	0.450	(0.442 − 2.211)	$0.0034^{\dagger}$
AZT-3TC-EFV	1.062	0.582	(-0.082 − 2.205)	0.0687
Age	-0.034	0.021	(-0.075 − 0.007)	0.1020
Weight	0.042	0.020	(0.003 − 0.082)	$0.0355^{\dagger}$

†*Indicates the significance of covariates at* 5% *level of significance.*

Table A.5: Multivariable linear-mixed effects model containing Main effects and Interaction

Variable	Estimate	Std.Err	95% CI	p-value
Intercept	9.579	0.526	(8.546 − 10.613)	$< 0.0001^{\dagger}$
Time	0.553	0.053	(0.449 − 0.657)	$< 0.0001^{\dagger}$
Sex				
Female(ref)				
Male	-0.844	0.372	(-1.576 − -0.113)	$0.0237^{\dagger}$
Functional Status				
Ambulatory (ref)				
Bedridden	-2.377	0.783	(-3.916 − -0.838)	$0.0025^{\dagger}$
Working	0.986	0.467	(0.068 − 1.905)	$0.0354^{\dagger}$
WHO Clinical Stage				
Stage IV (ref)				
Stage I	0.809	0.739	(-0.643 − 2.260)	0.2742
Stage II	0.646	0.625	(-0.582 − 1.875)	0.3015
Stage III	0.930	0.468	(0.011 − 1.849)	$0.0474^{\dagger}$
Regimen Type				
Others (ref)				
d4T-Based	0.116	0.725	(-1.309 − 1.542)	0.8730
AZT-3TC-NVP	0.743	0.477	(-0.195 − 1.682)	0.1203
AZT-3TC-EFV	1.325	0.598	(0.150 − 2.500)	$0.0271^{\dagger}$
Age	-0.014	0.021	(-0.056 − 0.027)	$0.4950^{\ddagger}$
Weight	0.018	0.022	(-0.025 − 0.060)	$0.4153^{\ddagger}$
Time×Bedridden	0.059	0.105	(-0.148 − 0.266)	$0.5769^{\ddagger}$
Time×Working	0.011	0.054	(-0.094 − 0.117)	$0.8359^{\ddagger}$
Time×Stage I	0.021	0.074	(-0.124 − 0.166)	$0.7766^{\ddagger}$
Time×Stage II	0.021	0.067	(-0.111 − 0.154)	$0.7505^{\ddagger}$
Time×Stage III	-0.083	0.050	(-0.181 − 0.015)	$0.0983^{\ddagger}$
Time×d4T-Based	-0.075	0.087	(-0.246 − 0.095)	0.3859
Time×AZT-3TC-NVP	-0.151	0.055	(-0.259 − -0.043)	$0.0060^{\dagger}$
Time×AZT-3TC-EFV	-0.133	0.064	(-0.260 − -0.006)	$0.0398^{\dagger}$

†*Indicates the significance of covariates at* 5% *level of significance.*
‡*Indicates that, the variables excluded from the model for good at this step.*

Table A.6: Univariable Cox proportional hazards model for HIV-infected patients under HAART

Variable	Estimate	Std.Err	HR (95% CI)	p-value
Sex				
Female(ref)				
Male	0.271	0.208	1.311 (0.872 − 1.973)	0.19340
Functional Status				
Ambulatory (ref)				
Bedridden	0.940	0.276	2.560 (1.491 − 4.394)	0.0006†
Working	-1.241	0.236	0.290 (0.182 − 0.459)	$< 0.0001^{\dagger}$
WHO Clinical Stage				
Stage IV (ref)				
Stage I	-0.938	0.303	0.391 (0.216 − 0.709)	0.0019†
Stage II	-1.454	0.438	0.234 (0.099 − 0.551)	0.0009†
Stage III	-0.524	0.238	0.592 (0.371 − 0.943)	0.0274†
Regimen Type				
Others (ref)				
d4T-Based	-0.227	0.797	0.780 (0.406 − 1.562)	0.50853
AZT-3TC-NVP	-1.056	0.348	0.348 (0.216 − 0.558)	$< 0.0001^{\dagger}$
AZT-3TC-EFV	-0.796	0.451	0.451 (0.240 − 0.849)	0.01362†
Age	0.005	0.012	1.005 (0.982 − 1.029)	0.65710
Weight	-0.076	0.013	0.927 (0.903 − 0.951)	$< 0.0001^{\dagger}$

†*Indicates the significance of covariates at* 5% *level of significance.*

Table A.7: Multivariable Cox proportional hazards model for HIV-infected patients under HAART

Variable	Estimate	Std.Err	HR (95% CI)	p-value
Sex				
Female(ref)				
Male	0.690	0.235	1.994 (1.259 − 3.158)	0.0032†
Functional Status				
Ambulatory (ref)				
Bedridden	1.152	0.298	3.164 (1.766 − 5.670)	$< 0.0001^{\dagger}$
Working	-0.824	0.267	0.438 (0.260 − 0.739)	0.0019†
WHO Clinical Stage				
Stage IV (ref)				
Stage I	0.122	0.340	1.130 (0.580 − 2.199)	0.7193‡
Stage II	-0.105	0.480	0.900 (0.351 − 2.306)	0.8262‡
Stage III	0.320	0.267	1.378 (0.815 − 2.326)	0.2312‡
Regimen Type				
Others (ref)				
d4T-Based	-0.389	0.350	0.677 (0.341 − 1.345)	0.2659
AZT-3TC-NVP	-0.455	0.260	0.634 (0.381 − 1.056)	0.0802
AZT-3TC-EFV	-0.703	0.341	0.495 (0.254 − 0.965)	0.0389†
Age	-0.001	0.012	0.999 (0.976 − 1.023)	0.9239‡
Weight	-0.057	0.014	0.944 (0.917 − 0.971)	$< 0.0001^{\dagger}$

†*Indicates the significance of covariates at* 5% *level of significance.*
‡*Indicates that, the variables excluded from the model for good at this step.*

Appendix B

Appendix B: Summary Results for Selected Figures

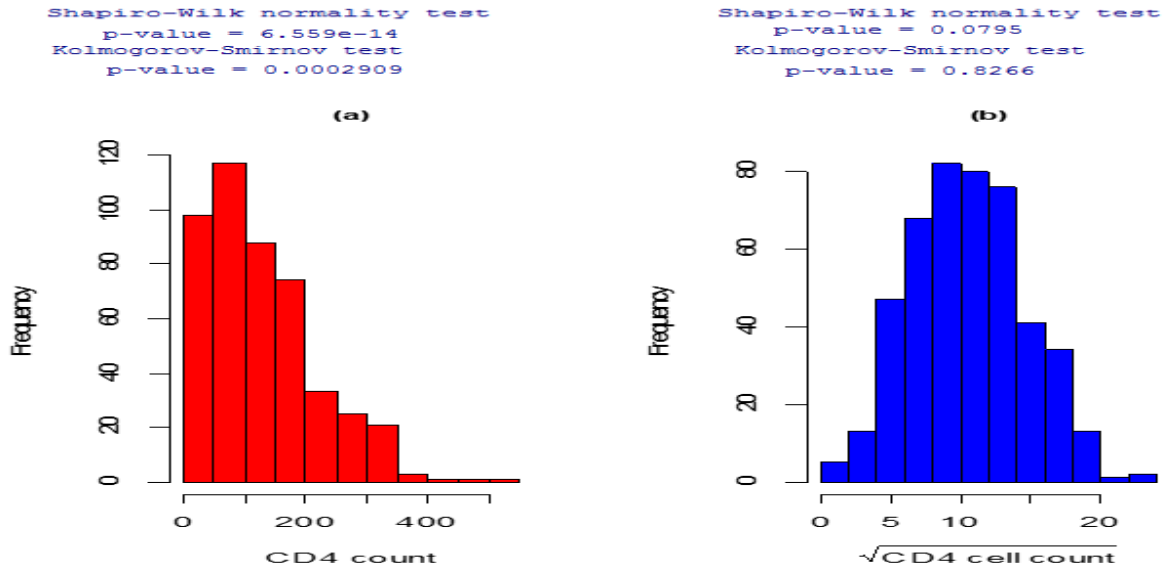

Figure B.1: Histogram of the actual CD4 cells count (a) and the square root CD4 cells count (b) at Baseline

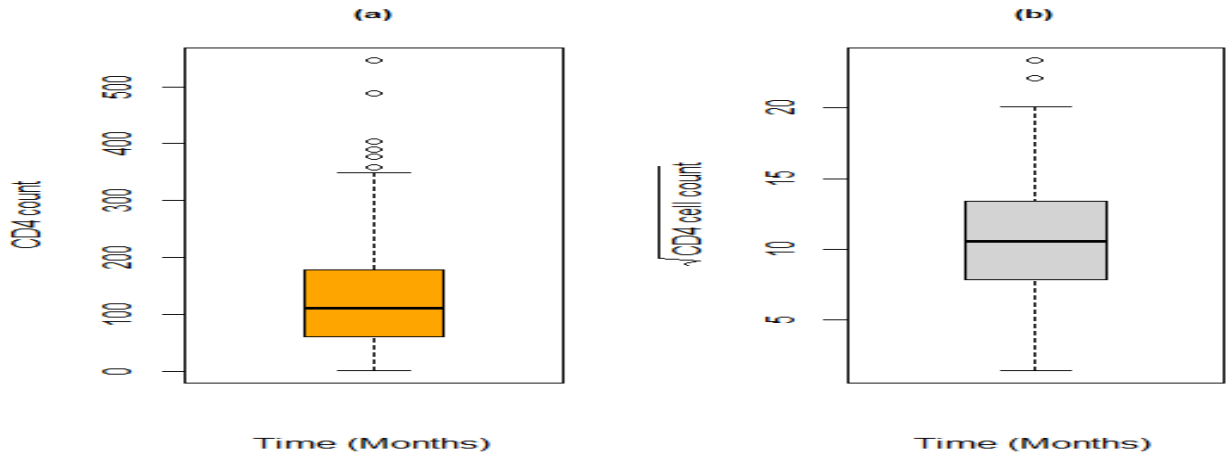

Figure B.2: Boxplots of the actual CD4 cells count (a) and the square root CD4 cells count (b) at Baseline

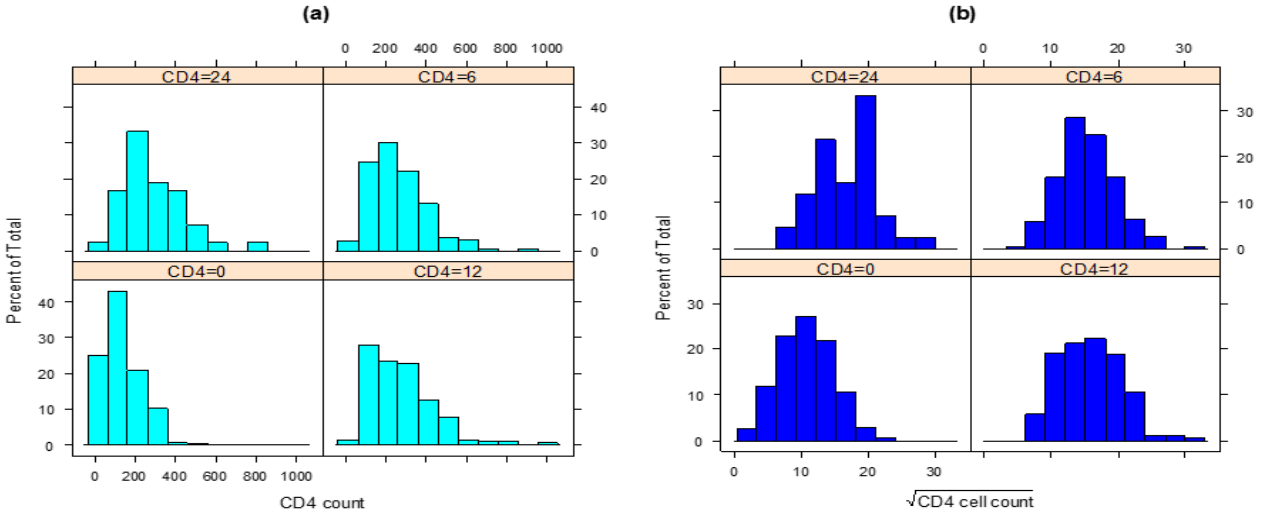

Figure B.3: Histogram of the actual CD4 cells count (a) and the square root CD4 cells count (b) over time

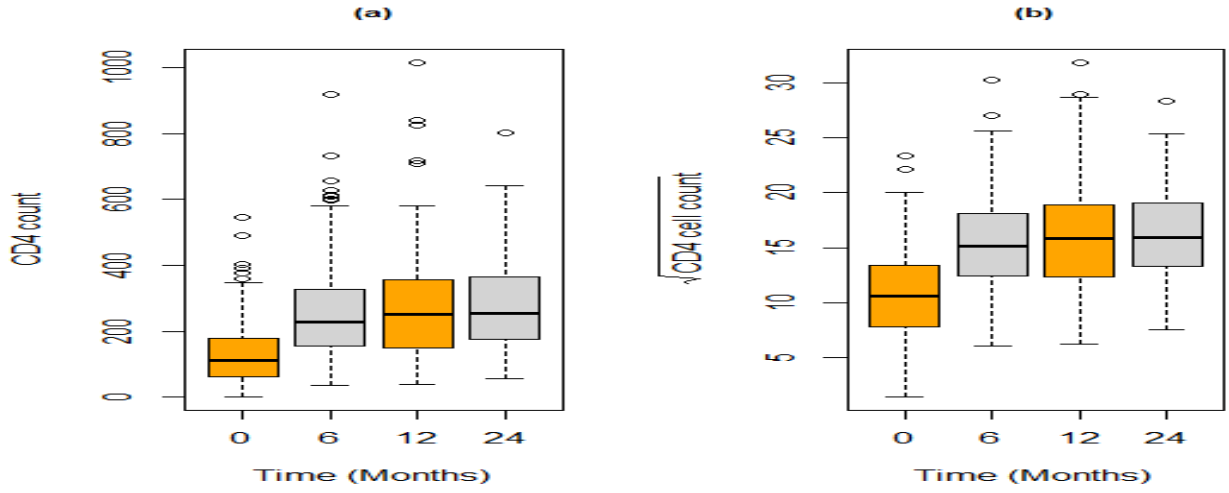

Figure B.4: Boxplots of the actual CD4 cells count (a) and the square root CD4 cells count (b) over time

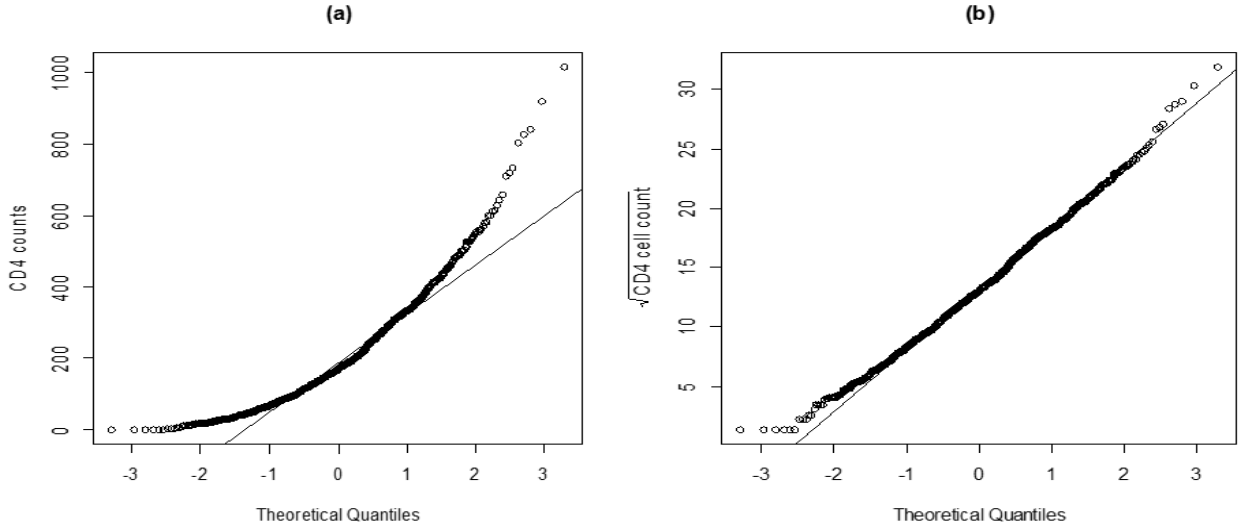

Figure B.5: Normal Q-Q Plot for actual CD4 cells count (a) and the Square root CD4 cells count (b) of HIV-infected patients under HAART

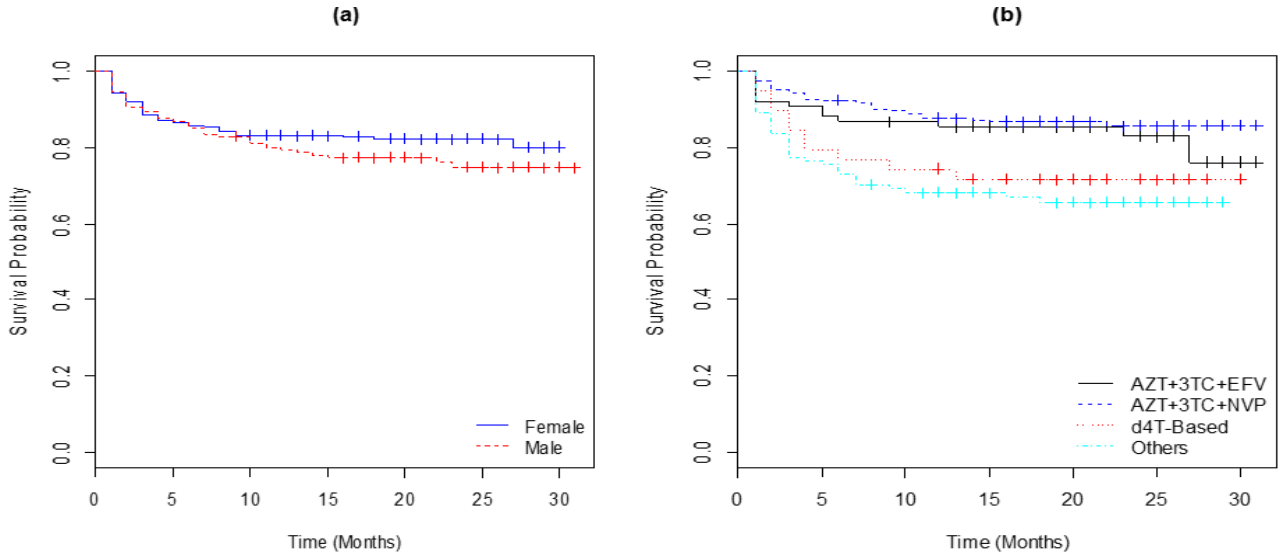

Figure B.6: Kaplan-Meier Survival plots of Sex (a) and Regimen type (b) of HIV-infected patients under HAART

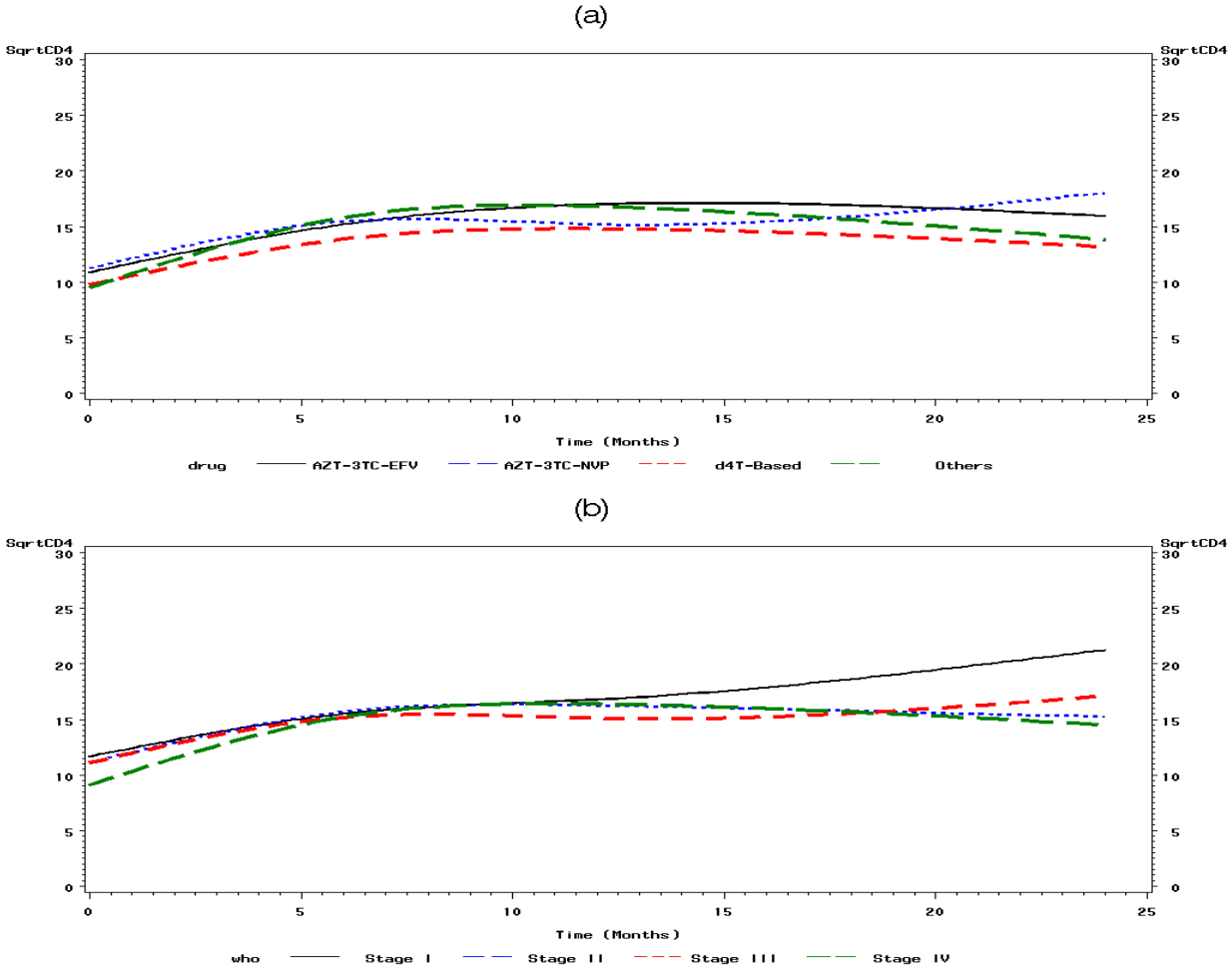

Figure B.7: The Mean profile plots of Regimen type (a) and WHO clinical stage (b) for HIV-infected patients under HAART