Modelling CD4 Count and Mortality in a Cohort of Patients Initiated on HAART

UNIVERSITY OF
KWAZULU - NATAL

INYUVESI
YAKWAZULU-NATALI

Nobuhle Nokubonga Mchunu

November, 2018

Modelling CD4 Count and Mortality in a Cohort of Patients Initiated on HAART

by

Nobuhle Nokubonga Mchunu

A thesis submitted to the
University of KwaZulu-Natal
in fulfilment of the requirements for the degree
of
MASTER OF SCIENCE
in
STATISTICS

Thesis Supervisor: Prof Henry G Mwambi

Thesis Co-supervisor: Dr Tarylee Reddy

Thesis SecCo-supervisor: Dr Nonhlanhla Yende-Zuma

UNIVERSITY OF KWAZULU-NATAL

SCHOOL OF MATHEMATICS, STATISTICS AND COMPUTER SCIENCE

PIETERMARITZBURG CAMPUS, SOUTH AFRICA

Declaration - Plagiarism

I, Nobuhle Nokubonga Mchunu, declare that

1. The research reported in this thesis, except where otherwise indicated, is my original research.

2. This thesis has not been submitted for any degree or examination at any other university.

3. This thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowlegded as being sourced from other persons.

4. This thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then

   (a) their words have been re-written but the general information attributed to them has been referenced, or

   (b) where their exact words have been used, then their writing has been placed in italics and referenced.

5. This thesis does not contain text, graphics or tables copied and pasted from the internet, unless specifically acknowledged, and the source being detailed in the thesis and in the reference sections.

13/03/2019

Nobuhle Nokubonga Mchunu (Student) Date

13/03/2019

Prof Henry G Mwambi (Supervisor) Date

13/03/2019

Dr Tarylee Reddy (Co-supervisor) Date

13/03/2019

Dr Nonhlanhla Yende-Zuma (SecCo-supervisor) Date

Disclaimer

This document describes work undertaken as a Masters programme of study at the University of KwaZulu-Natal (UKZN). All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the institution.

Dedication

To my late mother , Fungani Mngoma (may her beautiful soul rest in peace), and to my son Lusakhanya Zuma, this is for you both. God knows how much I love you

Abstract

Longitudinally measured data and time-to-event or survival data are often associated in some ways, and are traditionally analyzed separately (Asar et al., 2015). However, separate analyses are not applicable in this case because they may lead to inefficient or biased results. To remedy this, joint models optimally incorporate all available information (longitudinal and survival data) simultaneously (Wulfsohn & Tsiatis, 1997). Furthermore incorporating all sources of data improves the predictive capability of the joint model and lead to more informative inferences for the purpose of decision-making (Seyoum & Temesgen, 2017). The primary goal of this analysis was to determine the effect of repeatedly measured CD4 counts on mortality. The standard time-to-event models require that the time-dependent covariates of interest are external; where the value of the covariate at a future time point is not affected by the occurrence of the event. This requirement would not be fulfilled in this setting, since the repeatedly measured outcome is directly related to the mortality mechanism. Hence, a joint modeling approach was required.

We applied the methods developed in this thesis to the CAPRISA AIDS Treatment program (CAT). We also sought to determine if the patients' baseline BMI (Body mass index), baseline age, gender, baseline viral load, baseline CD8 count, baseline TB status and clinic site, influence the evolution of the CD4 count over time. Various linear mixed models were fitted to the CD4 count, adjusting for repeated measurements, as well as including intercept and slope as random effects. Different types of covariance structures were assessed and the spatial spherical correlation structure was found to be the best fit. The Cox PH model was employed to model mortality. Finally the joint model for longitudinal and time-to-event data was fitted.

Out of the 4014 patients, 1457 (36.30%) were male. There were more patients presenting without TB at ART initiation, 3042 (75.78%) compared to those with prevalent TB, 972 (24.22%). Results from the multivariable random effects model showed that the patients gender, age, baseline viral load and baseline CD8 cell count had statistically significant influences on the rate of change in CD4 cell count over time.

The un-adjusted and adjusted hazards regression both found CD4:CD8 ratio, viral load, gender and age of patients to be significant predictors of mortality. The result from the joint model in this study indicated that CD4 count change due to HAART and mortality had been influenced jointly by gender, age, baseline viral load, baseline CD8 count, time (in years) , CD4:CD8 ratio and by the interaction effects of time (in years) with TB status, baseline viral load and baseline CD8 cell count. CD4 count proved to be significantly associated with mortality, after adjusting for age, gender and other potential confounders

Model diagnostics were performed for validating model assumptions, and our joint model fitted quite well with fairly good diagnostic attributes. The methods that were developed in this thesis were applied to the CAPRISA AIDS Treatment program (CAT) between June 2004 to December 2013.

Acknowledgements

First and foremost I would like to thank God for giving me the strength and resilience to finish this thesis in record time through all the hardships I faced during this period.

I greatly appreciate my supervisors, Professor Henry G Mwambi, and Dr. Tarylee Reddy, for their constructive suggestions and comments on my research work. Their selfless support during my period of study at the University of Kwazulu-Natal is highly appreciated. I appreciate all their contributions of time, ideas, and funding to make my MSc experience productive and stimulating. In the same breath, I would like to thank DELTAS (Wellcome Trust's Developing Excellence in Leadership Training and Science) SSACAB (sub-Saharan Africa Consortium for Advanced Biostatistics) for partially funding my studies

My deepest gratitude to the Centre for the AIDS Programme of Research In South Africa (CAPRISA) for giving me the opportunity to use their world class data for the purposes of this thesis and for the fellowship that was awarded to me and for the invaluable experience I gained from working as a statistics fellow. In the same breath, I also want to express my sincere appreciation to Dr Nonhlanhla Yende-Zuma for being such a great mentor and co-supervisor to me during my fellowship at CAPRISA, I learnt so much from her. The passion and enthusiasm she has for the research was contagious and motivational for me, even during tough times in my MSc pursuit. I owe a great deal of thanks to the entire CAPRISA team especially the statistics department for being so supportive and very influential.

I dedicate the work of this thesis to the loving memory of my late mother, Fungani Mngoma who passed away so suddenly in March 2018. Thank you so much for being my pillar of strength and for always encouraging me to excel in everything that I do. The desire to make you proud spurred me on to complete this thesis.

Last but not least, I would like to thank all my friends and family for their love and support. To my best buddy Thembalethu Zuma thank you so much for always being there for me. Thank you all so much, this project would not have been what it is!

Contents

Page

List of Figures

List of Tables

Abbreviations

AFT	Accelerated failure time
AIC	Akaike information criterion
AIDS	Acquired Immune Deficiency Syndrome
AR	Autoregressive
ART	Antiretroviral therapy
ARV	Antiretroviral therapy
BIC	Bayesian information criterion
BMI	Body Mass Index
CAT	CAPRISA AIDS Treatment Project
CS	Compound symmetry
HAART	Highly Active Antiretroviral Therapy
HIV	Human Immunodeficiency Virus
KM	Kaplan-Meier
LD	Likelihood displacement
LME	Linear mixed effects
LRT	Likelihood Ratio Test
MAR	Missing at random
MCAR	Missing completely at random
ML	Maximum likelihood
PH	Proportional Hazards
REML	Restricted maximum likelihood
RLD	restricted likelihood distances
UN	Unstructured

Chapter 1

Introduction

1.1 Background

According to Karim & Karim (2010), Acquired Immune Deficiency Syndrome (AIDS) was first reported in 1983 in South Africa . Back then it was mainly associated with homosexuals, blood transfusion recipients and hemophiliacs (Karim & Karim, 2010). According to Karim & Karim (2010) approximately 5.3 million South Africans were estimated to be HIV positive by the end of 2007. in December 2006, approximately 1.3 million people in the sub-Saharan Africa region were receiving antiretroviral therapy (ART) (Bennett et al., 2008).

According to UNAIDS (2017) there were approximately 36.7 million people worldwide living with HIV/AIDS at the end of 2016. Of these, 2.1 million were children (<15 years old). As of June 2017, 20.9 million people living with HIV were accessing antiretroviral therapy (ART) globally. By mid-2016, 182 million people were on antiretroviral treatment UNAIDS (2016), up from 15.8 million in June 2015, 7.5 million in 2010, and less than one million in 2000. There have been some notable achievements in disease management, including substantial improvements in access to condoms, expansion of tuberculosis control efforts, and scale-up of free antiretroviral therapy (ART) (Karim et al., 2005). Even though antiretroviral treatment is widely accessible, treating and caring for millions of South Africans infected with HIV still poses a huge challenge as there are people who refuse to take the antiretroviral (ARV) treatment because of religious reasons, others fear side effects and have difficulty integrating pill-taking into their lives amongst other reasons (Karim & Karim, 2010).

According to Gandhi et al. (2006) the rising incidence of TB has been attributed to HIV co-infection especially in developing countries. This is due to the fact that HIV greatly increases the risk of active tuberculosis disease and majority of the patients presenting with prevalent tuberculosis in South Africa, are co-infected with HIV, thus posing a need to recruit those patients infected with TB (Karim & Karim, 2010).

1.1.1 CD4 count

CD4+ T cells (also called T-helper cells) are white blood cells that play an important role in ones immune system. They alert other immune cells to the presence of viruses and bacteria in a persons body. Certain receptors on the CD4+ T cells make them prime targets for HIV. If one contracts an HIV infection, the virus will attack ones CD4+ T cells. This will cause the number of CD4+ T cells in a persons body to drop, thereby weakening ones immune system. For a long time, the CD4 lymphocyte count was by far the most widely used biological marker of the disease to assess the stage of infection and HIV progression (Karim et al., 2005). However, there is less emphasis on CD4 count, the viral load is now widely used to monitor disease progression.

1.1.2 Viral load

During the acute phase of HIV infection, the viral load, which refers to the actual number of HI virus in the blood is very high, several times higher than set point levels during established infection. The viral load is used to determine the rate of destruction of the immune system. Thus the more HIV present in ones blood (and therefore the higher your viral load), then the faster the CD4 count will drop, and the greater the risk of acquiring other opportunistic infections because of HIV. Furthermore, high viral load is associated with a greater risk of HIV transmission, hence the importance of recognizing and detecting acute infection (Karim et al., 2005) and start treatment immediately.

1.1.3 CD8 count

CD8+ T cells are called the killer cells because they recognize and kill cells that are infected with a virus. CD8 lymphocyte counts and CD4 lymphocyte counts have an inverse relationship with each other, in that during an untreated HIV, the former increases as the latter declines (Margolick et al., 1995).

1.1.4 CD4:CD8 ratio

Few studies have addressed the significance of the CD4:CD8 ratio in HIV infection. Before the introduction of the Highly Active ART (HAART) in 1996, the CD4:CD8 ratio was identified as a predictor of disease progression (Taylor et al., 1989). More recently, a low CD4:CD8 ratio at initiation of ART has been associated with the prevalence and volume of coronary plaques (Lo et al., 2010). Sainz et al. (2013) reported that an association between the CD4:CD8 ratio and immune activation in HIV-infected adults results in long-term viral suppression.

1.2 Justification of the study

In this study we used data from the Centre for the AIDS Programme of Research in South Africa (CAPRISA). The CAPRISA AIDS Treatment (CAT) programme enrolled HIV positive patients and initiated them on ART between June 2004 and August 2013. Eligibility criteria was in accordance with the Department of Health guidelines throughout. Males and females at least 14 years of age from urban (eThekwini) and rural (Vulindlela) sites were enrolled. ”Routine demographic and clinical data were recorded at baseline and at follow-up visits. Laboratory safety assessments and CD4+ cell counts and viral loads were conducted at baseline and every 6 months or as clinically indicated. Patients were regarded as lost to follow-up if they missed 3 consecutive scheduled visits and if all attempts to track them telephonically and physically had failed”. Information on the deaths was based on hospital chart notes, death certificates or oral reports from participant’s relatives.

1.2.1 Joint Models for longitudinal and time-to-event Data

This thesis aim to investigates the joint modelling approach of Rizopoulos (2012), which enables one to combine longitudinal and time-to-event data. Rizopoulos (2012) applied the method to data from an AIDS clinical trial. Their longitudinal outcome was CD4 count. Their main research question was to test for a treatment effect on survival after adjusting for the CD4 count.

Joint models of longitudinal and time-to-event data have received much attention in the literature dating back to the past two decades. Many other investigators have described methods for estimating parameters of similar models. De Gruttola & Tu (1994) and Tsiatis et al. (1995) considered the progression of CD4 lymphocyte counts and survival time in patient with AIDS. De Gruttola & Tu (1994) assumed that the joint distribution of time-dependent log CD4 counts and some transformation of survival times are multivariate normally distributed. This formulation allowed them to fit the model using a modified expectation maximization (EM) algorithm which was proposed by Laird & Ware (1982). Wulfsohn & Tsiatis (1997) used Cox proportional hazards models to model the hazard of death as a function of the conditional expectation of true log CD4 counts given the history of observed counts, thus relaxing the normality assumption for the survival time.

Self & Pawitan (1992) proposed a two-step method for parameter estimation in modelling the relationship between CD4:CD8 ratios and time to AIDS diagnosis. Their method differs from Wulfsohn & Tsiatis (1997), in that they conditioned on survival information when computing expected values of the covariates. They also used partial likelihood to obtain estimates of the disease risk parameters, but they derived the corresponding variances to account for the uncertainty in the expected covariate values. To obtain these variances, they made the simplifying assumption that the variance of the covariate random effects is fixed and known (Faucett & Thomas, 1996).

A year later, Pawitan & Self (1993) used maximum likelihood methods to jointly model immunologic markers, time of infection, and time to AIDS. According to Faucett & Thomas (1996) they decided to switch things up, by modelling the marker as a function of disease time rather than modelling time of disease as a function of the marker. They also considered fully parametric Weibull regression models for the times of disease and infection.

Faucett et al. (2002) developed an approach, based on multiple imputation, using auxiliary variables to recover information from censored observations in survival analysis. They applied this approach to data from an AIDS clinical trial comparing ZDV and placebo, in which CD4 count is the time dependent auxiliary variable. To facilitate imputation, a joint model was developed for the data, which included a hierarchical change-point model for CD4 counts and a time dependent proportional hazards model for the time to AIDS.

Joint modelling techniques have seen great advances in the recent years. Several investigators, among others, Ding & Wang (2008) proposed a nonparametric multiplicative random effects model for the longitudinal process, which has many applications and leads to a flexible yet parsimonious nonparametric random effects model. A proportional hazards model is then used to link the biomarkers and event time.

Rizopoulos et al. (2009) proposed a new computational approach for fitting joint models of longitudinal response with a time-to-event outcome that is based on the Laplace method for integrals that makes the consideration of high dimensional random effects structures feasible. Contrary to the standard Laplace approximation, their method requires fewer repeated measurements per individual to produce reliable results (Faucett & Thomas, 1996).

Most recently Andrinopoulou et al. (2014) proposed a joint model consisting of two longitudinal outcomes, one continuous (aortic gradient) and the other ordinal (aortic regurgitation), and two time-to-event outcomes (death and re-operation). According to Faucett & Thomas (1996), they use B-splines to allow for more flexibility for the average evolution and the subject-specific profiles of the continuous repeated outcome. However, a drawback to this method is that when adopting a nonlinear structure for the model, there may be difficulties when interpreting the results.

Although substantial research has been done in the area of CD4 count modeling using linear mixed models, very few studies have used joint models to model CD4 count and time to death. A study done by Seyoum & Temesgen (2017) showed that the joint models were simpler as compared to the separate longitudinal and time to event models as their effective number of parameters was smaller. The current thesis is aimed at assessing whether the CD4 count predicts mortality and viral load suppression through construction of a joint model for longitudinal and time to event data in patients initiated on ART.

Ye et al. (2008) proposed a joint model for longitudinal measurements and time-to-event data in which the longitudinal measurements are modeled with a semiparametric mixed model to allow for the complex patterns in longitudinal biomarker data. They proposed a two-stage regression calibration approach that is simpler to implement than a joint modeling approach. In the first stage of their approach, the mixed model is fitted without regard to the time-to-event data. In the second stage, the posterior expectation of an individual's random effects from the mixed-model are included as covariates in a Cox model.

Seyoum & Temesgen (2017) used joint models to detect determinants of CD4 count change and adherence to highly active antiretroviral therapy. They found joint modelling analysis to be more parsimonious as compared to separate analysis, since it reduced type I error and subject-specific analysis improved its model fit. Chen et al. (2014) proposed a joint model for longitudinal and survival data with time-varying covariates subject to detection limits and intermittent missingness at random. Their proposed method was shown to improve the precision of estimates as compared to alternative methods.

A study conducted by Werner (2010) proved that ″modelling HIV markers jointly with informative drop-out is crucial to account for the missing data incurred from participants leaving the study to initiate ARV treatment. In ignoring this drop-out, CD4 count is estimated to be higher than what it actually is″.

1.3 Aims and objectives

- Determine the effect of ART on CD4 count trajectories and determine whether the CD4 evolution is influenced on measured covariates
- To describe mortality rates and determine the predictors of mortality among patients initiated on ART
- To construct a joint model for CD4 count and mortality in patients initiated on ART

1.4 Methodology

Make use of the linear mixed models methodology to model the CD4 count, adjusting for repeated measurements, as well as including intercept and slope as random effects.

″Use data of patients who commenced Highly Active Antiretroviral Therapy (HAART) from the Center for the AIDS Programme of Research in South Africa (CAPRISA) in the AIDS Treatment Project (CAT) between June 2004 and August 2013, including two years of follow-up for each patient″.

Analysis was done using linear mixed models for longitudinal data, Survival analysis models for time-to-event data and joint models for longitudinal data and time-to-event data.

Analysis was conducted using SAS, version 9.4 (SAS Institute INC., Cary) and R version 3.5.1.

P-values less than 0.05 were considered statistically significant.

1.5 Structure of the thesis

Chapter 1 (Introduction)

Chapter 2 (Exploratory data analysis)

Chapter 3 (linear mixed models)

Chapter 4 (Survival analysis)

Chapter 5 (Joint models for longitudinal and time-to-event data)

Chapter 6 (Application to the CAT data)

Chapter 7 (Discussion and concluding remarks)

Chapter 2

Data and exploratory analysis

2.1 Introduction

The first step in any analysis is to conduct an exploratory analysis, or univariate analysis of the data to obtain a clear sense of the distributional characteristics of the outcome variable as well as all possible predictor variables with the aim of determining the relevant modelling approaches suitable for it (Yende, 2010). Data, analysis was performed using SAS, version 9.4 (SAS Institute INC., Cary) and R version 3.5.1. In this chapter we will explore the distributional properties of the variables from the CAT project which include age, gender, BMI, site, TB status, CD4 Count, baseline CD8 Count, baseline viral load and the CD4:CD8 ratio.

2.2 Data description

"This thesis will use data collected at two sites, eThekwini (urban) and Vulindlela (rural) in KwaZulu-Natal province of South Africa. The data is collected as part of HIV and AIDS research by Centre for the AIDS Programme of Research in South Africa (CAPRISA). The data is collected on HIV+ positive patients. The eThekwini site enrolled the first patient on HAART in October 2004 while Vulindlela site enrolled the first patient in June 2004.

The eThekwini site stopped enrolling patients into the programme in August 2013 and the Vulindlela site stopped enrolling patients into the programme in January 2012. Patients at the eThekwini site are recruited from the Prince Cyril Zulu Clinic of Communicable Disease which is the chest clinic adjacent to the CAPRISA clinic and sometimes patients present themselves for HIV testing. Patients at the Vulindlela site are recruited from the Mafakatini clinic which is situated near that site or present themselves for medication". The data in the current study will be referred throughout the thesis as the CAPRISA AIDS Treatment Project (CAT).

"Some patients came for their six monthly visits a month prior to the scheduled visit or sometime a month after the scheduled visit which is still acceptable. An additional complexity with the data is that of missing observations due to drop out for known reasons such as death, loss to follow up and relocation to other areas. In this treatment project we have more females accessing ARVs than males". The description of the variables is presented in Table 2.1.

Table 2.1: Variable description

Characteristic	Description	Type
Gender	0 : Female	
	1 : Male	Binary
TB status	0 : No TB	
	1 : Prevalent TB	Binary
Site	0: Vulindlela site	
	1 : EThekwini site	Binary
Ratio	0 : CD4_CD8< 0.05	
	1 : CD4_CD8≥ 0.05	Binary
Sqrtcd4	CD4 count (square root transformed)	Continuous
BMI	Baseline body mass index	Continuous
Age	Baseline age in years	Continuous
sqrtcd8	Baseline CD8 count (square root transformed)	Continuous
Logviral	baseline viral load ($\log_{10}$ transformed)	Continuous
CD4_CD8_ratio		Continuous

Table 2.2: Baseline characteristics

Characteristic	N initiated on ART	EThekwini	Vulindlela	Missing	p-value
Age (years), mean ±SD	4010	34.28 ± 9.35	35.08 ± 8.13	4	0.0003
Gender, n (%):	4014			–	$< .0001$
Male		765 (42.15)	692 (31.47)		
Female		1050 (57.85)	1507 (68.53)		
TB status, n (%):	4014			–	$< .0001$
No TB		1092 (60.17)	1950 (88.68)		
Prevalent TB		723 (39.83)	249 (11.32)		
Median body mass index (kg/m^2),(IQR) ^{a}:	3777	22.70 (6.40)	22.70 (6.20)	237	0.4154
CD4 count, (cells/μL), median(IQR)^b:	3632	125.00 (134.00)	125.00 (132.50)	382	0.0079
Baseline CD8 count, (cells/μL), median(IQR)^c:	2078	751.00 (672.00)	799.00 (653.00)	1936	$< .0001$
Baseline viral load (log copies/ml), mean ±SD:	4010	4.96 ± 0.83	4.94 ± 0.96	488	0.0219
CD4:CD8 ratio, (cells/μL), median(IQR)^d :	2085			1929	$< .0001$
CD4:CD8 < 0.05		0.03 (0.02)	0.03 (0.02)		
CD4:CD8 ≥ 0.05		0.15 (0.13)	0.16 (0.15)		

Table 2.2 illustrates the results for the baseline characteristics broken down by site. There were 4014 patients enrolled whose ages range from 14-76 (with the mean age being 34.64 years). There were more females than males from both sites. Out of the 4014, 2557 (63.70%) were females and 1457 (36.30%) were males. All patients had a mean weight of 23.8 kg/m^2 at baseline with minimum and maximum weight of 10.5 and 264.5 kg/m^2 respectively. There were more patients presenting without TB at ART initiation, 3042 (75.78%) compared to those with prevalent TB, 972 (24.22%).

Table 2.3: Distribution of patients baseline characteristics according to gender and site

Characteristic	Ethekwini	Vulindlela
Age (years), mean ±SD		
Gender		
Female	33.69 (7.78)	33.68 (9.36)
Male	36.98 (8.23)	35.62 (9.09)
Body mass index (kg/m^2), mean ±SD		
Gender		
Female	25.21 (5.80)	25.09 (8.54)
Male	21.69 (3.99)	21.32 (6.09)
CD4 count (cells/μL), mean ±SD		
Gender		
Female	146.12 (120.76)	151.28 (121.04)
Male	127.45 (101.43)	121.72 (90.42)

Table 2.3 shows that women from both sites are younger than men . The mean BMI for males and females within each site was almost the same. Women compared to men have a higher mean BMI at baseline from both sites. Furthermore on average women had a higher baseline CD4 count compared to men from both sites.

Table 2.4: Patients' frequency for termination reason

Reason	Frequency (%)
Transferred	3038 (75.69)
Death	414 (10.31)
Defaulted	396 (9.87)
Patient decision	78 (1.94)
Relocated	51 (1.27)
Other	43 (1.07)
Poor adherence	5 (0.12)

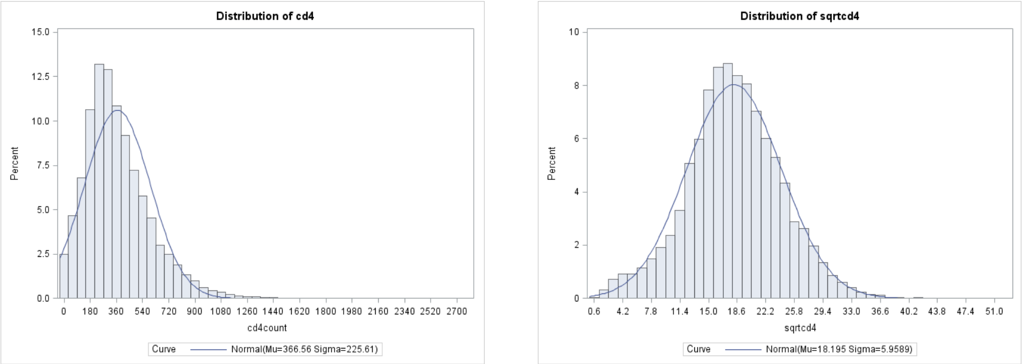
Of the 4014 patients, majority were transferred from the study accounting for 75.69%. Only 414 (10.31 %) people died.

2.3 Mortality

Majority of deaths occurred in patients without TB and patients from the Vulindlela site over 8195.87 person-years of follow-up. Patients presenting without TB had higher mortality rates compared to those with prevalent TB, 4.11 per 100 person-years (p-y), (95% CI: 3.68-4.58) vs. 0.94 per 100 p-y, (95% CI: 0.74-1.17); mortality rate ratio: 0.23, (95% CI: 0.18-0.29), $p < 0.0001$. Patients from the Vulindlela site had higher mortality rates comparedto those form the EThekwini site, 3.28 per 100 person-years (p-y), (95% CI: 2.90-3.70) vs. 1.77 per 100 p-y, (95% CI: 1.49-2.08); mortality rate ratio: 1.86, (95% CI: 1.52-2.27), $p < 0.0001$.

2.4 Distributional properties of CD4 count, CD8 count, viral load and CD4:CD8 ratio

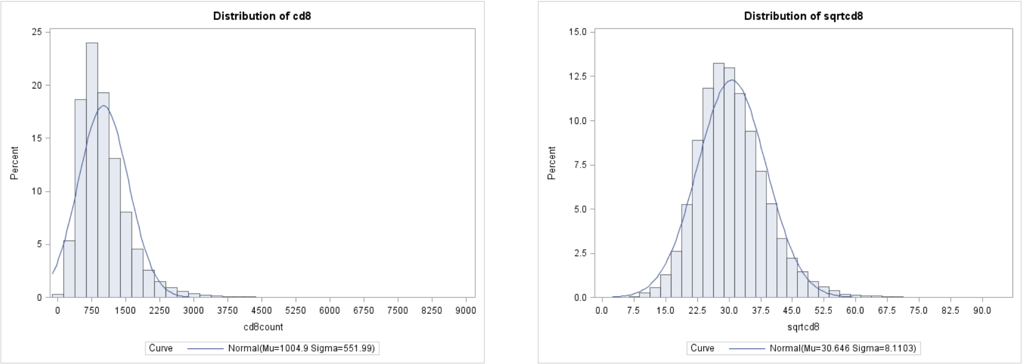
Figures 2.1, 2.2, 2.3 and 2.4 display the histograms that were plotted for our data to check the normality assumptions for CD4 counts, CD8 count, viral load and CD4:CD8 ratio respectively.



(a) CD4 count (cells/μL)

(b) Square root CD4 count (cells/μL)

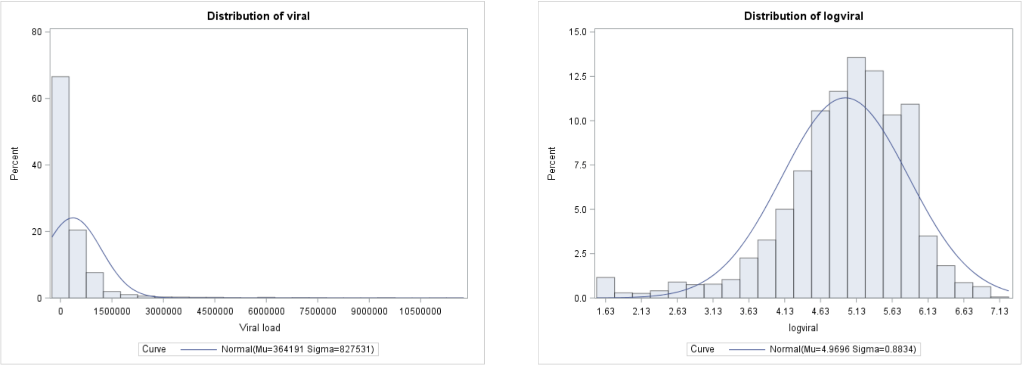
Figure 2.1 – Histogram for CD4 count (cells/μL) and square root transformed CD4 count (cells/μL)



(a) CD8 count (cells/μL)

(b) Square root CD8 count (cells/μL)

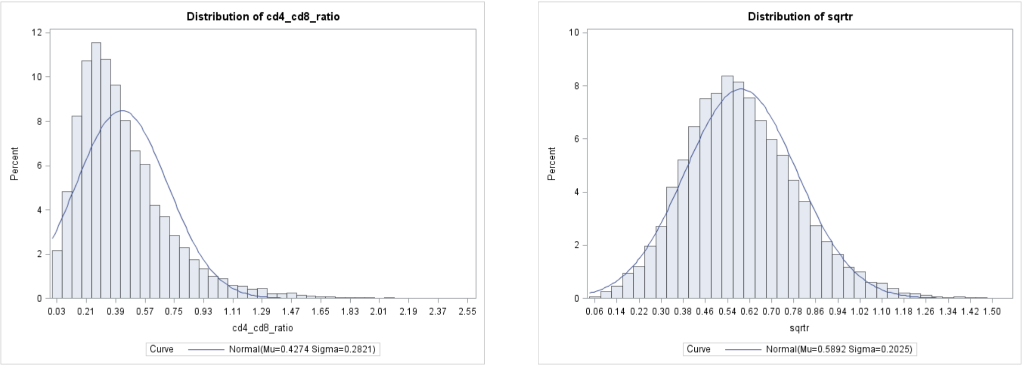
Figure 2.2 – Histogram for CD8 count (cells/μL) and square root transformed CD8 count (cells/μL)



(a) Viral load(copies/ml)

(b) $\log_{10}$ viral load(copies/ml)

Figure 2.3 – Histogram for viral load(copies/ml) and $\log_{10}$ transformed viral load

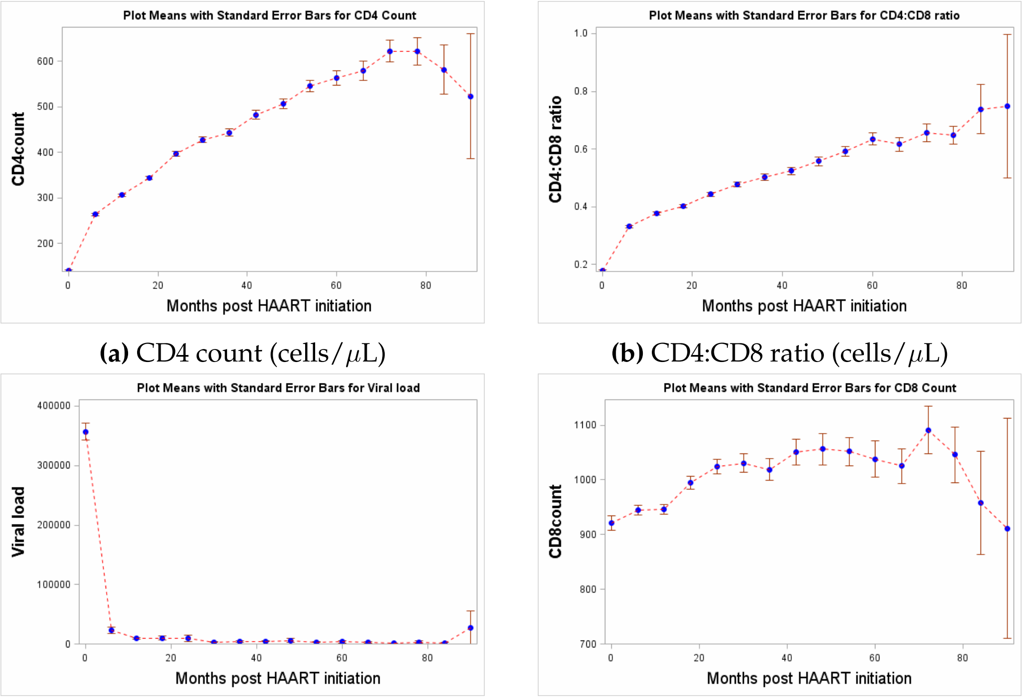


(a) CD4:CD8 ratio

(b) Square root CD4:CD8 ratio

Figure 2.4 – Histogram for CD4:CD8 ratio and square root transformed CD4:CD8 ratio

Figures 2.1a, 2.2a, 2.3a and 2.4a all show right-skewed histograms which violates the normality assumptions. There are several transformation methods available to normalize the data such as $\log_{10}$ transformation and the square root transformation. For this project we applied the square root transformation to CD4 count, CD8 count and CD4:CD8 ratio to normalize the data as can be seen in figures 2.1b, 2.2b and 2.4b because the square root transformation better approximates the normal distribution compared to the original CD4 count and CD4:CD8 ratio. The square root transformation is a commonly used transformation used when analyzing CD4 counts, as evidenced by previous research in similar cohorts (Yende (2010) ; Reddy et al. (2016); Wandeler et al. (2013); Reda et al. (2013) and De Beaudrap et al. (2009)). Throughout this project we will use square root transformed variables of CD4 Count, CD8 Count and CD4:CD8 ratio in the modelling processes.



(a) CD4 count (cells/μL)

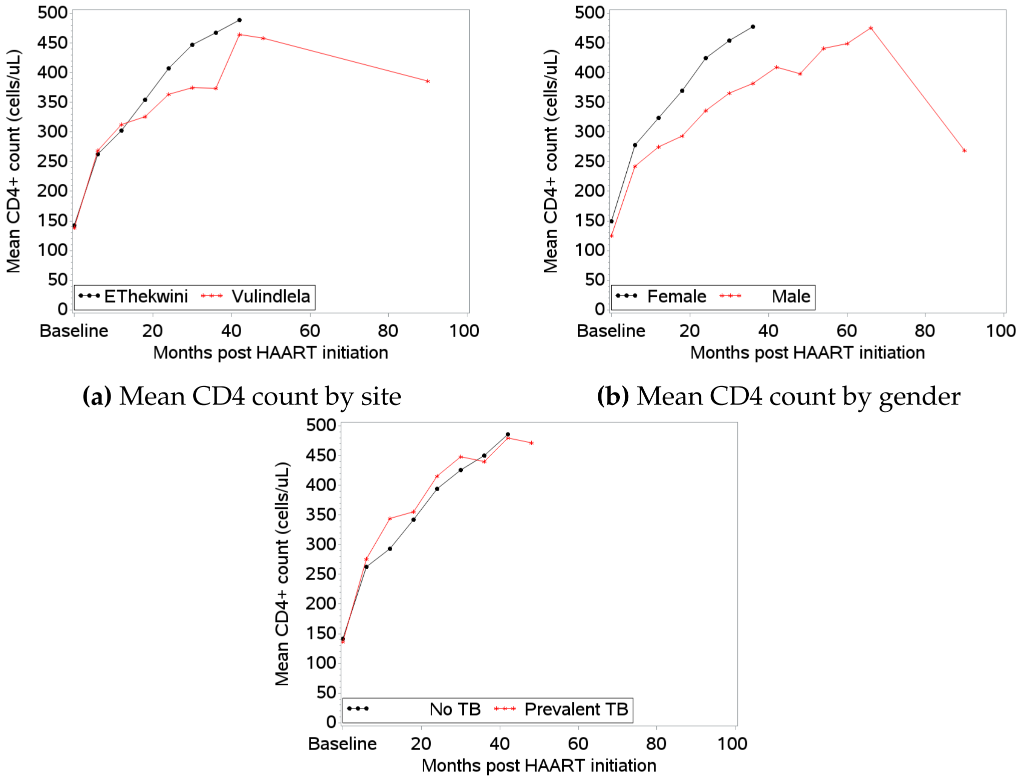
(b) CD4:CD8 ratio (cells/μL)

(c) viral load(copies/ml)

(d) CD8 count (cells/μL)

Figure 2.5 – Rate of change for CD4 count (cells/μL),CD8 count (cells/μL), CD4:CD8 ratio and viral load

Figure 2.5a and 2.5b suggests that CD4 count increases over time, after a patient has been initiated on HAART, this is exactly what we would expect and Figure 2.5c suggests that viral load decreases over time after a patient has been initiated on HAART. As can be seen on figure 2.5d, CD8 increase to a maximum then starts to decrease thereafter .



(a) Mean CD4 count by site

(b) Mean CD4 count by gender

(c) Mean CD4 count by TB status

Figure 2.6 – Rate of change for CD4 count (cells/μL) for site, gender and TB status

Figure 2.6a shows that after HAART initiation the EThekwini patients had a high rate of change for CD4 count and shows that female patients before and after HAART initiation had a higher rate of change for CD4 count compared to males regardless of site this is seen on figure 2.6b. Figure 2.6c shows that patients without TB start with higher mean CD4 count compared to those with prevalent TB but the rate of change in CD4 count is less compared to those with prevalent TB.

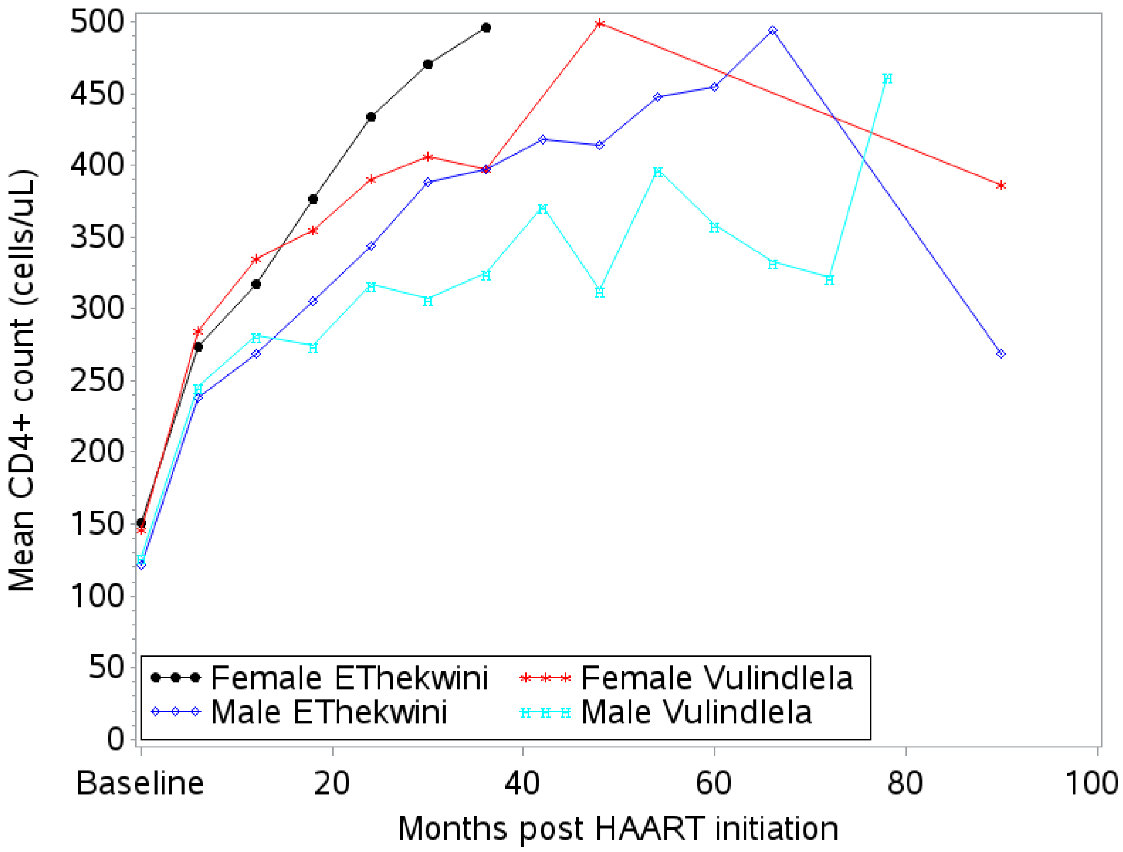
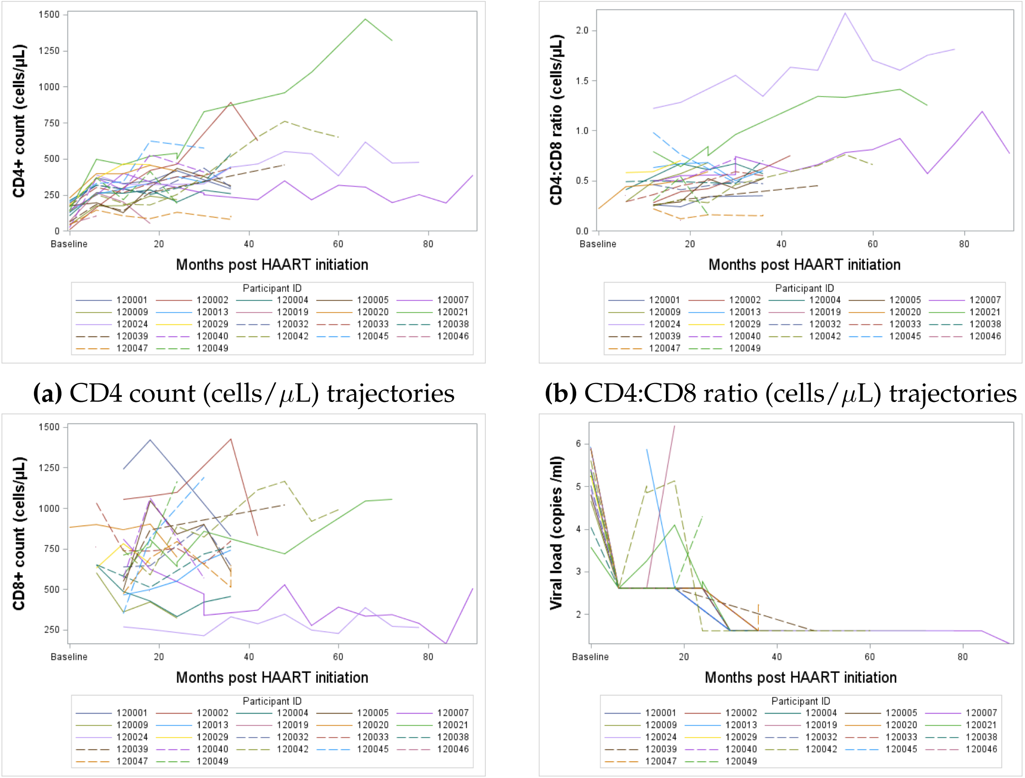


Figure 2.7 – Rate of change for CD4 count (cells/μL) for gender from both sites

The mean CD4 count at baseline for the eThekwini and Vulindlela site were 138.58 and 141.69 cells/μL respectively. Further women at both Vulindlela and eThekwini sites started with mean CD4 count of 139.32 and 139.47 cells/μL respectively, and men from Vulindlela and the eThekwini sites started with 147 and 137.36 respectively. Figure 2.7 illustrates these results.

2.5 Individual profiles (spaghetti plots)

The spaghetti plots are generally used to assess if there is any variation between and within subjects. We randomly selected 50 patients to construct such plots since graphs with all individual curves can be hard to distinguish for large sample size. Its important to note that randomly drawn subjects need not be representative and extreme curves are unlikely to be shown. Figures 2.8a to 2.8d indicated within and between patient variability in the rate of change of CD4 count, CD4:CD8 ratio, CD8 count, and viral load variability over time.



(a) CD4 count (cells/μL) trajectories

(b) CD4:CD8 ratio (cells/μL) trajectories

(c) CD8 count (cells/μL) trajectories

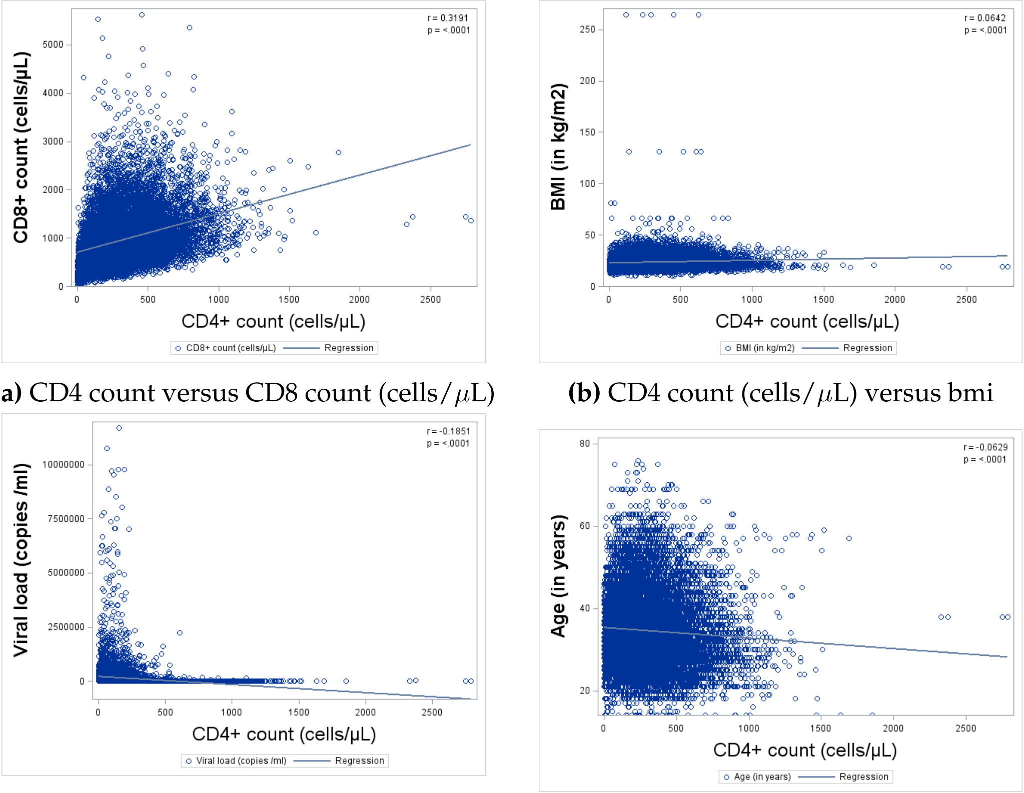
(d) viral load(copies/ml) trajectories

Figure 2.8 – Individual trajectories for a random sample of 50 patients, for CD4 count, CD4:CD8 ratio, CD8 count and viral load.

Figures 2.8a and 2.8b both portray the same qualitative features which depicts an increasing trend of CD4 count and the CD4:CD8 ratio over time after HAART initiation for 50 randomly selected patients. What can be observed is that generally there is evidence of between subjects variability as well as within subject variability. The subjects have large CD4 and CD4:CD8 evolutions over time, this suggests that perhaps linear mixed models with random intercepts and slopes could be plausible starting points. The thinning of the data toward later visit months suggests that trends at later times should be treated with caution Verbeke (1997). A decreasing trend of CD8 count and viral load over time after HAART initiation can be seen for 50 randomly selected patients in figures 2.8c and 2.8d respectively.

2.6 Scatter plots with sample correlation for CD4 count against covariates

A scatter plot shows graphically the relationship between two variables. We then used it to check for correlation between CD4 count and other continuous variables in our data, namely viral load, CD8 count, age and BMI. This is very helpful for model building.



(a) CD4 count versus CD8 count (cells/μL)

(b) CD4 count (cells/μL) versus bmi

(c) CD4 count (cells/μL) versus viral load(copies/ml)

(d) CD4 count (cells/μL) versus age (in years)

Figure 2.9 – Scatter plot of all CD4 count measurements versus CD8 count, BMI, viral load and age.

Figure 2.9a suggests a positive correlation between CD4 count and CD8 count. Hence, as CD4 count increases so does CD8 count. The Pearson's correlation coefficient was 0.3191 and this was statistically significant ($p< 0.0001$). As CD4 count increases so does BMI, this is shown by figure 2.9b. Figure 2.9c suggests a negative correlation between viral load and CD4 count. Hence, as viral load increases, CD4 count decreases, which is exactly what is expected given the relationship between these two variables in the absence of treatment. The Pearson's correlation coefficient was -0.1851 and this was statistically significant ($p< 0.0001$). Furthermore as age increases, CD4 count decreases 2.9d .

2.6.1 Summary

In this chapter patients baseline characteristics and distributional properties of the biomarkers were explored. The CD4 count, CD8 count and CD4:CD8 ratio were square root transformed and the viral load was transformed using a logarithm approximation. The spaghetti plots indicated some within and between patient variation which suggested that a model with both random intercepts and slopes could be plausible. The mean plots suggested an increase in the evolution of CD4 count over time after patients had been initiated on HAART.

Chapter 3

Linear mixed models

3.1 Introduction

In this chapter the theory of linear mixed models including two types of estimation methods and model diagnostics will be explored.

Linear mixed models are an extension of linear regression models which incorporate random effects in the structure for the mean, so that the data is allowed to display correlation and non-constant variability. Mixed models are commonly used for the analyses of longitudinal data where experimental units are followed over a period of time and they are called subjects. These subjects are regarded as a random sample from a larger population of subjects and hence any effects that are not constant for all subjects are regarded as random.

The linear mixed model is commonly used for analyzing continuous repeated measures from individuals ranging from social, economical, agriculture and biomedical applications (O'Brien & Fitzmaurice, 2004). The advantage of using such models is that they allow for unbalanced designs where all subjects do not require an equal number of observations and/or the same data collection occasions or visits (Zhang & Chen, 2013).

According to O'Brien & Fitzmaurice (2004), longitudinal data analysis is widely used for three reasons namely;

- To increase the sensitivity by making within-subject comparisons
- To study evolutions of outcomes of interest through time
- To use subject efficiency once they are enrolled in a study.

3.2 Model building

Consider a data set with N subjects. Let n_i denote the number of observations for the i^{th} subject. Let Y_i be the $n_i \times 1$ vector of observations for the i^{th} subject $(1 \leq i \leq N)$. Then the general linear mixed model is given by

$$Y_i = X_i\beta + Z_ib_i + e_i \tag{3.1}$$

where

- β is a $(p \times 1)$ vector which contains the parameters for the p fixed effects in the model including the constant term
- b_i is a $(q \times 1)$ vector with the random effects for the i^{th} subject in the data set.
- $X_i(n_i \times p)$ and $Z_i(n_i \times q)$ are the design matrices for the p fixed and q random effects respectively.
- e_i is a $n_i \times 1$ vector which contains the residual components for subject i

The random effects, b_i and e_i are assumed to be independent and are normally distributed with mean vector 0 and covariance matrices $D(q \times q)$ and $\sum_i (n_i \times n_i)$ respectively. Different structures for these covariance matrices are possible and will be briefly discussed in section 3.5. Thus

$$b_i \sim N(0, D)$$

and

$$e_i \sim N\Big(0, \Sigma_i\Big).$$

The distribution of b_i and e_i can jointly be written as;

$$\begin{bmatrix} b_i \\ e_i \end{bmatrix} \sim N\left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} D & 0 \\ 0 & \Sigma_i \end{bmatrix} \right\}. \tag{3.2}$$

An important distinction in the linear mixed model is between the conditional and marginal models for Y_i. The subject-specific mean of Y_i conditionally on b_i from equation 3.2 is given by

$$E(Y_i|b_i) = X_i\beta + Z_ib_i \tag{3.3}$$

and the conditional variance of Y_i given b_i is

$$Var(Y_i|b_i) = \Sigma_i \tag{3.4}$$

thus

$$(Y_i|b_i) \sim N\Big(X_i\beta + Z_ib_i, \Sigma_i\Big). \tag{3.5}$$

When we assume conditional independence the variance of Y_i given b_i is

$$Var(Y_i|b_i) = \sigma^2 I. \tag{3.6}$$

The marginal mean Y_i of when averaged over the distribution of random effects b_i is given by

$$\begin{aligned} E(Y_i) &= E\big[E(Y_i|b_i)\big] \\ &= E(X_i\beta + Z_ib_i) \\ &= X_i\beta + Z_iE(b_i) \\ &= X_i\beta \end{aligned} \tag{3.7}$$

and the marginal covariance matrix is

$$\begin{aligned} Var(Y_i) &= V_i \\ &= E\big[Var(Y_i|b_i)\big] + Var\big[E(Y_i|b_i)\big] \\ &= E[\Sigma_i] + Var(X_i\beta + Z_ib_i) \\ &= \Sigma_i + Z_iDZ_i^{'} \\ &= Z_iDZ_i^{'} + \Sigma_i \end{aligned} \tag{3.8}$$

thus the implied marginal model is given by

$$Y_i \sim N\Big(X_i\beta, Z_iDZ_i^{'} + \Sigma_i\Big). \tag{3.9}$$

In this interpretation, it becomes clear that the fixed effects enter only through the mean $E(Y_{ij})$, whereas the inclusion of subject-specific effects specifies the structure of the covariance between observations on the same unit (Antonio & Beirlant, 2007). It should be noted that intrinsically, the marginal model allows negative variance components provided V_i from 3.8 is positive semi-definite while in the conditional model negative components do not make sense.

3.3 Estimation of parameters in linear mixed models

According to Searle et al. (2008), the most often used methods of estimation in gaussian mixed models are maximum likelihood (ML) and restricted maximum likelihood (REML). The REML procedure is most popular when it comes to the estimation of variance components in mixed models assuming Gaussian random terms. REML maximizes the joint likelihood of all error contrasts rather than of all contrasts as in ordinary maximum likelihood (Gilmour et al., 1995).

3.3.1 Maximum likelihood estimation (MLE)

The method of maximum likelihood estimation was first introduced by (RA Fisher, 1922). He first presented the numerical procedure in 1912. Since then, this method has become one of the most important tools for estimation and inference available to statisticians. According to White (1982) a fundamental assumption underlying classical results on the properties of the maximum likelihood estimator is that the stochastic law which determines the behavior of the phenomena investigated (the ″true″ structure) is known to lie within a specified parametric family of probability distributions (the model). In other words, the probability model is assumed to be ″correctly specified.″ The drawback of the maximum likelihood estimator is that it does not take into account the degrees of freedom used in estimating fixed effects. Thus standard errors are underestimated and results in narrower confidence intervals hence a bigger chance to reject the null hypothesis. The MLE for σ^2 is obtained by maximizing the joint log-likelihood distribution given by

$$L_{ML}(\theta, Y) = -\frac{1}{2}\Big\{log|V| + (Y - X_i\hat{\beta})^{-1}V^{-1}(Y - X_i\hat{\beta})\Big\}. \tag{3.10}$$

This is the simple case of the linear model for independent observations and homogeneous variance, thus the MLE for σ^2 is

$$\hat{\sigma}^2_{ML} = \sum_{i=1}^{N} \frac{(Y_i - X_i\hat{\beta})'(Y_i - X_i\hat{\beta})}{N} \tag{3.11}$$

$\hat{\sigma}^2_{ML}$ is a biased estimator for σ^2 since

$$E(\hat{\sigma}^2_{ML} - \sigma^2) = -\frac{\sigma^2}{N} \neq 0 \tag{3.12}$$

where $\sigma^2, N > 0$, this implies that $\hat{\sigma}^2_{ML}$ underestimate σ^2.

3.3.2 Estimation of fixed effects parameters under ML

"Let β denote the vector of fixed effects and let α be a vector of all variance components in D and Σ_i, it then follows that the variance covariance matrix V_i of Y_i dependents on α. Thus we can let $\theta = (\beta', \alpha')'$ denote the vector of all parameters in the marginal model". Thus assuming the above assumptions holds, the marginal likelihood function is given by

$$L_{ML}(\theta) = \prod_{i=1}^{N} \Big\{ (2\pi)^{-\frac{n_i}{2}} |V_i|^{-\frac{1}{2}} \Big\} \times \Big\{ -\frac{1}{2}(Y_i - X_i\beta)' V_i^{-1} (Y_i - X_i\beta) \Big\} \tag{3.13}$$

where V_i is the matrix of variance components. According to Verbeke & Molenberghs (2000) the estimates for α_{ML} and β_{ML} can be obtained from maximizing $L_{ML}(\theta)$ with respect to θ that is with respect to α and β simultaneously. The log likelihood function for subject i is

$$l_i = logL_i = -\frac{n_i}{2} log(2\pi) - \frac{1}{2} log|V_i| \times \Big\{ -\frac{1}{2}(Y_i - X_i\beta)^{-1} V_i^{-1} (Y_i - X_i\beta \Big\}. \tag{3.14}$$

According to Harville (1977), if α is known then the maximum likelihood estimate of β is given by

$$\hat{\beta}(\alpha) = \Big\{ \sum_{i=1}^{N} X_i' W_i X_i \Big\}^{-1} \times \Big\{ \sum_{i=1}^{N} X_i W_i y_i \Big\} \tag{3.15}$$

where $W_i = V_i^{-1}$. According to O'Brien & Fitzmaurice (2004) the estimator of β that minimizes this expression is known as the generalized least squares (GLS) estimator of β, denoted by $\hat{\beta}$. If $E(Y_i)$ is correctly modeled, it can be shown that

$$E(\hat{\beta}) = \beta \tag{3.16}$$

and

$$\begin{aligned} Var(\hat{\beta}) &= \Big\{ \sum_{i=1}^{N} X_i' W_i X_i \Big\}^{-1} \Big\{ \sum_{i=1}^{N} X_i' W_i' Var(Y_i) W_i X_i \Big\} \Big\{ \sum_{i=1}^{N} X_i W_i X_i' \Big\}^{-1} \\ &= \Big\{ \sum_{i=1}^{N} X_i' W_i X_i \Big\}^{-1} \end{aligned} \tag{3.17}$$

provided $Var(Y_i)$ is truly given by V_i. In most cases, α is not known, and needs to be replaced by an estimate $\widehat{\alpha}$ and we can set $\widehat{V_i} = \widehat{W_i}^{-1}$ and estimate β by using the expression with W_i replaced with $\widehat{W_i}$ by (Verbeke & Molenberghs, 2000) .

3.3.3 Estimation of variance components under ML

The maximum likelihood procedure of Hartley & Rao (1967) is modified by adapting a transformation from Patterson & Thompson (1971) which partitions the likelihood render normality into two parts, one being free of the fixed effects. Recent developments promise to increase greatly the popularity of maximum likelihood as a technique for estimating variance components (Harville, 1977). Miller (1973) developed a satisfactory asymptotic theory for maximum likelihood estimators of variance components.

Consider a linear mixed model for one trait, represented by 3.1, the least squared equations given by

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix} \tag{3.18}$$

where $\hat{a}$ denote the vector of addictive effects, absorbing the fixed effects reduces the equations to

$$Z^{'}KZ\hat{a} = Z^{'}Ky \tag{3.19}$$

where

$$K = I - X^{'}(X^{'}X)^{-1}X^{'}. \tag{3.20}$$

According to Meyer (1991), a generalized inverse can be used when the inverse of $X^{'}X$ does not exist. To get the estimates of the variance components we consider method 3 of fitting constants by (Henderson, 1953). Thus

$$\hat{\sigma}_e^2 = \left\{ \frac{(y^{'}y) - \hat{a}^{'}Z^{'}y - \hat{b}^{'}X^{'}y}{r(X) - r(Z) + 1} \right\} \tag{3.21}$$

$$\hat{\sigma}_a^2 = \left\{ \frac{aZKy - \big(r(Z) - i\big)\hat{\sigma}_e^2}{tr(Z^{'}KZ)} \right\} \tag{3.22}$$

Where $r(X)$ and $r(Z)$ denote the column rank of X and Z respectively. N represents the number of observations, and tr is the trace operator. In this method any covariances between levels of a are ignored. When a and e are taken as having zero covariance that are from

$$V = ZG^{'}Z + R \tag{3.23}$$

and with V non-singular

$$VV^{-1} = I \tag{3.24}$$

and supposing that V is a square matrix having elements that are not functionally related.

$$\frac{\partial V^{-1}}{\partial \theta} = -V^{-1}\left(\frac{\partial V}{\partial \theta}\right)V^{-1} \tag{3.25}$$

and

$$\frac{\partial}{\partial \theta} log|V| = tr\left(V^{-1}\frac{\partial V}{\partial \theta}\right) \tag{3.26}$$

where elements of V are considered as function of θ. Using this result, we arrange the variance covariance components that occur in V as a vector $\theta_{h=1}^{v}$, where v is the total number of different components. Then to find the variance components estimates, we maximize the equation given by

$$L_{\theta_h} = log\left\{tr\left(V^{-1}\left(\frac{\partial V}{\partial \theta}\right)\right)\right\}^{-\frac{1}{2}} \times exp\left\{-\frac{1}{2}(Y - X\beta)^{'}V^{-1}(Y - X\beta)\right\} \tag{3.27}$$

$$l_{\theta_h} = log(L_{\theta_h}) = -\frac{1}{2}\left\{tr\left(V^{-1}\left(\frac{\partial V}{\partial \theta}\right)\right)\right\} - \frac{1}{2}\left\{(Y - X\beta)^{'}V^{-1}(Y - X\beta)\right\} \tag{3.28}$$

equating 3.28 to zero we get

$$tr\left[\widehat{V}^{-1}\left(\frac{\partial V}{\partial \theta_{h|\theta=\hat{\theta}}}\right)\right] = (Y - X\hat{\beta})^{'}\widehat{V}^{-1}\left(\frac{\partial V}{\partial \theta_{h|\theta=\hat{\theta}}}\right)(Y - X\hat{\beta}) \tag{3.29}$$

where

$$X\hat{\beta} = X(X^{'}\widehat{V}^{-1}X)^{-}X^{'}V^{-1} \tag{3.30}$$

we define

$$P = V^{-1} - V^{-1}X(X^{'}\widehat{V}^{-1}X)^{-}X^{'}V^{-1} \tag{3.31}$$

that is

$$V^{-1}(Y - X\hat{\beta}) = \hat{P}y \tag{3.32}$$

thus the maximum likelihood (ML) estimation equation is given by

$$tr\left[\widehat{V}^{-1}\left(\frac{\partial V}{\partial \theta_{h|\theta=\hat{\theta}}}\right)\right] = y^{'}\hat{P}\left(\frac{\partial V}{\partial \theta_{h|\theta=\hat{\theta}}}\right)\hat{P}y \tag{3.33}$$

To find the estimates of variance components, we consider the derivatives of ML equation in terms of D and Σ. We distinguish θ_d and θ_Σ as elements of θ that occurs in $Var(u) = D$ and $Var(e) = \Sigma$, respectively. Then

$$\frac{\partial V}{\partial \theta_d} = Z\left(\frac{\partial V}{\partial \theta_d}\right)Z^{'} \tag{3.34}$$

and

$$\frac{\partial V}{\partial \theta_\Sigma} = \frac{\partial D}{\partial \theta_\Sigma} \tag{3.35}$$

hence the ML equation becomes

$$tr\left[\widehat{V}^{-1}\left(\frac{\partial V}{\partial \theta_{h|\theta=\hat{\theta}}}\right)\right] = y^{'}\hat{P}\left(\frac{\partial D}{\partial \theta_{h|\theta=\hat{\theta}}}\right)\hat{P}y \tag{3.36}$$

for each parameter θ_d of D, and

$$tr\left[\widehat{V}^{-1}\left(\frac{\partial V}{\partial \theta_{h|\theta=\hat{\theta}}}\right)\right] = y^{'}\hat{P}\left(\frac{\partial \Sigma}{\partial \theta_{h|\theta=\hat{\theta}}}\right)\hat{P}y \tag{3.37}$$

for each parameter θ_Σ of Σ.

3.3.4 Restricted maximum likelihood estimation (REML)

Thompson Jr (1962) called this method restricted maximum likelihood or residual (marginal) maximum likelihood. Patterson & Thompson (1971) proposed an approach which takes into account the loss in degrees of freedom resulting from estimating fixed effects. Retaining the property of invariance under translation that ML estimators have, the REML estimators have the additional property of reducing to the analysis variance (ANOVA) estimators for many, if not all, cases of balanced data (equal subclass numbers). A computing algorithm is developed, adapting a transformation from (Hemmerle & Hartley, 1973), which reduces computing requirements to dealing with matrices having order equal to the dimension of the parameter space rather than that of the sample space. These same matrices also occur in the asymptotic sampling variances of the estimators.

3.3.5 REML estimation for the linear mixed model

Consider models where

$$Y \sim N(X\beta, V_i) \tag{3.38}$$

where

$$V_i = Z_i D Z_i^{'} + \Sigma_i \tag{3.39}$$

Combining the subject-specific sub-models we get

$$Y \sim N\big\{(X\beta), V(\alpha)\big\} \tag{3.40}$$

where

$$\begin{pmatrix} V_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & V_n \end{pmatrix} \tag{3.41}$$

According to Verbeke & Molenberghs (2000), maximization of the likelihood function of a set of error contrasts, gives the REML estimator for the variance components α and β given by

$$L_{REML}(\theta) = \left| \sum_{i=1}^{N} X_i^{'} W_i X_i (\hat{\alpha})^{\frac{-1}{2}} \right| \times L_{ML}(\theta) \tag{3.42}$$

with respect to $\theta = (\beta^{'}, \alpha^{'})^{'}$. The resulting estimates for β and α will be denoted by β_{REML} and α_{REML} respectively. $L_{REML}(\theta)$ can be seen as the penalized likelihood (Verbeke & Molenberghs, 2000).

3.3.6 Inference for the fixed effects

As stated previously if α is known the MLE of β is given by

$$\hat{\beta}(\alpha) = \Big\{ \sum_{i=1}^{N} X_i^{'} W_i X_i \Big\}^{-1} \Big\{ \sum_{i=1}^{N} X_i^{'} W_i Y_i \Big\}. \tag{3.43}$$

Under the marginal model above and conditional on α and $\hat{\beta}$ follows a multivariate normal distribution with mean vector given by

$$E[\hat{\beta}(\alpha)] = \Big\{\sum_{i=1}^{N} X_i' W_i X_i\Big\}^{-1}\Big\{\sum_{i=1}^{N} X_i' W_i E(Y_i)\Big\} \tag{3.44}$$

and we know that $E(Y_i) = X_i\beta$. Thus the above expression becomes

$$\begin{aligned} E[\hat{\beta}(\alpha)] &= \Big\{\sum_{i=1}^{N} X_i' W_i X_i\Big\}^{-1}\Big\{\sum_{i=1}^{N} X_i' W_i X_i \beta\Big\} \\ &= \beta \end{aligned} \tag{3.45}$$

and the covariance of β is then given by

$$\begin{aligned} V[\hat{\beta}(\alpha)] &= \Big\{\sum_{i=1}^{N} X_i' W_i X_i\Big\}^{-1}\Big\{\sum_{i=1}^{N} X_i' W_i Var(Y_i) W_i X_i\Big\}\Big\{\sum_{i=1}^{N} X_i' W_i X_i\Big\}^{-1} \\ &= \Big\{\sum_{i=1}^{N} X_i' W_i X_i\Big\}^{-1} \end{aligned} \tag{3.46}$$

provided $Var(Y_i)$ is truly given by V_i. Furthermore $W_i = V_i^{-1}$, and in most cases, α is not known, and needs to be replaced by an estimate $\hat{\alpha}$ and we can set $\hat{V_i} = \hat{W_i}^{-1}$ and estimate β by using the expression with W_i replaced with $\hat{W_i}$ by (Verbeke & Molenberghs, 2000).

3.3.7 Approximate Wald test

For any known matrix L, consider testing the hypothesis.

$$\begin{aligned} &H_0 : L\beta = 0 \\ &vs \\ &H_A : L\beta \neq 0 \end{aligned} \tag{3.47}$$

then the Wald test statistic is given by

$$W = \hat{\beta}' L' \Big[L \sum_{i=1}^{N} X_i' W_i X_i L'\Big]^{-1} L\hat{\beta} \tag{3.48}$$

"The asymptotic sum distribution W is chi-square distributed with rank (L) degrees of freedom". The deficiency of the Wald test is that the variability introduced by replacing α by some estimate (ML or REML) is not taken into account in the subsequent test, thus providing valid inferences only in sufficiently large samples. According to Gianola & Foulley (1990) there are at least 2 potential shortcomings of REML, first, REML estimates are the elements of the modal vector of the joint posterior distribution of the variance components. From a decision theoretic point of view, the optimum Bayes decision rule under quadratic loss in the posterior mean rather than the posterior mode. The mode of the marginal distribution of each variance component should provide a better approximation to the mean than a component of the joint mode. Second, if inferences about a single variance component are desired, the marginal distribution of this component should be used instead of the joint distribution of all components.

3.3.8 Inference for the variance components

The mean structure is usually of primary interest in the inference, however in a variety of applied statistical problems, there is a need for inference on variance components. This includes a variety of applied fields, for example, random-effects ANOVA models (Nelder, 1954), linear mixed models Verbeke & Molenberghs (2000), generalized linear and nonlinear (mixed) models (Jacqmin-Gadda & Commenges, 1995), over dispersion Cox (1983); Smith & Heitjan (1993); Ohara Hines (1997); and Lu (1997), clustering (Britton, 1997), and homogeneity in stratified analyses (Liang, 1987). A test for variance component helps in proving or establishing whether we do need the inclusion of random effects or not.

3.3.9 Approximate Wald test

"Asymptotically, ML and REML estimates of α are normally distributed with correct mean and inverse Fisher information matrix $I(\alpha)^{-1}$ as covariance". Hence approximate standard errors and Wald tests can easily be obtained for the variance components

3.3.10 The Likelihood ratio test (LRT)

The likelihood ratio test is ideal for the comparison of nested models with different covariance structure, but equal mean structures. When comparing two models it seems reasonable to consider the ratio of the likelihoods under the two models. This is the likelihood ratio. The better model has the greater likelihood. Let the hypothesis of interest be

$$H_0 : \alpha \in \Theta_{\alpha,0} \tag{3.49}$$

for some subspace $\Theta_{\alpha,0}$ of the parameter space Θ_α of the variance components α. Thus the test statistic is given by

$$LR = 2 \times [\ln(\hat{l}_{\text{alt}}) - \ln(\hat{l}_{\text{null}})] \tag{3.50}$$

Where $\hat{l}_{\text{alt}}$ and $\hat{l}_{\text{null}}$ denotes the likelihood for alternative model and likelihood for the null model respectively. The probability distribution of the test statistic is approximately a chi-squared (χ^2) distribution with degrees of freedom equal to $df_{\text{alt}} - df_{\text{null}}$, the number of free parameters for alternative and null models. There are many iterative algorithms that can be considered for computing the ML or REML estimates. The computations on each iteration of these algorithms are those associated with computing estimates of fixed and random effects for given values of the variance components.

3.4 The random coefficients model

Determining the relationship between the response variable and time, is often of importance in a study. This is achieved by the inclusion of time t_{ij} as a predictor in the model, with a corresponding slope, say β_t. Most likely the slope will vary with subject, so it is useful to fit the subject variable as the intercept and the subject*time interaction as the slope for each patient. These two terms could reasonably be assumed to arise at random from a distribution and, thus, would be specified as random effects. This gives rise to what is called a random coefficients model. The random coefficients model is often used if the repeated measurements do not occur at fixed intervals. This type of model is different from an ordinary random effects model because the subject and subject*time effects in the model are correlated. The random effects model must be adapted to this situation to allow for correlation among these random effects. This is done using the bivariate normal distribution. The bivariate random effect becomes

$$\begin{pmatrix} b_i \\ (b * t)i \end{pmatrix} \sim N(0, G) \tag{3.51}$$

where

$$\begin{pmatrix} \sigma_b^2 & \sigma_{b,b*t} \\ \sigma_{b,b*t} & \sigma_{b*t}^2 \end{pmatrix} \tag{3.52}$$

3.4.1 Random intercepts and slopes model

A model that includes both random intercepts and random slopes is likely the most realistic type of model, although it is also the most complex. In this model, both intercepts and slopes are allowed to vary across groups, meaning that they are different in different contexts (Cohen et al., 1983). The random intercept and slopes model is given by

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 t_{ij} + b_{i1} t_{ij} + b_{i0} + e_{ij} \\ &= \beta_0 + (\beta_1 + b_{1j}) t_{ij} + b_{0j} + e_{ij} \end{aligned} \tag{3.53}$$

$$\begin{bmatrix} b_{i0} \\ b_{i1} \end{bmatrix} \sim N(0, \Omega_b) \tag{3.54}$$

$$\Omega_b = \begin{bmatrix} \sigma^2_{b_0} & \sigma_{b_{01}} \\ \sigma_{b_{01}} & \sigma^2_{b_1} \end{bmatrix} \tag{3.55}$$

$$e_{ij} \sim N(0, \sigma^2_e) \tag{3.56}$$

$$b_i \sim MVN(0, \Sigma) \tag{3.57}$$

$$Var(Y_{ij}) = \sigma^2_1 + 2\sigma_{01} t_{ij} + \sigma^2_1 t_{ij} + \sigma^2 \tag{3.58}$$

$$Cov(Y_{ij}, Y_{ik}) = \sigma^2_0 + \sigma_{01}(t_{ij}, t_{ik}) + \sigma^2_1 t_{ij} t_{ik} \tag{3.59}$$

$$Cov(Y_{ij}, Y_{lk}) = 0 \tag{3.60}$$

More complex models possible, but harder to fit.

3.5 Types of covariance structures

The distinct nature of longitudinal mixed model analysis is the covariance structure of the observed data. Here measurements made on the same subject are likely to be more similar than measurements made on different individuals. That is, repeated measurements are correlated. The covariance among repeated measures must be modeled properly in order for the analysis to be valid. Note the covariance structure is not of primary interest in the analysis, however it plays a huge role for validity of inferences. Thus more effort is usually needed at the beginning of the statistical analysis to assess and identify the best covariance structure of the data.

Until recently, analysis techniques available in computer software only offered the user limited and inadequate choices. One choice was to ignore covariance structure and make invalid assumptions. Another was to avoid the covariance structure issue by analyzing transformed data or making adjustments to otherwise inadequate analyses. Ignoring covariance structure may result in erroneous inference, and avoiding it may result in inefficient inference. Recently available mixed model methodology permits the covariance structure to be incorporated into the statistical model.

There are several specific choices of the form of the working covariance structures, but the three most commonly used covariance structures are, compound symmetry (CS), unstructured (UN) and autoregressive (AR(1)).

3.5.1 Compound Symmetry (CS):

$$Cov(Y_{ijk}, Y_{ijl}) = \sigma_1^2 \tag{3.61}$$

if $k \neq l$ then

$$Var(Y_{ijk}) = \sigma_1^2 + \sigma^2 \tag{3.62}$$

Compound symmetric structure specifies that observations on the same subject have homogeneous covariance $Cov(Y_{ijk}, Y_{ijl}) = \sigma_1^2$ and homogeneous variance $Var(Y_{ijk}) = \sigma_1^2 + \sigma^2$. The correlation function is $\rho = \frac{\sigma_1^2}{\sigma_1^2+\sigma^2}$. Note that the correlation does not depend on the value of the lag, in the sense that the correlations between two observations are equal for all pairs of observations on the same subject. This is unrealistic in longitudinal data problem in the sense that observations closer to each other are more correlated than the ones which are further apart (Littell et al., 2000). Compound symmetric structure is sometimes called variance components structure, because the two parameters σ_1^2 and σ^2 represent between-subjects and within-subjects variances, respectively. This mix of between and within-subject variances logically motivates the form of $Var(Y_{ijk})$) in many situations and implies a non-negative correlation between pairs of within-subject observations. For example if we consider three repeated measures, the compound symmetric correlation structure is given by

$$CS = \begin{bmatrix} \sigma_1^2 + \sigma^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 + \sigma^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 + \sigma^2 \end{bmatrix} \tag{3.63}$$

3.5.2 Unstructured (UN)

All the variances and covariances are different, this is the most flexible covariance structure which leads to the unstructured pattern of correlations which assumes unconstrained pair-wise correlations where each correlation is estimated from the data (the most complex model). This lets the data dictate what they should be and requires the estimate of many parameters. The more data that are used to assess the covariance structure, the less data are left to estimate the parameters of linear models. An unstructured covariance structure with three repeated measures is given by

$$UN = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix} \tag{3.64}$$

The unstructured type of correlation has an immediate disadvantage because it increases the number of parameters to estimate in the overall model hence causing possible non-convergence problems, particularly those associated with boundary values. The best way adopted to reduce the number of parameters is to assume that all the covariances along the diagonal have a constant variance. Furthermore an analysis that uses this covariance matrix will be less powerful than an analysis that uses a less parametric but more realistic structure, however the problem with that is knowing what that proper structure is (Littell et al., 2000).

3.5.3 Autoregressive, order 1 [AR (1)]:

$$Cov(Y_{ijk}, Y_{ijl}) = \sigma^2 \times \rho^{|k-l|} \tag{3.65}$$

The AR (1) covariance structure specifies homogeneous variance $Var(Y_{ijk}) = \sigma^2$. Furthermore it specifies that covariances between observations on the same subject are not equal, but decrease toward zero with increasing lag. This covariance structure is relevant for repeated measures in time, and the term autoregressive is derived from time series analysis that assumes observations are related to their own past values or history through one, two, or a higher order autoregressive (AR) process. The correlation between two responses that are m measurements apart is ρ^m, since ρ is $-1 \leq \rho \leq 1$ or to be more realistic $0 < \rho < 1$, the greater the power $m \geq 1$, the smaller the magnitude. Thus the further measurements are apart, the lower their correlation. In the case of three repeated measurements AR(1) covariance structure is given by

$$AR(1) = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix} \tag{3.66}$$

Note that the AR (1) resolves some of the objectives to the use of the compound symmetry where only one correlation parameter is needed. AR (1) are a reasonable choice for evenly or equally spaced observations.

The heterogeneous versions of AR (1) and CS, are ARH (1), CSH respectively, and they are simple extensions which assume the variances along the diagonal of the matrix are not equal.

3.5.4 Spatial covariance structures

Spatial correlation structures are very useful for ”unequally spaced longitudinal data which can be viewed as a spatial process in one dimension” (Littell et al., 2000). The advantage of using spatial correlation structures is that they calculate the actual distance between measurements themselves without making any assumptions. Table 3.1 gives some of the spatial covariance structures, each one specifying and defining how fast the correlations decrease as functions of the distances between measurements d_{ij}.

Table 3.1: Spatial covariance structures

Structure	Example
Power	$\sigma^2 \begin{bmatrix} 1 & \rho^{d_{12}} & \rho^{d_{13}} \\ \rho^{d_{12}} & 1 & \rho^{d_{23}} \\ \rho^{d_{13}} & \rho^{d_{23}} & 1 \end{bmatrix}$
Linear	$\sigma^2 \begin{bmatrix} 1 & 1-\rho_{d_{12}} & 1-\rho_{d_{13}} \\ 1-\rho_{d_{12}} & 1 & 1-\rho_{d_{23}} \\ 1-\rho_{d_{13}} & 1-\rho_{d_{23}} & 1 \end{bmatrix}$
Exponential	$\sigma^2 \begin{bmatrix} 1 & exp(-d_{12}\backslash\rho) & exp(-d_{13}\backslash\rho) \\ exp(-d_{12}\backslash\rho) & 1 & exp(-d_{23}\backslash\rho) \\ exp(-d_{13}\backslash\rho) & exp(-d_{23}\backslash\rho) & 1 \end{bmatrix}$
Gaussian	$\sigma^2 \begin{bmatrix} 1 & exp(-d_{12}^2\backslash\rho^2) & exp(-d_{13}^2\backslash\rho^2) \\ exp(-d_{12}^2\backslash\rho^2) & 1 & exp(-d_{23}^2\backslash\rho^2) \\ exp(-d_{13}^2\backslash\rho^2) & exp(-d_{23}^2\backslash\rho^2) & 1 \end{bmatrix}$
Spherical	$\sigma^2 \begin{bmatrix} 1 & \left[1-(\frac{3d_{12}}{2\rho})+(\frac{d_{12}^3}{2\rho^3})\right] & \left[1-(\frac{3d_{13}}{2\rho})+(\frac{d_{13}^3}{2\rho^3})\right] \\ \left[1-(\frac{3d_{12}}{2\rho})+(\frac{d_{12}^3}{2\rho^3})\right] & 1 & \left[1-(\frac{3d_{12}}{2\rho})+(\frac{d_{12}^3}{2\rho^3})\right] \\ \left[1-(\frac{3d_{12}}{2\rho})+(\frac{d_{12}^3}{2\rho^3})\right] & \left[1-(\frac{3d_{12}}{2\rho})+(\frac{d_{12}^3}{2\rho^3})\right] & 1 \end{bmatrix}$

3.5.5 Choosing the best covariance structure

Ideally the covariance structure should be known from previous work or subject matter considerations. Selecting a structure that is too simple increases the Type I error rate and selecting a structure that is too complex sacrifices power and efficiency Littell et al. (2000).

To choose the best covariance structure use the information criteria (IC). Given a set of candidate covariance structures for the data, the preferred structure is the one with the minimum Akaike information criterion (AIC) value. AIC rewards goodness of fit (as assessed by the likelihood function), but it also includes a penalty that is an increasing function of the number of estimated parameters. The penalty discourages over fitting, because increasing the number of parameters in the model almost always improves the goodness of the fit. Alternatively one can use the Bayesian information criterion (BIC), which is an increasing function of the error variance σ_e^2 and an increasing function of k. That is, unexplained variation in the dependent variable and the number of explanatory variables increases the value of BIC. Hence, lower BIC implies either fewer explanatory variables, better fit, or both (Verbeke & Molenberghs, 2000). It is important to keep in mind that the BIC can be used to compare estimated models only when the numerical values of the dependent variable are identical for all estimates being compared. The models being compared need not be nested, unlike the case when models are being compared using an F-test or a likelihood ratio test. The interest in the covariance structure is not for its own right but for obtaining a good model for the covariance structure so that computations and inferences about the fixed effects are valid (Verbeke & Molenberghs, 2000).

Table 3.2: Information criteria for AIC and BIC

Criteria	Structure	Small is better	Large is better
AIC	$log(n)k - 2log(\hat{L})$	$-2l + 2d$	$l - d$
BIC	$2k - 2log(\hat{L})$	$-2l + 2dlog(n)$	$l - 0.5 * d * logn$

Where, $\hat{L}$ is the maximized value of the likelihood function of the model, n is the sample size and k is the number of free parameters to be estimated.

3.5.6 Some graphical guides

Consider fitting UN then plot the covariance for each starting time which can provide pertinent diagnostic information. That is, plot lag 1, covariance, lag 2, covariance, and so on, for the errors starting at 0, 1, 2 and so on. Then if there is a linearly declining covariance without increasing lags one might fit an AR (1) covariance structure, and if the lines overlay each other, then a constant variance would be appropriate otherwise the heterogeneous version of the structure would be best to use. However this is not the best method to use according to Kincaid (2005).

3.6 Model diagnostics

Sometimes the assumption of the random effects can be violated by longitudinal data. Thus it is very important to check the assumptions of the model after it has been fitted, that is, to check if the normality assumption for the random effects and the behavior of residuals is appropriate. In most cases, histograms and scatter plots of the random components and the residuals are often used for the diagnostic purposes. In particular, the scatter plots are used to pinpoint outlying observations which arise from subjects that seem to evolve differently from the other subjects in the sample, but the histograms of the residuals can be used to check for the normality of the random effects and the error terms. The empirical distribution of the data (the histogram) should be bell-shaped and resemble the normal distribution. This might be difficult to see if the sample is small. In this case one might proceed by regressing the data against the quantiles of a normal distribution with the same mean and variance as the sample. Lack of fit to the regression line suggests a departure from normality. One can also use the probability plots, and tests using the Shapiro-Wilk test and the Kolmogorov-Smirnov test to assess the normality assumptions. Specifically the W-statistic (in the Shapiro-Wilk test), suggested by Shapiro & Wilk (1965) has been shown to be a good omnibus test of normality.

3.6.1 Residual diagnostics

In order to validate model assumptions and to detect outliers and potentially influential data points residuals are employed. A residual is defined as the difference between an observed quantity and its predicted value. In the mixed model a marginal residual is the difference between the observed data and the estimated (marginal) mean that is

$$r_{mi} = y_i - x_i^{'}\hat{\beta} \tag{3.67}$$

and a conditional residual is the difference between the observed data and the predicted value of the observation that is,

$$r_{ci} = y_i - x_i^{'}\hat{\beta} - z_i^{'}\hat{b}_i \tag{3.68}$$

For a model without random effects b, the marginal and conditional residuals coincide. The name conditional residual stems from the fact that $x_i^{'}\hat{\beta} + z_i^{'}\hat{b}_i$ is the conditional mean of y_i. According to Schabenberger (2005), the raw residuals, r_{mi} and r_{ci} are usually not well suited for these purposes. "Even if the true model errors are uncorrelated and have equal variance, the residuals will exhibit correlations and their variances will differ. The interpretation of raw residuals is further made difficult if the variances of the observations differ. A data point with a smaller raw residual may be more troublesome than a data point with a large residual, if the variance of the former observation is less". Different types of residuals are presented in Table 3.3 .

Table 3.3: Summary of available residuals

Type of Residual	Marginal	Conditional
Raw	$r_{mi} = y_i - x_i^{'}\hat{\beta}$	$r_{ci} = y_i - x_i^{'}\hat{\beta} - z_i^{'}\hat{b}_i$
Studentized	$r_{mi}^{student} = \frac{r_{mi}}{\sqrt{\widehat{V}(r_{mi})}}$	$r_{ci}^{student} = \frac{r_{ci}}{\sqrt{\hat{V}[r_{ci}]}}$
Pearson	$r_{mi}^{pearson} = \frac{r_{mi}}{\sqrt{\hat{V}[Y_i]}}$	$r_{ci}^{pearson} = \frac{r_{ci}}{\sqrt{\hat{V}[Y_i \vert \hat{b}_i]}}$
Scaled	$\hat{C}_{rm}^{-1}$	

3.6.2 Influence diagnostics

Influential observations refer to those observations that appear to have fairly large influence on the parameter estimates.

3.6.3 Overall influence

Typically the subscript U denotes quantities obtained without the observations in the set U. An overall influence statistic measures the change in the objective function being minimized Schabenberger (2005) . Beckman et al. (1987) refers to it as the likelihood displacement (LD). According to Schabenberger (2005), the likelihood and restricted likelihood distances (RLD) are then given by

$$LD_{(U)} = 2\{l(\hat{\psi}) - l(\hat{\psi}_{(U)})\} \tag{3.69}$$

$$RLD_{(U)} = 2\{l_R(\hat{\psi}) - l_R(\hat{\psi}_{(U)})\} \tag{3.70}$$

According to Schabenberger (2005) ”the likelihood distance gives the amount by which the log-likelihood of the full data changes if one were to evaluate it at the reduced-data estimates. It is obtained by evaluating the likelihood function based on the full data set (containing all n observations) at the reduced-data estimates. The likelihood distance is a global, summary measure, expressing the joint influence of the observations in the set U on all parameters in that were subject to updating”.

3.6.4 Change in parameter estimates

Cooks distance measures the effect of deleting a given observation and was first introduced by Cook (1977). Schabenberger (2005) states that the difference between Cooks distance (D) and multivariate DFFITS (also known as MDFFITS) is that the latter uses an externalized estimate of the variance of the parameter estimates, while Cooks distance does not. For the fixed effects, the two statistics are

$$D(\beta) = (\hat{\beta} - \hat{\beta}_{(u)})' v\hat{a}r(\hat{\beta})^{-1}(\hat{\beta} - \hat{\beta}_{(u)})/rank(X) \tag{3.71}$$

$$MDFFITS = (\hat{\beta} - \hat{\beta}_{(u)})' v\hat{a}r(\widehat{\beta_{(u)}})^{-1}(\hat{\beta} - \hat{\beta}_{(u)})/rank(X) \tag{3.72}$$

For both statistics, large values, according to Schabenberger (2005) "indicate that the change in the parameter estimate is large relative to the variability of the estimate . If the covariance parameters are updated during influence analysis, similar statistics can be computed for $\hat{\theta}$. However, the $D(\theta)$ and $MDFFITS(\theta)$ statistics do not involve division by a matrix rank".

3.6.5 Change in precision of estimates

According to Schabenberger (2005) "the effect on the precision of estimates is separate from the effect on the point estimates. Data points that have a small Cooks D, for example, can still greatly affect hypothesis tests and confidence intervals, if their influence on the precision of the estimates is large".
where q denotes the rank of $Var(\theta)$. The variance matrix that is used in the computation of COVTRACE and COVRATIO for covariance parameters is obtained from the inverse Hessian matrix Schabenberger (2005).

3.6.6 Effect on fitted and predicted values

Following Schabenberger (2005) "the MIXED procedure computes the following statistics to measure influence on fitted and predicted values. The PRESS residual Allen (1974) is the difference between the observed value and the predicted (marginal) mean, where the predicted value is obtained without the observations in question. Formally,

$$\hat{e}_{i(U)} = y_i - x_i'\hat{\beta}_{(U)} \tag{3.73}$$

If you compute the influence of individual observations, the MIXED procedure reports these PRESS residuals. When removing sets of observations, PROC MIXED computes the PRESS statistic. This statistic is the sum of the squared PRESS residuals in a deletion set,

$$PRESS_{(U)} = \sum_{i \in U} \hat{e}_{i(U)} \tag{3.74}$$

The effect of observations on fitted values can be measured by the DFFITS statistic of Belsley et al. (1980). Recall that a DFFIT measures the change in predicted values due to removal of a single data point. If this change is standardized by the externally estimated standard error of the predicted value in the full data, then the DFFITS statistic is given by"

$$DFFITS_s = (\hat{y}_i - \hat{y}_{i(U)})/ese(\hat{y}_i) \tag{3.75}$$

3.6.7 Summary

The theory on linear mixed models and the estimation methods was discussed. Types of covariance structures were briefly described and the best structure can be selected by choosing the model with a structure that gives the lowest Akaike Information Criteria (AIC). Furthermore different types of residuals for model diagnostics were briefly discussed

Chapter 4

Survival analysis

4.1 Introduction

In this chapter we examine the survival analyses methodology, in particular the Cox proportional hazards model is discussed as well as the Kaplan-Meier and the log-rank test.

Survival analysis examines and models the time it takes for an event to occur and focuses on the distribution of survival times Fox (2002). According to Collett (2015), survival data are not symmetrically distributed and are strictly positively skewed. In a survival analysis, the time an individual has survived over some follow-up period is known as a survival time an event such as death is known as failures Kleinbaum & Klein (2005).The basic goals of survival analysis by Kleinbaum & Klein (2005) include;

- The estimation and interpretation of survivor and/or hazard functions from survival data
- Comparing survivor and/or hazard functions.

A key analytical problem in survival analysis is censoring. According to Kleinbaum & Klein (2005), censoring occurs when there is some information about individual survival time, but the survival time is not known exactly . There are three types of censoring, following Kleinbaum & Klein (2005) these types are;

1. Right censoring;
   This type of censoring occurs if the event occurs after the observed survival time (after the study is finished). It follows that right censored survival time is less than the actual survival time.

2. Left censoring;
   When the actual survival time is less than or equal to the observed survival time is known as left censoring.

3. interval censoring;
   interval censoring occurs when the individual is known to have experienced an event within an interval of time but the actual survival time is not known.

Some of the reasons for censoring includes an individual not experiencing the event before the end of the study, or lost to follow-up, perhaps that person has relocated, lastly a person could withdraw from the study or die but death not related to the event of interest (Kleinbaum & Klein, 2005).

4.2 The survivor function and the hazard function

Let T represent survival time. T can take any non-negative value and is regarded as a random variable with cumulative distribution function;

$$F(t) = P_r(T \leq t) = \int_0^t f(u)du \tag{4.1}$$

and a probability density function given by

$$f(t) = F^{'}(t) = -S^{'}, t \geq 0 \tag{4.2}$$

The survivor function is a decreasing function and it gives the probability that the random variable T exceeds the specified time t (Kleinbaum & Klein, 2005). The survivor function is given by

$$S(t) = P(T > t) = 1 - F(t) \tag{4.3}$$

The hazard function, denoted by $h(t)$, assesses the instantaneous risk of failure, given that the individual has survived up to time t (Fox, 2002).The function typically refers to a hazard rate, the instantaneous death rate, or the force of mortality (Collett, 2015), the hazard function is given by

$$\begin{aligned} h(t) &= \lim_{\delta t \to 0} \frac{P(t \leq T < t + \delta t | T \geq t)}{\delta t}; t \geq 0 \\ &= \lim_{\delta t \to 0} \frac{P(t \leq T < t + \delta t)}{\delta t P(T \geq t)} \\ &= \lim_{\delta t \to 0} \left[\frac{F(t + \delta t) - F(t)}{\delta t} \right] \times \left[\frac{1}{P(T \geq t)} \right] \\ &= \frac{f(t)}{S(t)} \end{aligned} \tag{4.4}$$

It then follows that

$$h(t) = \frac{f(t)}{S(t)} = \frac{-dlogS(t)}{dt} \tag{4.5}$$

and so

$$S(t) = exp\left[- \int_0^t h(u)du \right] = exp(-H(t)), t \geq 0, \tag{4.6}$$

all these functions give a mathematical equivalent specification of the distributions of the survival time T: If one of them is known, then the other two can be determined. Commonly used parametric functions include the Weibull and exponential distribution.

4.3 The Kaplan-Meier estimate of the survivor function

According to Collett (2015), the Kaplan-Meier (KM)estimate is a generalization of the empirical survivor function that accommodates censored observations and it is based on individual (ungrouped) survival times. Following Collett (2015) consider a sample with n individuals with observed survival times $t_1, \cdots, t_n$. Suppose that there are $m \leq n$ recorded event times then the ranked survival times are $t_{(1)} < t_{(2)} < \cdots < t_{(m)}$. Let d_j be the number of deaths at t_j. Let n_j be the number alive (those at risk) just before d_j for $j = 1, 2, \cdots, m$. The probability that an individual dies during the interval from $t_{(j)} - \delta$ to $t_{(j)}$ is estimated by d_j/n_j and corresponding estimated probability of survival through that interval is then $\big(n_j - d_j/n_j\big)$. Thus the Kaplan-Meier estimator is defined as follows

$$\hat{S}(t) = \prod_{t:t_{(j)} \leq t} \left(1 - \frac{d_j}{n_j} \right) \tag{4.7}$$

4.3.1 Non-parametric maximum likelihood

Consider the likelihood contribution of a case that experiences an event or censored at time t_j. Taking c_j to represent the number of cases censored between $t_{(j-1)}$ and $t_{(j)}$, and taking d_j to be the number of cases which die or experience the event at $t_{(j)}$, then by Hanagal (2011) the likelihood function is given by

$$L = \prod_{j=1}^{m} \Big[S(t_{(j-1)}) - S(t_{(j)})\Big]^{d_j} \Big[S(t_{(j)})\Big]^{c_j} \tag{4.8}$$

The conditional probability of surviving from $S(t_{(j-1)})$ to $S(t_{(j)})$ is $\pi_j = S(t_{(j)})/S(t_{(j-1)})$ where $S(t_{(j)})$ can be written as $S(t_{(j)}) = p_1 \times \cdots \times p_j$, thus the likelihood then becomes

$$L = \prod_{j=1}^{m} (1-p_j)^{d_j} p_j^{c_j} (p_1, \cdots, p_{j-1})^{d_j + c_j}. \tag{4.9}$$

Let $\sum_{j>i}(d_j + c_j)$ denote the total number exposed to risk at $t_{(j)}$, thus collecting the terms on each π_j, we get a binomial likelihood given by

$$L = \prod_{j=1}^{m} (1-p_j)^{d_j} p_j^{n_j - d_j} \tag{4.10}$$

The maximum likelihood estimator of π_j is then given by

$$\hat{p}_j = \frac{n_j - d_j}{n_j} = 1 - \frac{d_j}{n_j}.$$

The K-M estimator follows from multiplying these conditional probabilities.

$$var(\hat{p}_j) = \frac{p_j(1-p_j)}{n_j} \tag{4.11}$$

which can be estimated by

$$v\hat{a}r(\hat{p}_j) = \frac{\hat{p}_j(1-\hat{p}_j)}{n_j}. \tag{4.12}$$

According to Collett (2015), the Kaplan-Meier estimate of the survivor function of any value of t can be written as

$$\hat{S}(t) = \prod_{j=1}^{m} \hat{p}_j$$

for $j = 1, \ldots, m$, taking logarithms

$$\log\hat{S}(t) = \sum_{j=1}^{m} \log\hat{p}_j \tag{4.13}$$

and so the variance of $\log\hat{S}(t)$ is given by

$$var\big[\log\hat{S}(t)\big] = \sum_{j=1}^{m} var\big(\log\hat{p}_j\big) \tag{4.14}$$

To obtain the variance $\log\hat{p}$, we make use of a general result for the approximate variance of a function of a random variable given by

$$var\big[f(X)\big] \approx \left\{\frac{df(X)}{dX}\right\}^2 var(X). \tag{4.15}$$

Using equation 4.15, the approximate variance of $\log\hat{p}$ is $var(\hat{p}_j)/\hat{p}_j^2$, and using equation 4.11, the approximate estimated variance of $\log\hat{p}$ is $(1-\hat{p}_j)/(n_j\hat{p}_j)$, which on substitution for $\hat{p}_j$, reduces to

$$\frac{d_j}{n_j(n_j - d_j)}. \tag{4.16}$$

Equation 4.14 then becomes

$$var\big[\log\hat{S}(t)\big] \approx \sum_{j=1}^{m} \frac{d_j}{n_j(n_j - d_j)}. \tag{4.17}$$

and a further application of the result in equation 4.15 gives

$$var\big[\log\hat{S}(t)\big] \approx \frac{1}{\big[\hat{S}(t)\big]^2} var\Big[\hat{S}(t)\Big] \tag{4.18}$$

so that

$$var\big[\log\hat{S}(t)\big] = \sum_{j=1}^{m} \frac{d_j}{n_j(n_j - d_j)}. \tag{4.19}$$

Using delta method to get the variance of the survivor function from the variance of its log, we get

$$var\big(\hat{S}(t)\big) = \hat{S}(t)^2 \sum_{j=1}^{m} \frac{d_j}{n_j(n_j - d_j)}. \tag{4.20}$$

Finally, the standard error of the Kaplan-Meier estimate of the survivor function is

$$se\big[\hat{S}(t)\big] \approx \hat{S}(t)\left\{\sum_{j=1}^{m} \frac{d_j}{n_j(n_j - d_j)}\right\}^{\frac{1}{2}}. \tag{4.21}$$

This result is known as *Greenwood's formula*.

4.3.2 Limitations of Kaplan-Meier

- This method is mainly descriptive
- Does not control for covariates
- Suitable for categorical predictors
- Can not accommodate time-dependent variables

4.4 The Log- Rank test (Mantel-Haenszel statistic)

The log-rank test also known as the Mantel-Haenszel statistic is a large-sample chi-square test that uses a statistic that provides an overall comparison of the K-M curves being compared as its test criterion. The log-rank test makes use of observed versus expected cell counts over categories of outcomes (Kleinbaum & Klein, 2005). The categories for the logrank statistic are defined by each of the ordered failure times for the entire set of data being analyzed. Suppose the two groups are denoted by 1 and 2, and that there are k distinct times, $t_1 < t_2 < \cdots < t_k$ across the two groups. The test uses a conditional argument based on the number at risk of failing just prior to each observed failure time. Suppose that at time t_j there are d_j deaths and n_j at risk in total, with d_{1j} and d_{2j} deaths and n_{1j} and n_{2j} at risk in group 1 and 2 respectively such that $d_{1j} + d_{2j} = d_j$ and $n_{1j} + n_{2j} = n_j$, at each death time t_j. This scenario is summarized in a 2×2 table below

Table 4.1: Number of deaths at t_j in each of two groups of individuals

Group	Number of deaths	Number surviving	Number at risk (Total)
1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
2	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
Total	d_j	$n_j - d_j$	n_j

If the assumption that the two groups are the same is true, then according to Collett (2015), the expected number of deaths at any time follows the hyper-geometric distribution and is given by

$$E(d_{1j}) = e_{1j} = \frac{n_{1j} d_j}{n_j} \tag{4.22}$$

and the variance

$$var(d_{1j}) = v_{1j} = \frac{n_{1j}n_{2j}d_j(n_j - d_j)}{n_j^2(n_j - 1)}. \tag{4.23}$$

Summing the various measures over the death times gives

$$O_1 = \sum_j d_{1j}, \qquad E_1 = \sum_j e_{1j}, \qquad V_1 = \sum_j v_{1j}.$$

Thus the test statistic is then given by

$$\chi_1^2 = \frac{(O_1 - E_1)^2}{V_1}. \tag{4.24}$$

This statistic summarizes the extent to which the observed survival times in the two groups of data deviate from those expected under the null hypothesis of no group differences (Collett, 2015). The larger its values is, the greater the evidence against the null hypothesis. V_1 is the variance of the difference $O_1 - E_1$ assuming independent event times. Alternatively, assuming the deviations $d_{1j} - e_{1j}$ for $j = 1, 2, \cdots, k$, are independent,

$$Z = \frac{O_1 - E_1}{\sqrt{V_1}}. \tag{4.25}$$

should have an approximately standard normal distribution, and at 5% level of significance the null hypothesis is rejected if the observed Z is greater than 1.96. The log rank test can be generalized to test equality of death rates in $s > 2$ groups. The test statistic, with $(s - 1)$ degrees of freedom, would then be given by

$$\chi_1^{s-1} = \frac{(O_1 - E_1)^2}{V_1} + \frac{(O_2 - E_2)^2}{V_2} + \cdots \tag{4.26}$$

If the calculated value exceeds the table value at 5% significant level, we reject the null hypothesis of no group differences survivor or hazard functions.

4.5 The Cox proportional hazards (PH) model

The Cox proportional-hazards regression model is broadly applicable and the most common tool used for studying the dependency of survival time on predictor variables (Fox, 2002). The Cox proportional hazard model which was introduced by Cox (1972) is given by

$$\begin{aligned} h(t|X) &= h_0(t)exp(\beta_1 X_1 + \cdots + \beta_p X_p) \\ &= h_0(t)exp(\boldsymbol{\beta}^{'}\boldsymbol{X}) \end{aligned} \tag{4.27}$$

This model is referred to as a semi-parametric model since it does not assume any form of probability distribution. This model allows for fixed covariates that do not change over time (Cox, 1972) and parameters are estimated by maximizing the partial-likelihood (Cox, 1975). In equation 4.27, given a set of covariates in $\boldsymbol{X} = (X_1 \ldots X_p)^{'}$, $h_0(t)$ is known as a baseline hazard function, and $\boldsymbol{\beta} = (\beta_1 \ldots \beta_p)^{'}$ is a vector of regression coefficients. The key property of the model is that $h_0(t)$ is left completely unspecified. However in specifying the model in equation 4.27, the understanding is that link between the hazard and the covariates is correctly specified and all the covariates necessary are in the model. This is almost impossible in practice hence the model has been extended in many ways to account for unobserved covariates that could have been included in the model.

4.5.1 The Cox PH assumption

The PH assumption requires the hazard ratio to be constant over time, or equivalently, that the hazard for one individual is proportional to the hazard for any other individual, where the proportionality constant is independent of time Kleinbaum & Klein (2005), as shown in equation 4.28

$$\widehat{HR} = \frac{\hat{h}(t, \boldsymbol{X_1})}{\hat{h}(t, \boldsymbol{X_2})} = \hat{\theta}, \tag{4.28}$$

The hazard ratio of two individuals with different hazards and time-invariant covariates say X_1 and X_2 is

$$\begin{aligned} \widehat{HR} &= \frac{h_0(t)exp(\hat{\boldsymbol{\beta}}^{'}\boldsymbol{X_1})}{h_0(t)exp(\hat{\boldsymbol{\beta}}^{'}\boldsymbol{X_2})} \\ &= exp[\hat{\boldsymbol{\beta}}^{'}(\boldsymbol{X_2} - \boldsymbol{X_2})]. \end{aligned} \tag{4.29}$$

The hazard functions are proportional, implying that the ratio of the two hazards is constant regardless of time t.

4.5.2 The partial likelihood function for survival times

Suppose that k of the n individual survival times are uncensored and that $n - k$ are right censored. Let $t_{(1)} < t_{(2)} < \cdots < t_{(k)}$ and $X_1, \cdots, X_k$ denote the ordered, distinct event times and covariates respectively Lee & Wang (2003). $R(t_i)$ represents the set of subjects at risk at event time t_i. For a particular observed event at time $t_{(i)}$ given $R(t_i)$, the probability that the event is on the individual observed is

$$\frac{exp\big(\sum_{j=1}^{p}\boldsymbol{\beta}_j \boldsymbol{X}_{j(i)}\big)}{\sum_{l\in R(t_i)} exp\big(\sum_{j=1}^{p}\boldsymbol{\beta}_j \boldsymbol{X}_{j(l)}\big)} = \frac{exp(\boldsymbol{\beta}'\boldsymbol{X}_{(i)})}{\sum_{l\in R(t_i)} exp(\boldsymbol{\beta}'\boldsymbol{X}_{(l)})}. \tag{4.30}$$

Since each observed event time contributes a factor as given above, the overall partial likelihood function is

$$\begin{aligned} L\beta &= \prod_{i=1}^{k}\frac{exp\big(\sum_{j=1}^{p}\boldsymbol{\beta}_j \boldsymbol{X}_{j(i)}\big)}{\sum_{l\in R(t_i)} exp\big(\sum_{j=1}^{p}\boldsymbol{\beta}_j \boldsymbol{X}_{j(l)}\big)} \\ &= \prod_{i=1}^{k}\frac{exp(\boldsymbol{\beta}'\boldsymbol{X}_{(i)})}{\sum_{l\in R(t_i)} exp(\boldsymbol{\beta}'\boldsymbol{X}_{(l)})}. \end{aligned} \tag{4.31}$$

The maximum partial likelihood estimate of the regression coefficients can be found by setting to zero the derivative of the log of the expression in equation 4.31 by differentiating with respect to β and solving the resulting simultaneous equations iteratively using numerical methods such as the Newton-Raphson procedure.

According to Kleinbaum & Klein (2005), if time-dependent variables are considered, equation 4.29 will no longer satisfy the PH assumption, instead the extended Cox model with the partial likelihood evaluated at each event time in the form is used. The extended Cox model accommodates two types of time-dependent covariates, namely, time varying covariates resulting from repeated observations at different time points prior to the event or censoring and covariates whose values change according to a mathematical function of time. This partial likelihood is given by

$$\begin{aligned} L\beta &= \prod_{i=1}^{k}\frac{exp\big(\sum_{j=1}^{p}\boldsymbol{\beta}_j \boldsymbol{X}_{j(i)}(t_i)\big)}{\sum_{l\in R(t_i)} exp\big(\sum_{j=1}^{p}\boldsymbol{\beta}_j \boldsymbol{X}_{j(l)}(t_i)\big)} \\ &= \prod_{i=1}^{k}\frac{exp(\boldsymbol{\beta}'\boldsymbol{X}_{(i)}(t_i))}{\sum_{l\in R(t_i)} exp(\boldsymbol{\beta}'\boldsymbol{X}_{(l)}(t_i))}. \end{aligned} \tag{4.32}$$

4.5.3 Cox PH assumption model checking

The adequacy of the fitted model needs to be assessed after a model has been fitted to an observed set of survival data. Many model-checking procedures are based on graphical methods, adding time-dependent covariates as well as a formal test based on residuals. Residuals refer to the values that can be calculated for each individual in the study, and have the feature that their behavior is known, when the model is satisfactory (Collett, 2015)

4.5.4 Graphical methods

There are two graphical approaches for checking the PH assumption, namely, comparing log-log survival curves and comparing observed versus predicted survival curves.

- Comparing log-log survival curves

  Recall that the survival function is given by,

  $$S(t,X) = [S_0(t)]exp\big(\sum_{i=1}^{p} \boldsymbol{\beta}_i \boldsymbol{X}_i\big) \tag{4.33}$$

  where $S_0(t)$ denotes the baseline survival function. After taking logarithm twice we get

  $$ln[-lnS(t,X)] = \sum_{i=1}^{p} \boldsymbol{\beta}_i \boldsymbol{X}_i + ln[-lnS_0(t)] \tag{4.34}$$

  Then the difference in log-log curves corresponding to two different subjects $\boldsymbol{X}_1 = (x_{11}, \ldots, x_{1p})$ and $\boldsymbol{X}_2 = (x_{21}, \ldots, x_{2p})$ is given by

  $$ln[-lnS(t,\boldsymbol{X}_1)] - ln[-lnS(t,\boldsymbol{X}_2)] = \sum_{i=1}^{p} \boldsymbol{\beta}_i (x_{1i} - x_{2i}) \tag{4.35}$$

  which does not depend on t provided the two covariate vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_1$ are not time dependent. This relationship is very helpful to help us identify situations where we may have proportional hazards. If a PH model is appropriate for a given set of predictors, one should expect that empirical plots of log-log survival curves for different individuals to be approximately parallel (Kleinbaum & Klein, 2005). This method does not work well for continuous predictors or categorical predictors that have many levels because the group becomes cluttered. Furthermore, the curves are sparse when there are few time points and it may be difficult to tell how close to parallel is close enough (Therneau & Grambsch, 2013). "Similarly, although the PH assumption may not be violated, the log-minus-log curves are rarely perfectly parallel in practice, and tend to become sparse at longer time points, and thus less precise. It is not possible to quantify how close to parallel is close enough, and thus how proportional the hazards are. The decision to accept the PH hypothesis often depends on whether these curves cross each other. As a result, the decision to accept the PH hypothesis can be subjective and conservative" (Schemper, 1992), since one must have strong evidence "crossing lines" to conclude that the PH assumption is violated. To alleviate these limitations, Martinussen & Scheike (2007) suggest providing standard errors to these plots, however this approach can be computationally intensive and is not directly available in standard computer programs.

- Comparing observed versus predicted survival curves

  The use of observed versus predicted plots to assess the PH assumption is the graphical analog of the goodness-of-fit (GOF) testing approach. If for each category of the predictor being assessed, the observed and expected plots are close to one another, one can then conclude that the PH assumption is satisfied. If, however, one or more categories show quite discrepant observed and expected plots, one can conclude that the PH assumption is violated (Kleinbaum & Klein, 2005)

4.5.5 Residuals for the Cox regression model

Several residuals have been proposed for the assessment of the Cox PH model adequacy. For this project we will briefly discuss the Martingale residuals, Deviance residuals, Schoenfeld residuals and Score residuals. Others include Cox-Snell residuals and Modified Cox-Snell residuals but we will not discuss these here.

- Martingale residuals

  $$r_{M_i} = \delta_i - r_{C_i} \tag{4.36}$$

  Eqauation 4.36 is known as a martingale residuals, since they can be derived using martingale methods (Collett, 2015). Following Collett (2015), Martingale residuals can take values between $-\infty$ and ∞, with the residuals for censored observations, where $\delta_i = 0$ being negative. It can be shown that these residuals sum to zero and in large samples the Martingale residuals are uncorrelated with one another and have an expected value of zero (Collett, 2015). In this respect, they have properties similar to those possessed by residuals encountered in linear regression analysis, however they are not symmetrically distributed about zero, even when the fitted model is correct.

- Deviance residuals

  The Deviance residuals are symmetrically distributed about zero and were first introduced by Therneau et al. (1990)

$$r_{D_i} = sgn(r_{M_i})\big[-2\big(r_{M_i} + \delta_i log(\delta_i - r_{M_i})\big)\big]^{\frac{1}{2}}, \tag{4.37}$$

  Where r_{M_i} is the martingale residual for the i^{th} individual, and the function $sgn(\cdot)$ is the sign function. This is the function that takes the value $+1$ if its argument is positive and -1 if negative. Thus $sgn(r_{M_i})$ ensures that the deviance residuals have the same sign as the Martingale residuals (Collett, 2015). The deviance is a statistic that is used to summarize the extent to which the fit of a model of current interest deviates from that of a model which is a perfect fit to the data. This latter model is called the saturated or full model in which β coefficients are allowed to be different for each individual. The statistic is given by

$$D = -2(log\hat{L}_c - log\hat{L}_f),$$

  where $\hat{L}_c$ is the maximized partial likelihood under the current model and $\hat{L}_f$ is the maximized partial likelihood under the full model. The smaller the value of the deviance, the better the model (Collett, 2015).

- Schoenfeld Residuals

  These residuals are different to the ones mentioned above in a sense that, there is not a single value of the residual for each individual, but a set of values, one for each explanatory variable included in the fitted Cox regression model. Schoenfeld Residuals were originally known as partial residuals. The i^{th} partial or Schoenfeld Residual for X_j, the X_j explanatory variable in the model is given by

$$r_{p_{ji}} = \delta_i(x_{ji} - \hat{a}_{ji}), \tag{4.38}$$

  where x_{ji} is the value of the j^{th} explanatory variable, $j = 1, \ldots, p$ for the i^{th} individual in the study,

$$\hat{a}_{ji} = \frac{\sum_{l \in R(t_i)} x_{ji} exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_l)}{\sum_{l \in R(t_i)} exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_l)} \tag{4.39}$$

  and $R(t_i)$ is the set of all individuals at risk at time t_i. Note that non-zero values of these residuals only arise for uncensored observations. Moreover, if the largest observation in a sample of survival times is uncensored, the value of $\hat{a}_{ji}$ for that observation, from equation 4.39, will be equal to x_{ji} and so $r_{pji} = 0$ (Collett, 2015).

The i^{th} Schoenfeld residual, for the explanatory variable X_j is an estimate of the i^{th} component of the first derivative of the logarithm of the partial likelihood function with respect to β_j which is given by

$$\frac{\partial logL(\beta)}{\partial\beta_j} = \sum_{i=1}^{n} \delta_i(x_{ji} - \hat{a}_{ji}), \tag{4.40}$$

where

$$a_{ji} = \frac{\sum_l x_{jl} exp(\boldsymbol{\beta}' \boldsymbol{x}_l)}{\sum_{l\in R(t_i)} exp(\boldsymbol{\beta}' \boldsymbol{x}_l)} \tag{4.41}$$

The i^{th} term in this summation, evaluated at $\hat{\beta}$ is then the Schoenfeld residual for X_j given in equation 4.38. Since the estimates of the β's are such that

$$\left(\frac{\partial logL(\beta)}{\partial\beta_j}\right)|_{\hat{\beta}} = 0,$$

the Schoenfeld residuals must sum to zero. These residuals also have the property that, in a large sample, the expected values of r_{pji} is zero and they are uncorrelated with one other. The scaled version of the Schoenfeld residuals, proposed by Grambsch & Therneau (1994) is more effective in detecting departures from the assumed model and is given by

$$\boldsymbol{r}^*_{p_i} = rVar(\hat{\beta})\boldsymbol{r}_{p_i},$$

where

$$\boldsymbol{r}_{p_i} = (r_{p1i}, \ldots, r_{ppi})'$$

and r is the number of deaths among the n individuals, and $Var(\hat{\beta})$ is the variance-covariance matrix of the parameter estimates in the fitted Cox regression model. These scaled Schoenfeld residuals are therefore quite straight forward to compute (Collett, 2015).

- Score Residuals

Just like the Schoenfeld Residuals, the Score Residuals are also obtained from the first derivative of the logarithm of the partial likelihood function with respect to the parameter β_j, $j = 1, \cdots, p$. However, equation 4.40 is now expressed as

$$\frac{\partial logL(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \left\{ \delta_i(x_{ji} - a_{ji}) + exp(\boldsymbol{\beta}' \boldsymbol{x}_i) \sum_{t_r \leq t_i} \frac{(a_{jr} - x_{ji})\delta_r}{\sum_{l \in R(t_r)} exp(\boldsymbol{\beta}' \boldsymbol{x}_l)} \right\} \tag{4.42}$$

where x_{ji} value of the j^{th} explanatory variable. δ_i is the event indicator which is zero for censored observations and unity otherwise, a_{ji} is given in equation 4.41, and R_{t_r} is the risk set at time t_r. In this formulation, the contribution of the i^{th} observation to the derivative only depends on information up to time t_i. In other words, if the study was actually concluded at time t_i the i^{th} component of the derivative would be unaffected. Residuals are then obtained as the estimated value of the n component of the derivative. From equation 4.42 the first derivative of the logarithm of the partial likelihood function, with respect to β_j, is the efficient score for β_j and so by Collett (2015) these residuals are known as score residuals. From equation 4.42, the i^{th} score residual, $i = 1, \cdots, n$ for the j^{th} explanatory variable in the model, X_j, is given by

$$r_{S_{ji}} = \delta_i(x_{ji} - a_{ji}) + exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i) \sum_{t_r \leq t_i} \frac{(\hat{a}_{jr} - x_{ji})\delta_r}{\sum_{l \in R(t_r)} exp(\boldsymbol{\beta}' \boldsymbol{x}_l)} \tag{4.43}$$

Using equation 4.38, this may be written in the form

$$r_{S_{ji}} = r_{p_{ji}} + exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i) \sum_{t_r \leq t_i} \frac{(\hat{a}_{jr} - x_{ji})\delta_r}{\sum_{l \in R(t_r)} exp(\boldsymbol{\beta}' \boldsymbol{x}_l)} \tag{4.44}$$

Which according to Collett (2015) shows that the score Residuals are modifications of the Schoenfeld Residuals.

4.5.6 Strategies of dealing with non-proportionality

If nonproportional hazards are detected, then by Therneau & Grambsch (2000), the researcher may choose between the following options to address the violation;

- If changes in the coefficient over time appears very small or if it appears the outliers are driving the changes in the coefficient, one can ignore the non-proportionality. In large datasets, very small departures from proportional hazards can be detected.
- One can also stratify the model by the non-proportional covariate. The advantage of stratification is that it allows each stratum to have its own baseline

hazard, which ultimately solves the non-proportionality problem. The drawback of stratifying is that, one cannot test whether the stratifying variable itself affects the hazard rate significantly.

- Consider including covariate interactions with time as predictors in the Cox model. Thus a significant interaction indicates violation of the PH assumption.
- Lastly one can consider using alternative regression models such as accelerated failure time models and additive hazard regression models. In section 4.6 below we briefly describe additive hazard regression models.

4.6 Additive hazard regression model

The additive hazards model, also known as the additive Aalen model, originally suggested by Aalen (1980), is a simple and flexible nonparametric model with well understood properties. This model was proposed as an alternative method for modeling survival data, when the proportional hazard assumption for the Cox proportional hazards model is violated. However, this model has not gained much popularity in practice, according to Scheike (2006) perhaps this is because the model only contains nonparametric terms, and that the handling of these terms for inferential purposes has not been fully developed yet. Other reasons could be lack of familiarity with the models and lack of knowledge on how to implement the models using existing software Xie et al. (2013). For Aalens model the hazard rate at time t for an individual i with vector of covariates $\boldsymbol{x}_i(t) = (x_{i1}(t), \cdots, x_{ip}(t))'$ takes the form

$$\alpha(t|\boldsymbol{x}_i) = \beta_0(t) + \beta_1 x_{i1}(t) + \cdots + \beta_p x_{ip}(t). \tag{4.45}$$

where β_0 is the baseline hazard corresponding to the hazard rate of an individual with all covariates identically equal to zero, while $\beta_j(t)$ the coefficients in the above model can be interpreted as the change in hazard at time t, corresponding to a unit increase in the j^{th} covariate. Note that $\beta_j(t)$ in 4.45 allow the effects of the covariates to change over time.

According to Xie et al. (2013), unlike the proportional hazards model which estimates hazard ratios, an additive model estimates the difference in hazards: the change in hazard function due to the exposure of interest or stated more simply the absolute difference in the instantaneous failure rate per unit of change in the exposure variable. Further details about this model can be found in (Aalen, 1980; Scheike, 2006 and Hosmer et al., 2002)

4.6.1 summary

The Kaplan-Meier and log-rank formulation was discussed. The Cox proportional hazards model was examined, including residuals used for checking the PH assumption. Furthermore strategies for dealing with non proportionality were briefly described. The additive hazard regression model as an alternative to model survival data when the PH assumption fails was briefly described.

Chapter 5

Joint models for longitudinal and time-to-event data

5.1 Introduction

This chapter will describe how to jointly model time-to-event data with longitudinal data following an approach by Rizopoulos (2012).

5.2 Joint model formulation

According to Wu et al. (2012) joint models for longitudinal and time-to-event data are typically required in the following situations:

- Survival models with measurement errors in time-dependent covariates;
- Longitudinal models with informative dropouts;
- Longitudinal and survival processes governed by a common latent process and
- The use of external information for more efficient inference.

Furthermore joint models of longitudinal data and time-to-event data entail, improving inference for a time-to-event outcome, whilst taking account of an intermittently and error-prone measured endogenous time-dependent variable Wulfsohn & Tsiatis (1997) and studying the relationship between the two correlated processes Henderson et al. (2000).

Formulating a standard joint modelling framework, follows a typical setup where you have a linear mixed-effects (LME) model for the longitudinal data and a Cox proportional hazards (PH) model for the time-to-event data, with the two models sharing some random effects Wu et al. (2012).This is the so called shared parameter model approach

5.2.1 The survival sub-model

We shall follow the model formulation by Rizopoulos (2012), thus consider a longitudinal study with n individuals in the sample. The objective is to model the time to an event of interest or survival time. The main aim is to measure the association between the longitudinal marker level and the risk for an event, while accounting for the special features of the former Rizopoulos (2012). To achieve this let $m_i(t)$ denote the true and unobserved value of the longitudinal outcome at time t. Note that $m_i(t)$ is different from $y_i(t)$, where the latter is contaminated with the measurement error value of the longitudinal outcome at time t. A relative risk model is postulated to quantify the association between $m_i(t)$ and the risk of an event given by

$$\begin{aligned} h_i(t|\mathcal{M}_i(t), w_i) &= \lim_{\delta t \to 0} P_r\Big\{t \leq T_i^* < t + dt | T_i^* \geq t, \mathcal{M}_i(t), w_i\Big\} \\ &= h_0(t) \times exp\big\{\gamma^T w_i + \alpha m_i(t), \big\}, \qquad t > 0 \end{aligned} \tag{5.1}$$

where $\mathcal{M}_i(t) = \big\{m_i(s), 0 \leq s < t\big\}$ denotes the history of the true unobserved longitudinal process up to time point t, $h_0(\cdot)$ denotes the baseline risk function, and w_i is a vector of baseline covariates, with a corresponding vector of regression coefficients γ. α quantifies the effect of the underlying longitudinal outcome to the risk of an event (Rizopoulos, 2012).

Usually in standard survival analysis $h_0(\cdot)$ is left completely unspecified, however according to Hsieh et al. (2006), following this route in the joint modelling framework may lead to an underestimation of the standard errors of the parameter estimates. Several approaches in the literature have been proposed to flexibly model $h_0(\cdot)$. According to Rizopoulos (2012) typically used distributions for specification of the baseline risk function include the Weibull, the log-normal and the Gamma. Practically, two approaches that work quite satisfactory are the piecewise-constant models and regression splines;

1. Under the piecewise-constant model, the baseline risk function takes the form:

$$h_0(\cdot) = \sum_{q=1}^{Q} \xi_q I(v_{q-1} < t \leq v_q), \tag{5.2}$$

where $0 = v_0 < v_1 < \cdots < v_Q$ denotes a split of the time scale, with v_Q being larger than the largest observed time, and ξ_q denotes the value of the hazard in the interval $(v_{q-1}, v_q]$. According to Rizopoulos (2012), the specification of the baseline hazard become more flexible as the number of knots increases. In the limiting case where each interval v_{q-1}, v_q contains only a single true event time (assuming no ties), model 5.2 is equivalent to a standard survival analysis where $h_0(\cdot)$ is left completely unspecified and estimating it using nonparametric maximum likelihood.

2. For the regression splines model the log baseline risk function is expanded into B-spline basis functions for cubic splines with the following form:

$$logh_0(t) = k_0 + \sum_{d=1}^{m} k_d B_d(t, q), \tag{5.3}$$

where $k = (k_0, k_1, \cdots, k_m)^{'}$ are the spline coefficients, q denotes the degree of the B-splines basis function $B(\cdot)$, and $m = \ddot{m} + q - 1$, denoting the number of interior knots (Rizopoulos, 2012). Similar to the piecewise-constant model, increasing the number of knots increases the flexibility in approximating $h_0(\cdot)$, (Rizopoulos, 2012).

5.2.2 The longitudinal sub-model

According to Rizopoulos (2012), to measure the effect of the longitudinal covariate to the risk for an event, $m_i(t)$ needs to be estimated and successfully reconstruct the complete longitudinal history $\mathcal{M}_i(t)$ for each subject. In order for this to work a suitable mixed-effects model is postulated to describe the subject-specific time evolutions. Following Rizopoulos (2012) consider a linear-mixed effect model with normally distributed outcomes

$$\begin{cases} y_i(t) = m_i(t) + e_i(t) \\ m_i(t) = x_i^{'}(t)\beta + z_i^{'}(t)b_i \\ b_i \sim N(0, D), \quad e_i(t) \sim N(0, \sigma^2) \end{cases} \tag{5.4}$$

Where $x_i(t)$ and $z_i(t)$ are the design vectors for the fixed effects β and random effects b_i, respectively. Note that the design vectors $x_i(t)$ and $z_i(t)$ as well as the error terms $e_i(t)$ are time dependent. Furthermore, assume that the error terms are mutually independent, independent of the random effects and normally distributed with mean zero and variance σ^2 Rizopoulos (2012). According to Wu et al. (2012), in survival models, some time-dependent covariates may be measured with errors. Thus the mixed model plays an important role in accounting for that measurement error problem by postulating that the observed level of the longitudinal outcome $y_i(t)$ equals the true level $m_i(t)$ plus a random term. Moreover, the time structure in the definitions of $x_i(t)$ and $z_i(t)$, and the use of subject-specific random effects allows to reconstruct the complete path of the time-dependent process $\mathcal{M}_i(t)$ for each subject Rizopoulos (2012). The survival function $S_i(t)$ depends on the whole history of the true marker levels, therefore obtaining a good estimate of $\mathcal{M}_i(t)$ is crucial for an accurate estimation of $S_i(t)$ (Rizopoulos, 2012). To flexibly model the subject-specific longitudinal profiles, several authors have considered spline-based approaches in the joint models framework (Rizopoulos, 2012). Alternatively, highly nonlinear shapes of subject-specific evolutions can be modelled using the linear mixed model approach by incorporating an additional stochastic term that aims to capture remaining serial correlation in the observed measurements not captured by the random effects (Rizopoulos, 2012). Thus, the linear mixed model is then given by

$$y_i(t) = m_i(t) + u_i(t)e_i(t) \tag{5.5}$$

where $u_i(t)$ is a stochastic process with mean zero, independent of b_i and $e_i(t)$, and $m_i(t)$ has the same mixed-effects model structure as 5.4 (Rizopoulos, 2012).

5.3 Estimation of joint models

According to Wu et al. (2012), there are two commonly used approaches for inference of joint models:

1. Two-stage methods
2. Likelihood methods

5.3.1 Two-stage methods

- In the first stage, the linear mixed effects model is fitted to the longitudinal covariate data, that is, the covariate is modeled using growth curve models with random effects (Laird & Ware, 1982). At each event time, the individual random effects are estimated using empirical Bayes methodology (Wulfsohn & Tsiatis, 1997)
- In the second stage, the survival model is fitted separately, the modeled value is then substituted into the partial likelihood for the Cox model with time dependent covariates, and the partial likelihood is then maximized Tsiatis et al. (1995). This modeling approach has been advocated on the basis that it reduces the bias of the parameter estimate in the Cox model (Wulfsohn & Tsiatis, 1997).

According to Wu et al. (2012), this method is fairly simple and can be implemented with existing software. The limitation with the two-stage methods, however, is that they may lead to biased inferences

Following Wu et al. (2012), the bias in the estimation of the longitudinal model parameters caused by ignoring the informative truncations from the events may depend on the strength of the association between the longitudinal process and the survival process. The bias resulting from ignoring the estimation uncertainty in Stage 1 may depend on the magnitude of measurement errors in covariates. To address these issues, various modified two-stage methods have been proposed, leading to better two-stage methods.

5.3.2 Likelihood methods

Wulfsohn & Tsiatis, 1997; Henderson et al., 2000 and Hsieh et al., 2006 proposed a semi-parametric maximum likelihood method to estimate joint models (Rizopoulos, 2012). According to (Rizopoulos, 2012), the maximum likelihood estimates are derived as the modes of the log-likelihood function corresponding to the joint distribution of the observed outcomes T_i, δ_i, y_i. "Assume that the vector of time-independent random effects b_i underlies both the longitudinal and survival process. 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)". This means the distribution can be factored into a product of two components as

$$p(T_i, \delta_i, y_i | b_i, \theta) = p(T_i, \delta_i | b_i, \theta) p(y_i | b_i, \theta) \tag{5.6}$$

and

$$p(y_i | b_i, \theta) = \prod_j p(y_i(t_{ij}) | b_i, \theta), \tag{5.7}$$

with $\theta = (\theta_t^{'}, \theta_y^{'}, \theta_b^{'})^{'}$ denoting the full parameter vector, with θ_t denoting the parameters for the event time outcome, θ_y the parameters for the longitudinal outcomes and θ_b the unique parameters of the random-effects covariance matrix, and y_i is the $n_i \times 1$ vector of the longitudinal responses of the i^{th} subject. Furthermore, assume that given the event history, the censoring mechanism and the visiting process are independent of the true event times and future longitudinal measurements (Rizopoulos, 2012). Under these assumptions the log-likelihood contribution for the i^{th} subject takes the following form

$$\begin{aligned} logp(T_i, \delta_i, y_i; \theta) &= log \int p(T_i, \delta_i, y_i; \theta) db_i \\ &= log \int p(T_i, \delta_i | b_i; \theta_t, \beta) \Big[\prod_j p\{y_i(t_{ij}) | b_i \theta_y\} \Big] P(b_i; \theta_b) db_i, \end{aligned} \tag{5.8}$$

with the conditional density for the survival part $p(T_i, \delta_i | b_i; \theta_t, \beta))$ taking the form

$$\begin{aligned} p(T_i, \delta_i | b_i; \theta_t, \beta) &= h_i(T_i | \mathcal{M}_i(T_i); \theta_t, \beta)^{\delta_i} S_i((T_i | \mathcal{M}_i(T_i; \theta_t, \beta) \\ &= \Big[h_0(T_i) exp\{\gamma' w_i + \alpha m_i(T_i)\} \Big]^{\delta_i} \\ &\times exp\Big(- \int_0^{T_i} h_0(s) exp\{\gamma' w_i + \alpha m_i(s)\} ds \Big), \end{aligned} \tag{5.9}$$

where $h_0(\cdot)$ can be any positive function of time, such as the piecewise-constant model 5.2, or the B-spline model 5.3 or the hazard function of any known distribution, and the survival function given by 5.1. The joint density for the longitudinal responses together with the random effects is given by

$$\begin{aligned} p(y_i | b_i) p(b_i; \theta) &= \prod_j p(y_i(t_{ij}) | b_i; \theta_y p(b_i; \theta_b)) \\ &= (2\pi\sigma^2)^{-n_i/2} exp||y_i - X_i\beta - Z_i b_i||^2 / 2\sigma^2 \\ &\times (2\pi)^{-q_b/2} det(D)^{-1/2} exp(-b_i^{'} D^{-1} b_i / 2), \end{aligned} \tag{5.10}$$

where q_b denotes the dimensionality of the random-effects vector, and $||x||$ denotes the Euclidean vector norm (Rizopoulos, 2012). Maximization of the log-likelihood function $l(\theta) = \sum_i logp(T_i, \delta_i, y_i; \theta))$ with respect to θ can be achieved using standard algorithms, such as the Expectation-Maximization (EM; Dempster et al. (1977)) algorithm or the Newton-Raphson algorithm or any of its variants (Hunter & Lange, 2004). In the joint modelling literature the EM algorithm has been traditionally preferred (treating the random effects as ”missing data”), mainly due to the fact that in the M-step some of the parameters have closed-form updates Rizopoulos (2012). However, according to Rizopoulos (2012), a serious drawback of the EM algorithm is its linear convergence rate that results in slow convergence especially near the maximum.

5.4 Asymptotic inference for joint models

5.4.1 Hypothesis testing

In general, if one is interested in testing the null hypothesis

$$H_0 : \theta = \theta_0 \qquad versus \qquad H_a : \theta \neq \theta_0, \tag{5.11}$$

According to Rizopoulos (2012), we could use the standard asymptotic likelihood inference tests in Table 5.1 below

Table 5.1: Standard asymptotic likelihood inference tests

Test	Parameter
Likelihood ratio test	$LRT = -2\big\{l\big(\hat{\theta}_0\big) - l\big(\hat{\theta}\big)\big\}$
Score test	$U = S^T(\hat{\theta}_0)\big\{I(\hat{\theta}_0)\big\}^{-1}S(\hat{\theta}_0)$
Wald test	$W = (\hat{\theta} - \theta_0)^T I(\hat{\theta})(\hat{\theta} - \theta_0)$

where $\hat{\theta}_0$ and $\hat{\theta}$ denote the maximum likelihood estimates under the null and alternative hypothesis, respectively. $S(\cdot)$ and $I(\cdot)$ denote the score function and the observed information matrix of the model under the alternative hypothesis. Under the null hypothesis, the asymptotic distribution of each of these tests is based on a chi-squared distribution on p degrees of freedom, where p denotes the number of parameters being tested (Rizopoulos, 2012). For a single parameter θ_j the Wald test is equivalent to $(\hat{\theta}_j - \theta_{0j})/\widehat{s.e}(\hat{\theta}_j)$, which follows an asymptotic standard normal distribution. According to Rizopoulos (2012), "these test statistics are approximately low-order Taylor series expansion of each other, and they are asymptotically equivalent. However, in practice, when we are dealing with finite samples, they usually differ. In this case, the likelihood ratio test is generally considered the most reliable compared to the Wald test. The Score and the Wald test require fitting the model only under the null and alternative hypotheses, respectively, whereas the likelihood ratio test requires to fit the joint model under both hypothesis, and thus it is a bit more computationally expensive" (Rizopoulos, 2012). In the presence of missing data in the variable of interest, then the score test will be more efficient since it requires fitting the model under only the null hypothesis and therefore, avoids a case-wise deletion of missing values.

5.5 Estimation of the random effects

The random effects b_i in the joint model framework are used as a construct to describe the heterogeneity in the subject longitudinal evolutions and to build the association between the longitudinal and event time processes. To derive predictions for either outcome, an estimate of the random effects vector b_i is required. Since the random effects are assumed to be random variables, then according to Rizopoulos (2012), it is natural to estimate them using Bayesian paradigm. In particular, assuming that $p(b_i;\theta)$ is the prior distribution, and that $p(T_i,\delta_i|b_i;\theta)p(y_i|b_i;\theta)$ is the conditional likelihood part, the corresponding posterior distribution can be be derived as follows

$$\begin{aligned} p(b_i|T_i,\delta_i,y_i;\theta) &= \frac{p(T_i,\delta_i|b_i;\theta)p(y_i|b_i;\theta)p(b_i;\theta)}{p(T_i,\delta_i,y_i;\theta)} \\ &\propto p(T_i,\delta_i|b_i;\theta)p(y_i|b_i;\theta)p(b_i;\theta). \end{aligned} \tag{5.12}$$

In mixed models this 5.12 is a multivariate normal distribution, whereas in the joint model framework it does not have a closed-form solution and it has to be numerically computed (Rizopoulos, 2012). To describe this posterior distribution, standard summary measures are often utilized. For its location the mean or the mode are typically used, defined as

$$\begin{cases} \bar{b}_i = \int b_i \text{p}(b_i|T_i,\delta_i,y_i;\theta)db_i, \text{and} \\ \hat{b}_i = \text{arg max}_b\{\text{log p}(b|T_i,\delta_i,y_i;\theta)\}, \end{cases} \tag{5.13}$$

respectively, and as a measure of dispersion we may use the posterior variance or the inverse Hessian matrix of the random effects, i.e.,

$$\begin{cases} var(b_i) = \int (b_i - \bar{b}_i)^2 p(b_i|T_i,\delta_i,y_i;\theta)db_i, and \\ H_i = \{-\frac{\partial^2 logp(b|T_i,\delta_i,y_i;\theta)}{\partial b^T \partial b}|_{b=\hat{b}_i}\}^{-1}. \end{cases} \tag{5.14}$$

For the estimation of 5.13 and 5.14, an empirical Bayes approach is employed where θ is replaced by $\hat{\theta}$ (Rizopoulos, 2012).

5.6 Advantages of joint models

Often longitudinally measured data and time-to-event or survival data are associated in some ways. For example, in this study patients infected with HIV were monitored until they developed AIDS or died, and markers such as the CD4 lymphocyte count and CD8 lymphocyte count or the estimated viral load were regularly measured. Thus, the association between time to event and the longitudinal trajectories. Separate analyses of longitudinal data and survival data are not applicable in this case because they may lead to inefficient or biased results. Joint models, on the other hand, provide valid and efficient inferences by optimally incorporating all available information (longitudinal and survival data) simultaneously (Wulfsohn & Tsiatis, 1997). Several other advantages of jointly modelling longitudinal and survival data taken from (Faucett & Thomas, 1996) are listed below

- Jointly modelling longitudinal (covariate) and survival data reduces bias in parameter estimates due to measurement error and informative censoring
- The model allows for unequally spaced measurements, or missing covariate data and censoring of survival times
- ”In a survival analysis setting, where the covariate of interest is time-dependent, either the entire history of the covariate for every subject, or, minimally, measurements of the covariate at each time of disease occurrence for all subjects in the corresponding risk set, are necessary. This extensive measurement of covariate is rarely, if ever, executed and the values obtained are typically subject to measurement error. Thus by modelling the covariates over time, we can enhance the survival analysis since we can interpolate covariate values between the observed measurements to the specific times of disease occurrence, with the use of the entire covariate history of the subjects”.
- Furthermore, according to Faucett & Thomas (1996), after accounting for measurement error, the standard error of the relative risk estimate will reflect correctly the uncertainty in the measurements of the covariate. Conversely, utilizing the survival data in the covariate tracking (longitudinal) model will yield improved covariate tracking parameter estimates by allowing adjustment for informative right censoring of the repeated measurements by the disease process.
- The joint model has the distinct advantage of simultaneously modelling two response variables (for example in this study, CD4+ count and time-to-death), this allows the researcher some degree of flexibility (Ramroop, 2010).

5.7 Joint model diagnostics

5.7.1 Residuals for the longitudinal part

In the standard linear mixed-effects model, two types of residuals are often used, namely subject-specific (conditional) residuals and the marginal (population averaged) residuals, see section 3.6.1 for more details. According to Rizopoulos (2012), conditional residuals predict the conditional errors $e_i(t)$, and can be used for checking the homoscedasticity and normality assumptions, whereas marginal residuals predict the marginal errors $y_i - X_i\beta = Z_i b_i + e_i$, and can be used to investigate misspecification of the mean structure $X_i\beta$ as well as to validate the assumptions for the within-subjects covariance structure V_i. These two residuals can also be used to check the assumptions of the longitudinal part of a joint model as well (Rizopoulos, 2012).

5.7.2 Residuals for the survival part

According to Rizopoulos (2012) martingale residuals are the standard type of residuals for the relative risk submodel of the joint model. These residuals are given by

$$\begin{aligned} r_i^{tm}(t) &= N_i(t) - \int_0^t R_i(s) h_i(s|\hat{\mathcal{M}}_i(s); \hat{\theta}) ds \\ &= N_i(t) - \int_0^t R_i(s) \hat{h}_0(s) exp\{\hat{\gamma}^T w_i + \hat{\alpha}\hat{m}_i(s)\} ds \end{aligned} \tag{5.15}$$

where "$N_i(t)$ is the counting process denoting the number of events for subject i by time t, $R_i(t)$ is the left continuous at risk process with $R_i(t) = 1$ if subject i is at risk at time t, and $R_i(t) = 0$ otherwise. $\hat{m}_i(t) = x_i^T(t)\hat{\beta} + z_i^T(t)\hat{b}_i$ and $\hat{h}_0(\cdot)$ denotes the estimated baseline risk function (Rizopoulos, 2012). $r_i^{tm}(t)$ can be seen as the difference between the observed number of events and predicted or expected number of events by the same time based on the fitted model". According to Rizopoulos (2012), martingale residuals are mainly used for a direct identification of excess events and for evaluating whether the appropriate functional form for a covariate of interest has been used in the model. Alternatively, one can use Cox-Snell residuals which are calculated as the value of the estimated cumulative risk function evaluated at the observed event time T_i, that is

$$\begin{aligned} r_i^{tcs} &= \int_0^{T_i} (h_i(s|\hat{\mathcal{M}}_i(s); \hat{\theta}) ds \\ &= \int_0^{T_i} \hat{h}_0(s) exp(\hat{\gamma}^T w_i + \hat{\alpha}\hat{m}_i(s)) ds, \end{aligned} \tag{5.16}$$

and thus, $r_i^{tcs} = N_i(T_i) - r_i^{tm}(T_i)$. "According to the probability integral transform, when the assumed model fits the data well we expect that the probability of survival time t, i.e., $S(t) = Pr(T_i^* > t)$ will have a uniform distribution in [0.1], and therefore the cumulative hazard, defined as $\mathcal{H}(t) = -logS(t)$ will have a unit exponential distribution. This identity implies that we can check the overall goodness of fit of our relative risk sub-model by checking whether the Cox-Snell residuals r_i^{tcs} are unit exponentially distributed. However, according to Rizopoulos (2012), a complexity in the practical use of these residuals is that they are evaluated at the observed event time T_i, and thus when T_i is censored, r_i^{tcs} will be censored as well. Hence in order to check the fit of the model, while accounting for the fact that r_i^{tcs} is actually a censored sample from a unit exponential distribution, we compare the survival function of the unit exponential distribution, $E_{exp}(t) = exp(-t)$, with the Kaplan-Meier estimate of the survival function of r_i^{tcs}" (Rizopoulos, 2012).

5.7.3 Summary

The joint model methodology by Rizopoulos (2012) was examined. The advantages of joint models over separate analyses were discussed

Chapter 6

Application

6.1 Modelling square root CD4 count

In this chapter various linear mixed models will be fitted to determine the effect of different variables on CD4 count (square root transformed). The focus will be based on describing the mean model for square root CD4 count while also trying to capture the best correlation structure of the repeated measurements within a subject. This is because misspecification of the covariance structure for repeated measures in longitudinal analysis may lead to biased estimates of the model parameters. Given a set of candidate covariance structures for the data, the preferred structure is the one with the minimum Akaike information criterion (AIC) value. Generally, to develop the best fitting model for the data is not an easy task. Model selection is one of the major challenges faced in data analysis, the likelihood ratio test will be used as a selection criteria when comparing two nested models. Model building will be done using the stepwise procedure in SAS, version 9.4 (SAS Institute INC., Cary) and R version 3.5.1.

6.2 Univariable models

All the models for CD4 count outcome were fitted using the MIXED procedure in SAS software. The advantage of using linear mixed models is that it uses all available data including incomplete case. This method yields a consistent estimator of precision, even if the covariance is misspecified (Verbeke & Molenberghs, 2000).

6.2.1 Random effects model

Various covariance structures were used to model the within and between variation. Maximum likelihood (ML) and restricted maximum likelihood method (REML) are the estimation methods to be used. Independence structure, autoregressive and Toeplitz covariance structures were inappropriate for our data because: (1) Independence structure assumes that repeated measures are uncorrelated which is unrealistic with a longitudinal study. (2) Toeplitz structure assumes that correlations between equally distant points is constant which most likely not the case with our study given we had fairly long individual sequence of observations. (3) Autoregressive structure assumes that the measurements are equally spaced which is not the case with our study because of unplanned visits and missing values which exacerbated the coarseness of the data. We also tested the unstructured covariance structure as it is the most flexible, however it did not converge. We compared the AIC's of compound symmetry, spatial spherical, spatial power, spatial exponential, spatial Gaussian, and spatial linear covariance structures. We used AIC instead of the likelihood ratio test because the chosen structures were not nested within each other and also the likelihood ratio test would not be valid under REML.

$$Y_{ij} = \beta_0 + b_{0i} + (\beta_1 + b_{1i})t_{ij} + e_{ij} \tag{6.1}$$

Table 6.1: Covariance Parameter Estimates for the univariable random effects model

Effect	Estimate	Standard Error	Z Value	Pr > Z
UN(1,1)	13.2125	1.3727	9.62	$< .0001$
UN(2,1)	-3.4465	0.3361	-10.25	$< .0001$
UN(2,2)	2.5496	0.2591	9.84	$< .0001$
Variance	11.1209	1.4933	7.45	$< .0001$
SP(SPH)	31.8709	4.0896	7.79	$< .0001$
Residual	2.4560	0.1777	13.82	$< .0001$

Table 6.2: Solution for fixed effects for the univariable random effects model

Effect	Estimate	Standard Error	DF	t Value	Pr > $\|t\|$
Intercept	13.5911	0.1203	1645	113.01	$< .0001$
times_years	2.6385	0.06253	1488	42.20	$< .0001$

Table 6.1 shows that the random intercepts and slopes are significantly different from zero. The covariance between the random intercept and slope is given by UN (2,1), which is negative and statistically significant. The AIC for this model is 41972.9 which is smaller than 42484.8 the AIC for the marginal model, implying that the random effects model fits our data well. The results in Table 6.2 give us an average intercept and slope over time where β_0 =13.5911 is the average square root CD4

count at baseline and this is an estimate of the population mean. β_1 =2.6385 is an estimate for the population rate of increase of square root CD4 count. The intercept and slope for each patient is given by $(\beta_0 + b_{0i})$ and $(\beta_1 + b_{1i})$ respectively. The next model is where the relationship between CD4 count and gender will be explored. The model is given by:

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1 G_i + \beta_2 t_{ij} + \beta_3 G_i t_{ij} + b_{1i} t_{ij} + e_{ij} \tag{6.2}$$

where

$$\begin{cases} G_i = 1 & \text{if male} \\ G_i = 0 & \text{if female} \end{cases} \tag{6.3}$$

Table 6.3: Covariance Parameter Estimates for gender under the univariable random effects model

Effect	Estimate	Standard Error	Z Value	Pr > Z
UN(1,1)	12.1469	1.3546	8.97	$< .0001$
UN(2,1)	-3.3361	0.3306	-10.09	$< .0001$
UN(2,2)	2.4857	0.2576	9.65	$< .0001$
Variance	11.3363	1.4906	7.61	$< .0001$
SP(SPH)	31.9933	3.9368	8.13	$< .0001$
Residual	2.4205	0.1780	13.60	$< .0001$

Table 6.4: Solution for fixed effects under the univariable random effects model

Effect		Estimate	Standard Error	DF	t Value	Pr > $\|t\|$
Intercept		14.3020	0.1477	1644	96.83	$< .0001$
times_years		2.5582	0.07564	1487	33.82	$< .0001$
Gender	Men	-2.0794	0.2438	4698	-8.53	$< .0001$
times_years*gender	Men	0.1525	0.1323	4698	1.15	0.2491

In this model, the mean rate of change for males and females is given by β_2 and $(\beta_2+\beta_3)$ respectively. The results for this model are given by Table 6.3 and Table 6.4. Table 6.4 show that male and female patients have mean square root CD4 count of 12.2226 (14.3020-2.0794) and 14.3020 respectively. They also, on average gain square root CD4 count at the rate of 2.7107 (2.5582+0.1525) and 2.5582 respectively. The intercepts are statistically significantly different with females on average having a higher mean rate of change in square root CD4 count than males. The results are in line with exploratory data analysis in chapter 2, Figure 2.6b where we showed that females had higher mean CD4 count over time. The next sub-model fitted assesses whether the mean rate of change is the same in the two sites and this the model is given by;

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1 S_i + \beta_2 t_{ij} + \beta_3 S_i t_{ij} + b_{1i} t_{ij} + e_{ij} \tag{6.4}$$

$$\begin{cases} S_i = 1 & \text{If EThekwini site} \\ S_i = 0 & \text{If Vulindlela site} \end{cases} \tag{6.5}$$

Table 6.5: Covariance Parameter Estimates for site under the univariable random effects model

Effect	Estimate	Standard Error	Z Value	Pr > Z
UN(1,1)	12.9013	1.1582	11.14	$< .0001$
UN(2,1)	-3.3159	0.3284	-10.10	$< .0001$
UN(2,2)	2.4653	0.2438	10.11	$< .0001$
Variance	11.2772	1.1762	9.59	$< .0001$
SP(SPH)	32.4517	2.8763	11.28	$< .0001$
Residual	2.4612	0.1764	13.95	$< .0001$

Table 6.6: Solution for fixed effects for site under the univariable random effects model

Effect		Estimate	Standard Error	DF	t Value	Pr > $\|t\|$
Intercept		13.8685	0.1393	1643	99.58	$< .0001$
times_years		2.4638	0.06870	1488	35.86	$< .0001$
Site	EThekwini	-1.0591	0.2740	4698	-3.87	0.0001
times_years*site	EThekwini	0.7363	0.1596	4698	4.61	$< .0001$

The results in Table 6.6 show that the interaction term times_years*site is significant, implying that the inclusion of random intercept and slope in the model allows the rate of change from both sites to differ significantly at 5% level of significance. Patients from the EThekwini and Vulindlela site have mean square root CD4 count of 12.8094 (13.8685-1.0591) and 13.8685 respectively. They also, on average gain square root CD4 count at the rate of 3.2001 (2.4638+0.7363) and 2.4638 respectively. The intercepts and slopes are statistically significantly different with patients from the EThekwini site on average having a higher mean rate of change in square root CD4 count than those from the Vulindlela site. The next sub-model fitted assesses whether the mean rate of change is the same for patients without TB and those with prevalent TB and the model is given by:

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1 B_i + \beta_2 t_{ij} + \beta_3 B_i t_{ij} + b_{1i} t_{ij} + e_{ij} \tag{6.6}$$

$$\begin{cases} B_i = 1 & \text{if prevalent TB} \\ B_i = 0 & \text{if no TB} \end{cases} \tag{6.7}$$

Table 6.7: Covariance Parameter Estimates for TB status under the univariable random effects model

Effect	Estimate	Standard Error	Z Value	Pr > Z
UN(1,1)	13.1707	1.3945	9.44	$< .0001$
UN(2,1)	-3.2084	0.3261	-9.84	$< .0001$
UN(2,2)	2.3894	0.2463	9.70	$< .0001$
Variance	10.7989	1.5044	7.18	$< .0001$
SP(SPH)	31.4333	4.3450	7.23	$< .0001$
Residual	2.4772	0.1751	14.14	$< .0001$

Table 6.8: Solution for fixed effects for TB status under the univariable random effects model

Effect		Estimate	Standard Error	DF	t Value	Pr > $\|t\|$
Intercept		12.5455	0.2573	1644	48.76	$< .0001$
times_years		3.6831	0.1593	1488	23.12	$< .0001$
TB status	No TB	1.3420	0.2904	4697	4.62	$< .0001$
times_years*TB status	No TB	-1.3007	0.1724	4697	-7.54	$< .0001$

Table 6.8 shows patients with no TB and those with prevalent TB have mean square root CD4 count of 13.8875 (12.5455+1.3420) and 12.5455 respectively. They also, on average gain square root CD4 count at the rate of 2.3824 (3.6831-1.3007) and 3.6831 respectively. The intercepts and slopes are statistically significantly different with patients with prevalent TB on average having a higher mean rate of change in square root CD4 count than those without TB. The next set of models will explore the relationship between square root CD4 count and the four continuous variables namely

age, BMI, baseline CD8 count and baseline viral load separately.

Table 6.9: Covariance Parameter Estimates for age under the univariable random effects model

Effect	Estimate	Standard Error	Z Value	Pr > Z
UN(1,1)	13.3158	1.3289	10.02	$< .0001$
UN(2,1)	-3.4821	0.3372	-10.33	$< .0001$
UN(2,2)	2.5675	0.2586	9.93	$< .0001$
Variance	10.9673	1.4248	7.70	$< .0001$
SP(SPH)	31.5365	3.9099	8.07	$< .0001$
Residual	2.4643	0.1773	13.90	$< .0001$

Table 6.10: Solution for fixed effects for age under the univariable random effects model

Effect	Estimate	Standard Error	DF	t Value	Pr > $\|t\|$
Intercept	14.0840	0.4731	1643	29.77	$< .0001$
times_years	3.0727	0.2592	1488	11.86	$< .0001$
Age	-0.01443	0.01315	4698	-1.10	0.2724
times_years*age	-0.01243	0.007121	4698	-1.75	0.0809

Looking at Table 6.10, the estimate for age is not statistically significantly different from zero (P-value = 0.2724). This implies that the regression of square root CD4 count on age is not statistically significant given time (times_years) and times_years*age are in the model. This means that younger and older patients started HAART with almost the same CD4 count. However, the interaction times_years*age is negative and also not statistically significant (P-value = 0.0809). Continous variables re frequently modified into categorical variables because interpreting their interaction term is very tricky (Van Walraven & Hart, 2008).

Table 6.11: Covariance Parameter Estimates for BMI under the univariable random effects model

Effect	Estimate	Standard Error	Z Value	Pr > Z
UN(1,1)	12.8752	1.3663	9.42	$< .0001$
UN(2,1)	-3.3638	0.3331	-10.10	$< .0001$
UN(2,2)	2.5227	0.2583	9.77	$< .0001$
Variance	11.1887	1.4885	7.52	$< .0001$
SP(SPH)	31.9776	4.0305	7.93	$< .0001$
Residual	2.4482	0.1774	13.80	$< .0001$

Table 6.12: Solution for fixed effects for BMI under the univariable random effects model

Effect	Estimate	Standard Error	DF	t Value	Pr > $\|t\|$
Intercept	12.0300	0.8347	1642	14.41	$< .0001$
times_years	3.0585	0.2971	1488	10.29	$< .0001$
BMI	0.06529	0.03481	4699	1.88	0.0607
times_years*bmi	-0.01767	0.01208	4699	-1.46	0.1435

The parameter estimate of the effect of BMI on square root CD4 count is 0.06529 and it is not statistically significant (P-value = 0.0607). Implying that for every unit increase in BMI the mean square root CD4 count increases by 0.06529 units, while setting all the other explanatory variables constant .This is an indication of a positive correlation between mean square root CD4 count and the BMI of that individual at that time.

Table 6.13: Covariance Parameter Estimates for square root CD8 count under the univariable random effects model

Effect	Estimate	Standard Error	Z Value	Pr > Z
UN(1,1)	9.2726	1.2502	7.42	$< .0001$
UN(2,1)	-2.1488	0.2846	-7.55	$< .0001$
UN(2,2)	2.0961	0.2232	9.39	$< .0001$
Variance	11.5492	1.3432	8.60	$< .0001$
SP(SPH)	33.5600	3.6296	9.25	$< .0001$
Residual	2.4713	0.1725	14.32	$< .0001$

Table 6.14: Solution for fixed effects for square root CD8 count under the univariable random effects model

Effect	Estimate	Standard Error	DF	t Value	Pr > $\|t\|$
Intercept	7.5405	0.3967	1643	19.01	$< .0001$
times_years	4.8707	0.2064	1488	23.60	$< .0001$
sqrtcd8	0.2096	0.01321	4698	15.86	$< .0001$
times_years*sqrtcd8	-0.07796	0.006272	4698	-12.43	$< .0001$

For every unit increase in baseline CD8 count, the square root CD4 count increases by 0.2096 units subject to other effects held constant in the model. The interaction term times_years*sqrtcd8 is statistically significant (P< .0001) meaning that the rate of change is different for everyone for different baseline square root CD8 values 6.14.

Table 6.15: Covariance Parameter Estimates for $\log_{10}$ under the univariable random effects model

Effect	Estimate	Standard Error	Z Value	Pr > Z
UN(1,1)	12.6457	1.1254	11.24	$< .0001$
UN(2,1)	-3.2545	0.3266	-9.96	$< .0001$
UN(2,2)	2.4860	0.2432	10.22	$< .0001$
Variance	11.1154	1.1248	9.88	$< .0001$
SP(SPH)	32.1992	2.7797	11.58	$< .0001$
Residual	2.4867	0.1754	14.18	$< .0001$

Table 6.16: Solution for fixed effects for $\log_{10}$ under the univariable random effects model

Effect	Estimate	Standard Error	DF	t Value	Pr > $\|t\|$
Intercept	17.4233	0.6518	1642	26.73	$< .0001$
times_years	1.1834	0.3516	1488	3.37	0.0008
Logviral	-0.7717	0.1282	4699	-6.02	$< .0001$
times_years*logviral	0.2921	0.07014	4699	4.16	$< .0001$

Table 6.16 shows that for every unit increase in baseline $\log_{10}$ viral load the square root CD4 count decreases by 0.7717 units, subject to other effects held constant in the model. Implying that there is a negative correlation between CD4 count and viral

load. The interaction term times_years*logviral is statistically significant (P$<$ .0001) meaning that the rate of change is different for everyone for different baseline log viral load values.

6.2.2 Marginal model

We fitted a marginal model investigating the effect of time on the square root CD4 count. This model is given by

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + e_{ij} \tag{6.8}$$

where Y_{ij} is the j^{th} square root CD4 count measurement for the i^{th} subject for $i = 1, 2, \cdots, N$ and $j = 1, 2, \cdots, n_i$. β_0 is the intercept, t_{ij} represents the years post infection at the i^{th} visit, while β_1 is the slope estimate for the change in square root CD4 count for every one visit increase. The e_{ij} is the random error associated with the j^{th} measurement for subject i. In this model individuals have the same mean response over time.

Table 6.17: Covariance Parameter Estimates for the univariable marginal model

Cov Parm	Estimate	Standard Error	Z Value	Pr > Z
Variance	26.5566	1.0857	24.46	$<$ .0001
SP(SPH)	54.2824	2.2200	24.45	$<$ .0001
Residual	2.2512	0.2618	8.60	$<$ .0001

Table 6.17 presents the covariance parameter estimates. These are estimates for random effects portion of the model. The variance component for patients is highly significant between patient variation and the residual variance is also significant at 5% level of significance

Table 6.18: Solution for fixed effects for the univariable marginal model

Effect	Estimate	Standard Error	DF	t Value	Pr > $\|t\|$
Intercept	14.6774	0.1490	1645	98.48	$<$ .0001
times_years	1.3610	0.08246	6186	16.51	$<$ .0001

The results in Table 6.18 present an average intercept and slope over time. $\beta_0 = 14.6774$ is the average intercept across patients and $\beta_1 = 1.3610$ is the average slope across all patients. In other words, the average square root CD4 count at baseline is 14.6774. Hence the average person with a square root CD4 count of 14.6774 gained square root CD4 count of 1.3610 per visit holding other variables fixed. This is in line with Figure 2.5a in Chapter 2. We compared the marginal model to the random effects model and we found that the AIC for this model is 42484.8 which is bigger than that of the random effects model (41972.9), implying that the random effects model is a better fit for our data.

6.3 Multivariable models

6.3.1 Marginal model

We fitted a marginal model to the square root CD4 count with all the variables in the model. Table 6.20 gives a summary of log Likelihood, AIC, AICC and BIC for the chosen structures using ML and REML as methods of estimation respectively. As illustrated in Table 6.20, spatial spherical structure is the best covariance structure for our analysis under both ML and REML methods of estimation with AIC= 41865.5 for ML and 41906.9 for REML. The model was then fitted under ML so that insignificant fixed effects will be deleted one at a time starting with the most insignificant one and the results are presented in table 7.1 and 7.2. The variables TB status, age, BMI and the interaction terms times_years*gender and BMI*times_years are statistically insignificant. We will start by removing the most insignificant term which is BMI*times_years, however TB status will not be removed from the model because it is involved in a significant interaction and age will not be removed as well because it is an important variable. According to Hallahan (2003), variables with subject matter importance should be kept in the model . After removing all the statistically insignificant terms, the AIC decreased from 41865.5 to 41864.4 and the likelihood ratio test significantly increased from 4852.65 to 4857.48 (P-value$<$ 0.0001) indicating a much better model fit. The final model was fitted using the REML algorithm and the results for this model are given in table 6.19, 6.21 and 6.22

Table 6.19: Covariance Parameter Estimates under the marginal multivariable model

Cov Parm	Estimate	Standard Error	Z Value	Pr > Z
Variance	22.4176	0.7150	31.35	< .0001
SP(SPH)	52.7733	1.9216	27.46	< .0001
Residual	2.4434	0.1659	14.72	< .0001

Table 6.20: Fit statistics for different covariance structures by ML and REML under the marginal multivariable model

Covariance structure	-2 Log Likelihood	AIC	BIC
Maximum Likelihood (ML)			
SP(SPH)	41827.5	**41865.5**	41968.2
SP(POW)	41899.2	41937.2	42039.9
SP(EXP)	41899.2	41937.2	42039.9
SP(GAU)	41876.8	41914.8	42017.5
SP(LIN)	41845.2	41883.2	41985.9
CS	43434.6	43470.6	43567.9
Restricted Maximum Likelihood (REML)			
SP(SPH)	41900.9	**41906.9**	41923.1
SP(POW)	41972.9	41978.9	41995.1
SP(EXP)	41972.9	41978.9	41995.1
SP(GAU)	41949.1	41955.1	41971.3
SP(LIN)	41919.3	41925.3	41941.5
CS	43512.4	43516.4	43527.2

Table 6.21: Solution for Fixed Effects under the marginal multivariable model

Effect		Estimate	Standard Error	DF	t Value	Pr > $\|t\|$
Intercept		12.4819	1.0194	1639	12.24	$< .0001$
Site	EThekwini	-0.9703	0.3086	1639	-3.14	0.0017
Gender	Men	-1.8233	0.2057	1639	-8.87	$< .0001$
TB status	No TB	0.6263	0.3463	1639	1.81	0.0707
Age		-0.00105	0.01358	1639	-0.08	0.9384
logviral		-0.7469	0.1402	1639	-5.33	$< .0001$
sqrtcd8		0.2049	0.01398	1639	14.66	$< .0001$
times-years		3.1199	0.5809	6181	5.37	$< .0001$
times-years*site	EThekwini	0.5955	0.1885	6181	3.15	0.0016
times-years*TB status	No TB	-0.7921	0.2503	6181	-3.16	0.0016
times-years*age		-0.01628	0.007607	6181	-2.14	0.0323
times-years*logviral		0.3048	0.08102	6181	3.76	0.0002
times-years*sqrtcd8		-0.06364	0.006724	6181	-9.46	$< .0001$

Table 6.22: Type 3 Tests of Fixed Effects under the marginal multivariable model

Effect	Num DF	Den DF	F Value	Pr > F
Site	1	1639	9.89	0.0017
Gender	1	1639	78.61	< .0001
TB status	1	1639	3.27	0.0707
Age	1	1639	0.01	0.9384
logviral	1	1639	28.39	< .0001
sqrtcd8	1	1639	214.80	< .0001
times_years	1	6181	32.24	< .0001
times_years*site	1	6181	9.98	0.0016
times_years*TB status	1	6181	10.01	0.0016
age1*times_years	1	6181	4.58	0.0323
logviral*times_years	1	6181	14.15	0.0002
sqrtcd8*times_years	1	6181	89.57	< .0001

6.3.2 Random effects model

Just like in the marginal model we compared the covariance structures for the random effects model and the unstructured covariance structure was the best fit for the random intercept and slope but the spatial spherical structure was found to be the best fit for the repeated measurements under both ML and REML with AIC=41532.2 for ML and 41569.4 for REML, the results are illustrated in Table 7.3.

Table 6.23: Covariance Parameter Estimates under the random effects multivariable model

Cov Parm	Estimate	Standard Error	Z Value	Pr > Z
UN(1,1)	8.2187	1.1685	7.03	< .0001
UN(2,1)	-1.8707	0.2704	-6.92	< .0001
UN(2,2)	1.9056	0.2074	9.19	< .0001
Variance	11.3421	1.2339	9.19	< .0001
SP(SPH)	33.4894	3.3765	9.92	< .0001
Residual	2.4975	0.1680	14.87	< .0001

The AIC for this model is 41569.4 which is smaller than 41906.9 that we used when selecting the spatial spherical covariance structure in table 6.20 under the marginal model. Just like in the marginal model, this model was also fitted under ML first and the results are illustrated in Table 7.4 and 7.5. The variables prev, age1, BMI and the interaction terms, times_years*site, times_years*gender, age*times_years and BMI*times_years are statistically insignificant. We will start by removing the most insignificant term which is BMI*times_years, however TB status will not be removed from the model because it is involved in a significant interaction and age will not be removed as well because it is an important variable. After removing all the statistically insignificant terms, the AIC decreased from 41532.2 to 41530.0 and the likelihood ratio test significantly increased from 5191.91 to 5206.26 (P-value$<$ 0.0001) indicating a much better model fit. The final model was fitted using the REML algorithm and the results for this model are given in table 6.24, 6.25 and 6.26

Table 6.24: Covariance Parameter Estimates under the final random effects multivariable model

Cov Parm	Estimate	Standard Error	Z Value	Pr > Z
UN(1,1)	12.9874	0.6773	19.18	$<$ .0001
UN(2,1)	-2.5179	0.2827	-8.91	$<$ .0001
UN(2,2)	2.5753	0.2076	12.40	$<$ .0001
Variance	6.1359	1.3262	18.81	$<$ .0001
SP(SPH)	1.7897	0.1099	16.29	$<$ .0001
Residual	2.9862	0.1504	19.85	$<$ .0001

Table 6.25: Solution for Fixed Effects under the final random effects multivariable model

Effect		Estimate	Standard Error	DF	t Value	Pr > $\|t\|$
Intercept		12.2626	0.8472	1637	14.47	$<$ .0001
Site	EThekwini	-0.2964	0.2369	4698	-1.25	0.2109
Gender	Men	-1.6942	0.2054	4698	-8.25	$<$ .0001
TB status	No TB	0.5314	0.2848	4698	1.87	0.0621
Age		-0.02673	0.01066	4698	-2.51	0.0122
logviral		-0.6009	0.1204	4698	-4.99	$<$ .0001
sqrtcd8		0.1903	0.01285	4698	14.80	$<$ .0001
times_years		4.8535	0.4411	1487	11.00	$<$ .0001
times_years*TB status	No TB	-1.1629	0.1697	4698	-6.85	$<$ .0001
times_years*logviral		0.2045	0.06903	4698	2.96	0.0031
times_years*sqrtcd8		-0.07192	0.006586	4698	-10.92	$<$ .0001

Looking at Table 6.25, β_0=12.2626 is the average random intercept across patients. In other words, the average square root CD4 count at baseline is 12.2626. β_1=4.8535 is the average random slope across all patients. Hence the average person with a square root CD4 count of 12.2626 gained square root CD4+ count of 4.8535 per visit. Hence for every year post HAART initiation, CD4 count on average increases significantly by 4.8535 square root cells (P-value$<$ .0001) , showing that over time CD4 count increases after HAART initiation, which is what is expected. Patients from the EThekwini site start with low mean square root CD4 count compared to those from Vulindlela site but this is not significant (P-value= 0.2109). Males and older people on average have a significantly lower mean rate of change in square root CD4 count. Patients presenting without TB at ART initiation started HAART with high mean CD4 count compared to those with prevalent TB but their rate of change in CD4 count is significantly less compared to those with prevalent TB (P-value$<$ .0001). For a unit increase in baseline log viral load, CD4 count significantly decreases by 39.91% (P-value$<$ .0001) subject to other effects held constant in the model. Implying that there is a negative correlation between CD4 count and viral load. For every unit increase in baseline square root CD8 count, the square root CD4 count increases by 0.1903 units subject to other effects held constant in the model.

Table 6.26: Type 3 Tests of Fixed Effects under the final random effects multivariable model

Effect	Num DF	Den DF	F Value	Pr > F
Site	1	4698	1.57	0.2109
Gender	1	4698	68.07	$<$ .0001
TB status	1	4698	3.48	0.0621
Age	1	4698	6.29	0.0122
logviral	1	4698	24.93	$<$ .0001
sqrtcd8	1	4698	219.17	$<$ .0001
times_years	1	4698	105.03	$<$ .0001
times_years*TB status	1	4698	46.97	$<$ .0001
logviral*times_years	1	4698	8.78	0.0031
sqrtcd8*times_years	1	4698	119.24	$<$ .0001

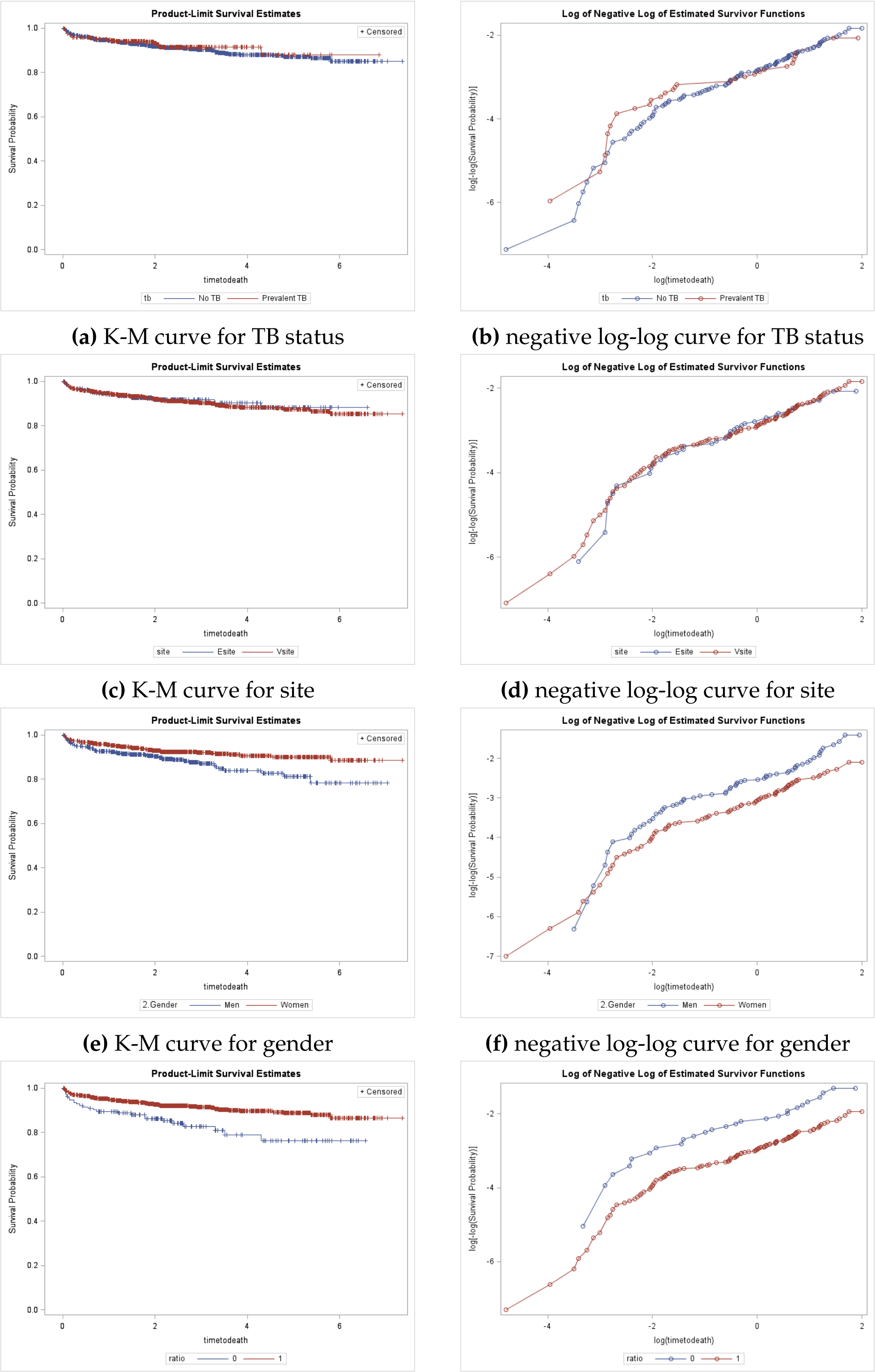
The AIC for this model is 41634.2 which is a considerable reduction compared to 41893.1, the AIC for the marginal model with the same covariance structure in section 6.3.1. Variance estimates for the random effects associated with the intercept and spatial spherical are statistically significant. The code for this model can be found in Appendix A

6.4 Survival analysis : modelling mortality

This section examines the survival of HIV patients co-infected with TB who were enrolled in the CAPRISA AIDS Treatment Project (CAT). First we test for the Cox PH assumption and those variables that violate the PH assumption will be excluded from the model. There after we will fit the Cox PH model in an attempt to model the relationship between mortality and the covariates described in Chapter 2.

The test for equality of strata (namely site, TB status and gender) in Table 7.6 both yield highly insignificant Chi-square test statistics for the log rank and Wilcoxon. The Wilcoxon test is a variation of the log rank test weighting the observed minus expected score of the jth failure time by n_j (the number still at risk at the jth failure time). Both the log rank and Wilcoxon test yield significant chi-square test statistics for gender.

We plotted the Kaplan-Meier curves and the negative log-log curves for the categorical variables, namely TB status, site, gender and ratio, in an attempt to assess for the Cox PH assumption. The Kaplan-Meier survival curves appear to steadily drift apart for gender and ratio but for TB status and site they seem to be crossing. The Log-log curves looked approximately parallel for gender, and ratio. Again, plots for TB status and site tended to cross more than once which makes it is difficult to conclude whether the assumption has been violated. Figure 6.1c indicates that patients from the EThekwini site have a lower probability of death when compared to those from the Vulindlela site. Figure 6.1e shows that men have an increased probability of death compared to women and those patients presenting without TB at ART initiation have a lower survival prognosis compared to those with prevalent TB. Figure 6.1h shows that patients with CD4:CD8 count ratio of < 0.05 have an elevated hazard of dying due to HIV compared to those with a CD4:CD8 count ratio ≥ 0.05



(a) K-M curve for TB status

(b) negative log-log curve for TB status

(c) K-M curve for site

(d) negative log-log curve for site

(e) K-M curve for gender

(f) negative log-log curve for gender

(g) K-M curve for ratio

(h) negative log-log curve for ratio

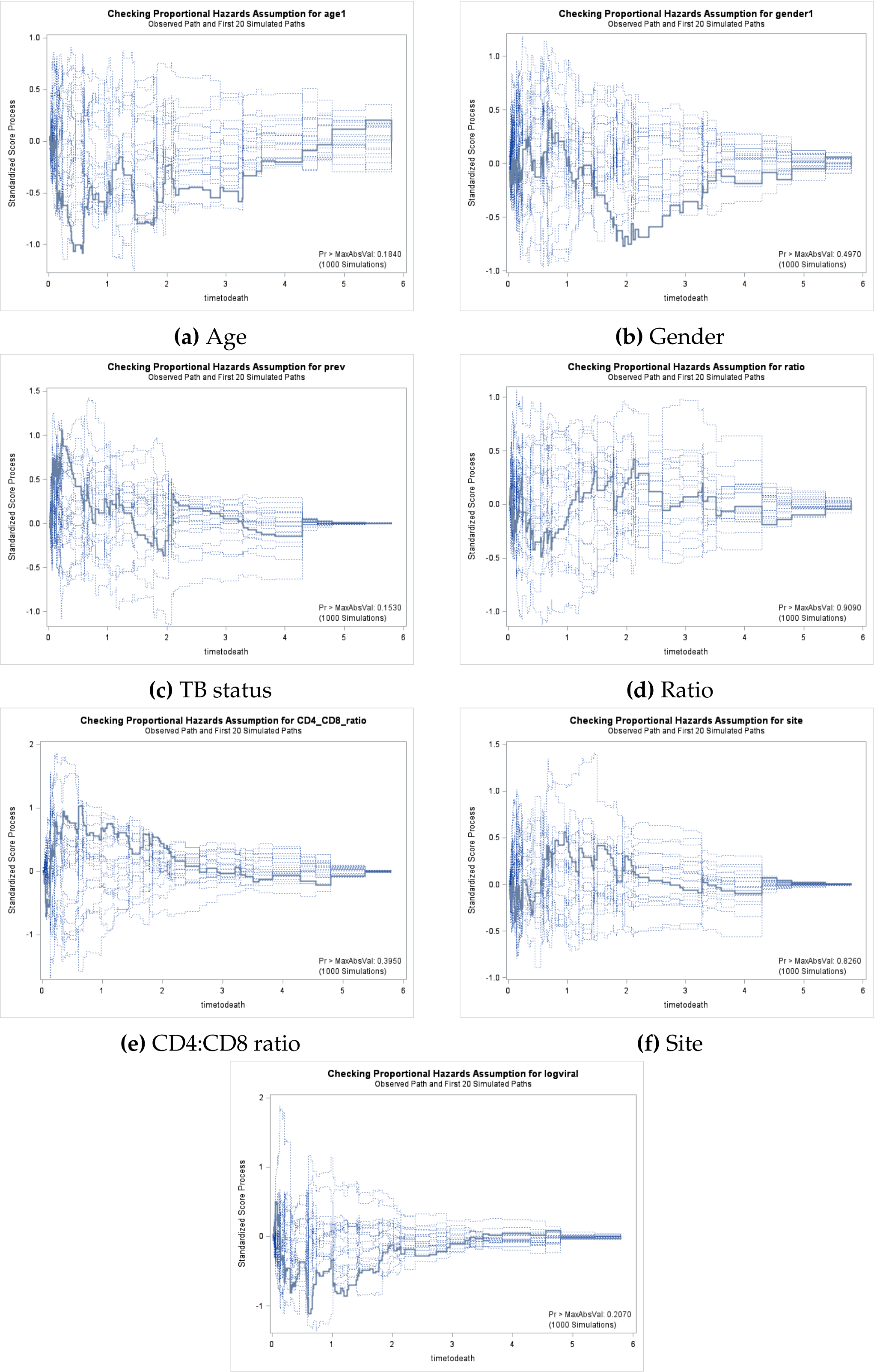
Figure 6.1 – Kaplan-Meier curves and negative log-log curves for TB status, site, gender and ratio

Since some of the curves in Figure 6.1 cross more than once it is difficult to conclude whether the assumption has been violated. The formal test ”coxzph” in R gives an estimate of the time-dependent coefficient $b(t)$ and tests its significance. If the proportional hazards assumption is true, $b(t)$ will be a constant and graphically a horizontal line. Thus a statistical test was applied to the current problem to test for time dependence of the regression coefficients, the result of which appears in Table 6.27.

Table 6.27: An investigation of the proportional hazards assumption

Variable	Effect	rho	chisq	P-value
Ratio	≥ 0.05	0.0117	0.0193	0.89
TB status	Prevalent TB	-0.0654	0.607	0.436
Gender	Men	0.0559	0.448	0.503
Site	Esite	-0.0178	0.0445	0.833
CD4:CD8 ratio		-0.0561	0.701	0.403
Age		0.122	2.28	0.131
logviral		0.0658	0.809	0.368
sqrtcd8		0.37	32.5	$< .0001$
sqrtcd4		0.282	10.4	0.00127
BMI		0.149	7.67	0.00562

Looking at Table 6.27 and p-values, there seems to be no evidence to suggest that the proportional hazards assumption is violated for variables ratio, TB status, gender, site, CD4:CD8 ratio, age and logviral. However there seems be some serious violation of the PH assumption for variables sqrtcd8, sqrtcd4 and BMI and these results are supported by the plots of the score process in figure 6.2. The cumulative sum of Schoenfeld residuals, or equivalently the observed score process can also be used to assess proportional hazards (Lin et al., 1993). Graphically, the solid lines represent the observed standardized score, while dotted lines represent 20 simulated sets of standardized score under the null hypothesis that PH assumption holds. A solid line that falls significantly outside the boundaries set up collectively by the dotted lines suggest that observed standardized score do not conform to the expected standardized score. These plots can then be used to assess when the lack of fit is present. In particular, an observed score well above the simulated process is an indication of an effect higher than the average one, and conversely. Variables that violate the PH assumption were not included in our model.



(a) Age

(b) Gender

(c) TB status

(d) Ratio

(e) CD4:CD8 ratio

(f) Site

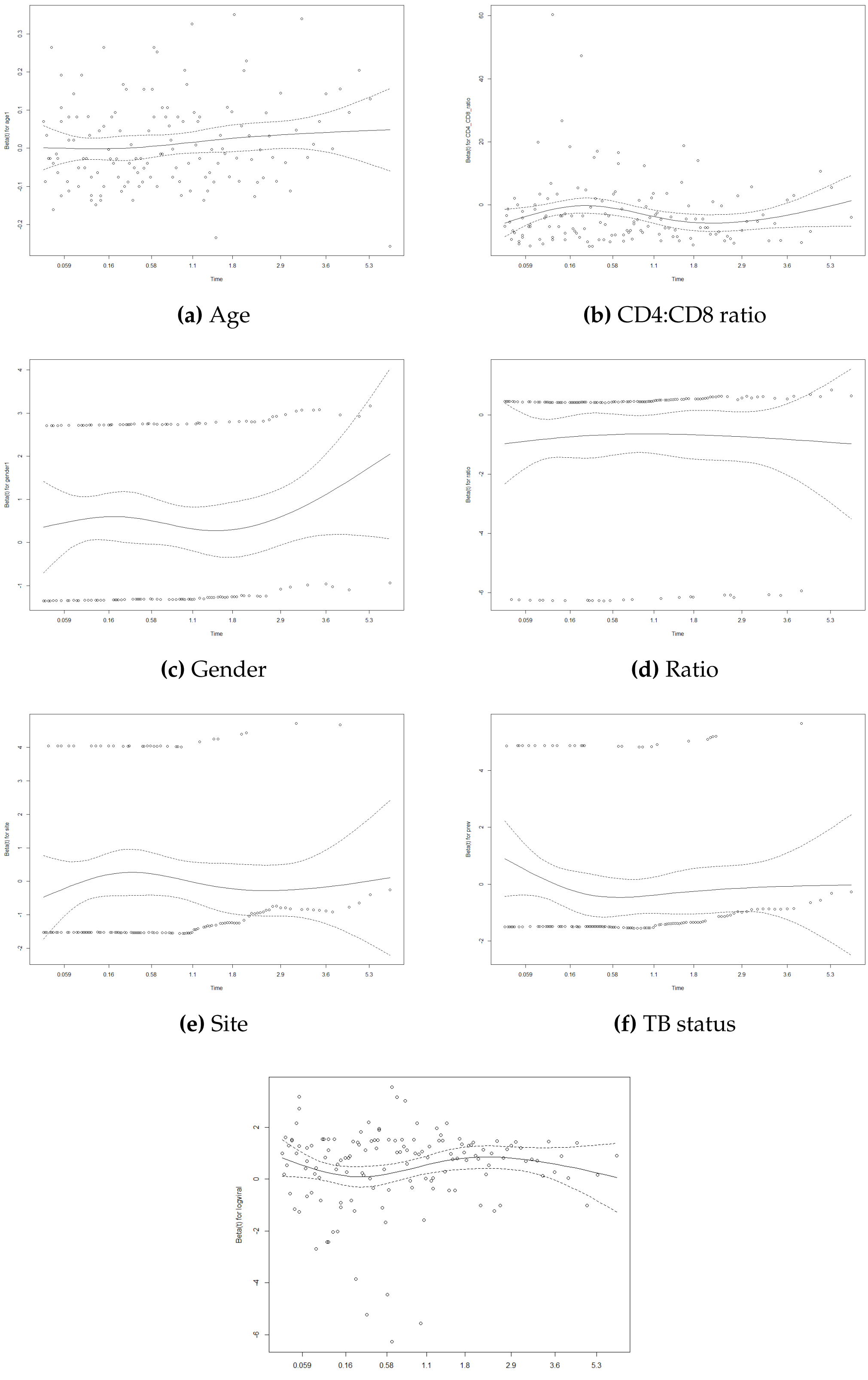
(g) Logviral

Figure 6.2 – Score process for the variables that satisfy the PH assumption

Table 6.28: The un-adjusted hazard ratios from fitting the Cox-proportional hazard model

	Effect	Estimate	HR (95% CI)	P-value
Un-adjusted hazards regression				
Site	EThekwini	-0.03	0.97 (0.66-1.43)	0.8850
Gender	Men	0.52	1.69 (1.21-2.35)	0.0020
Ratio	≥ 0.05	-0.68	0.50 (0.33-0.77)	0.0016
TB status	Prevalent TB	-0.12	0.89 (0.59-1.34)	0.5730
Logviral		0.45	1.57 (1.26-1.97)	< 0.0001
Age		0.01	1.01 (0.99-1.03)	0.1850
CD4:CD8 ratio		-2.91	0.05 (0.01-0.21)	< 0.0001
Adjusted hazards regression				
Site	EThekwini	0.02	1.02 (0.67-1.55)	0.9190
Gender	Men	0.42	1.53(1.09-2.14)	0.0132
TB status	Prevalent TB	-0.33	0.72 (0.46-1.13)	0.1566
Logviral		0.41	1.51 (1.20-1.90)	0.0004
Age		0.01	1.01 (0.99-1.03)	0.2939
CD4:CD8 ratio		-2.57	0.08 (0.02-0.30)	0.0002

Un-adjusted Cox proportional hazards regression analysis showed that patients from the EThekwini site (urban) had a 3% reduced risk of death compared to those from the Vulindlela site (rural), HR: 0.97, 95% CI: 0.66±1.43, P-value = 0.8850, meaning the site effect was not significant. Patients presenting with prevalent TB at ART initiation had a 12% reduced risk of death compared to those without TB, HR: 0.89, 95% CI: 0.59±1.34, P-value = 0.5730, meaning the TB status at baseline was not significant. (table 6.28). In addition male patients, older people or those with a higher mean $\log_{10}$ viral load had a significantly higher risk of death (table 6.28). Further exploration of Table 6.28 shows that for the adjusted proportional hazards regression analysis site, TB status and age have no significant effect in predicting death, but gender, Log_{10} viral load, and CD4:CD8 ratio has a statistically significant effect on predicting death. The code for the adjusted hazards model can be found in Appendix B



(a) Age

(b) CD4:CD8 ratio

(c) Gender

(d) Ratio

(e) Site

(f) TB status

(g) Logviral

Figure 6.3 – Time trend of the hazard ratio for the variables that satisfy the PH assumption

Figure 6.3 shows plots for Schoenfield residuals against time for all the variables that satisfy the PH assumption. The solid line is a smoothing-spline fit to the plot, with the broken lines representing standard error bands around the fit. This plot reaffirms the finding in Table 6.27, that the assumption of proportional hazards appears to be supported for the variables ratio, TB status, gender, site, CD4:CD8 ratio, age and logviral. Thus, we are confident that the proportional hazards model assumed in our analysis is justified.

6.5 Jointly modelling the CD4 count and mortality

In this section a joint model for CD4 count and mortality will be fitted using R version 3.5.1. This model will be fitted using **JM** package in R, where we first fit the linear mixed-effects and Cox models separately, and then supply the returned objects as main arguments in function ″jointmodel()″. The joint model fitted by ″jointmodel()″ has the exact same structure for the longitudinal and survival submodels as these two separately fitted models, with the addition that in the survival submodel the effect of the estimated ″true″ longitudinal outcome $m_i(t)$ is included in the linear predictor. The Cox model for the survival submodel needs to be fitted in the dataset containing only the survival information (that is, ″single row per patient″). The main argument ″timeVar″ of ″jointmodel()″ is used to specify the name of the time variable in the linear mixed-effects model, which is required in the internal computations of $m_i(t)$. The ″method″ argument specifies the type of baseline risk function, which in this case is assumed to be piece-wise constant, and the numerical integration approach. A detailed output of the fitted model is produced by the R function ″summary()″ that returns, among others, the parameter estimates, the standard errors, and asymptotic Wald tests for both the longitudinal and survival submodels. In the results for the event process, the parameter labeled ″Assoct″ is, in fact, parameter α in (5.1) that measures the association between $m_i(t)$ (that is, in our case of the ″true square root CD4 count) and the risk for death″.

Table 6.29: Joint model for longitudinal CD4 count and time to death

	Effect	Estimate	Standard Error	z-value	P-value
Longitudinal Process					
Intercept		12.6518	0.7700	16.4307	$< .0001$
times_years		4.1446	0.4222	9.8174	$< .0001$
Gender	Men	-1.6809	0.2061	-8.1544	$< .0001$
Logviral		-0.6082	0.1112	5.4687	$< .0001$
Age		-0.0263	0.0111	-2.3731	0.0176
sqrtcd8		0.1838	0.0116	15.8974	$< .0001$
times_years*prev	Prevalent TB	1.0438	0.1408	7.4151	$< .0001$
times_years*logviral		0.1554	0.0707	2.1971	0.0280
times_years*sqrtcd8		-0.0703	0.0070	-10.1147	$< .0001$
Event Process					
Gender	Men	0.0396	0.1767	0.2240	0.8227
Logviral		0.4074	0.1229	3.3146	0.0009
Age		0.0178	0.0100	1.7836	0.0745
CD4:CD8 ratio		1.2047	0.6405	1.8809	0.0600
Assoct		-0.3194	0.0309	-10.3314	$< .0001$

Table 6.30: Confidence intervals for the event process

Variable	Effect	Estimate ($\exp(-\alpha)$)	95% CI
Gender	Men	1.0404	(0.7359, 1.4708)
CD4_CD8_ratio		3.3356	(0.9506, 11.7042)
logviral		1.5029	(1.1812, 1.9123)
Age		1.0179	(0.9982, 1.0380)
Assoct		0.7266	(0.6839, 0.7720)

The joint model finds a significantly strong association between the square root CD4 count and the risk for death, with a unit decrease in the marker corresponding to a $\exp(-\alpha)$= 0.73 increase in the risk for death (95% CI: 0.68, 0.77). The parameter labelled ”Assoct” is the parameter that actually measures the association between CD4 count and the risk of dying due to HIV/AIDS. These results are statistically significant indicating that indeed CD4 count is a good predictor of mortality and in fact confirms that an increase in CD4 counts is associated with better survival. The code for the adjusted hazards model can be found in Appendix B

6.6 Joint Model Diagnostics

This section examines model diagnostics as they are a prerequisite step in validating model assumptions. Just like in the separate analysis, assumptions are assessed using residual plots. Model diagnostics for the joint models have not received much attention in the literature, with the only the exception being the conditional residuals of Dobson & Henderson (2003) and the multiple imputation residuals of Rizopoulos et al. (2010).

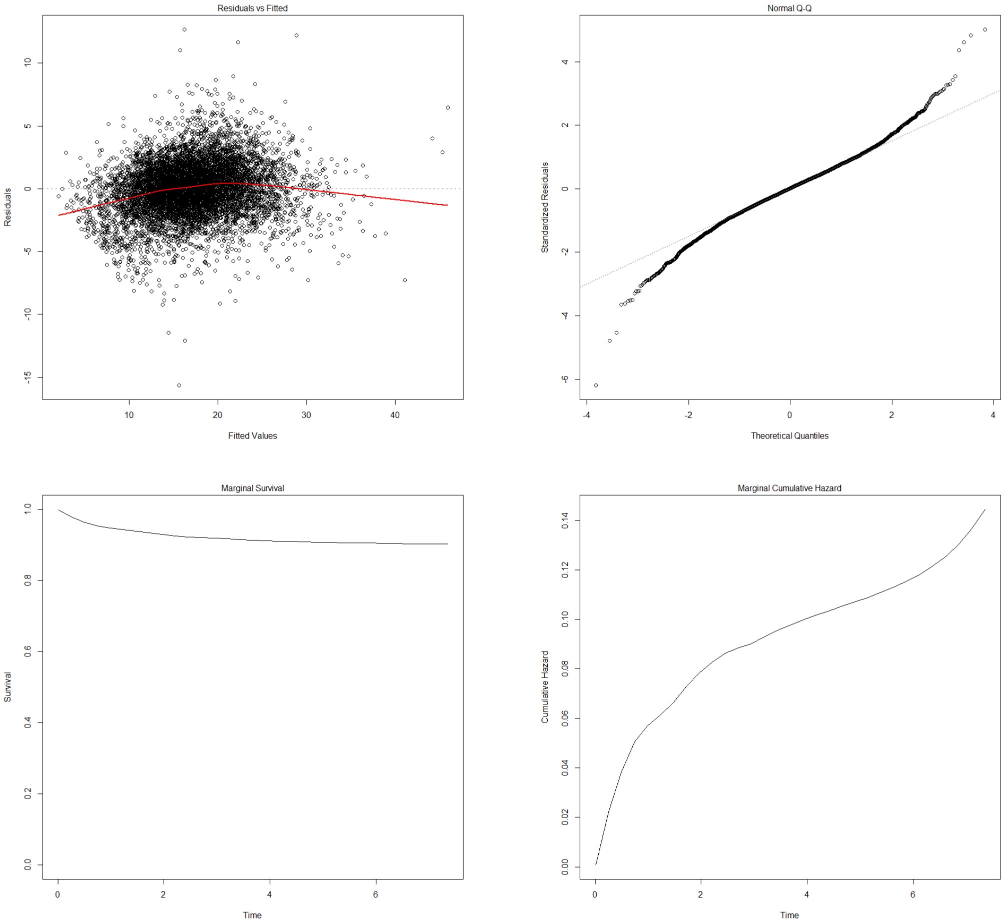


Figure 6.4 – Default diagnostic plots for the joint model fitted to the CAT dataset

Figure 6.4 presents the plots for the subject-specific residuals versus the corresponding fitted values, the Q-Q plot of the subject-specific residuals, and the marginal survival and cumulative risk functions for the event process. The residuals are scattered randomly around the 0 line indicating that the assumption that the relationship is linear is reasonable, also the residuals roughly form a horizontal line around 0 suggesting that the variances of the error terms are equal.The Q-Q plot is symmetric, with deviations from the Gaussian distribution occurring in both the left and right tails.

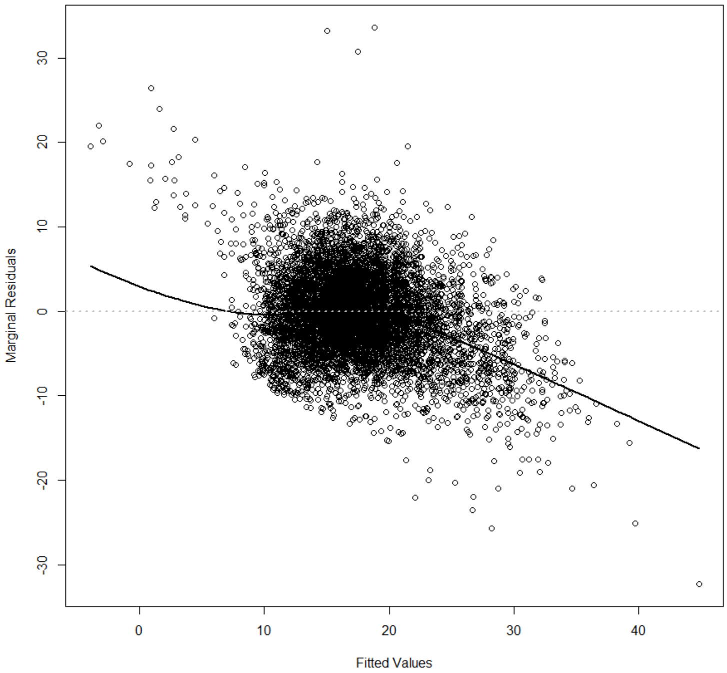


Figure 6.5 – Marginal standardized residuals versus fitted values for the longitudinal outcome for the CAT dataset

In Figure 6.5 we observe that the fitted loess curve in the plot of the standardized marginal residuals versus the fitted values shows a systematic trend with more negetive residuals for large fitted values. This maybe an indication that the form of the design matrix of the fixed effects X is not the appropriate one. However its important to note that low levels of CD4 count indicate a worsening of a patients condition resulting in higher death rates, which is why we cannot conclude solely from this figure that the lack-of-fit is attributed to a misspecification of X. Thus according to Rizopoulos (2012) who encountered the same problem for both the AIDS and PBC data explains that ”residuals based on the observed data alone can be proven misleading when it comes to validating the joint model's assumption”.

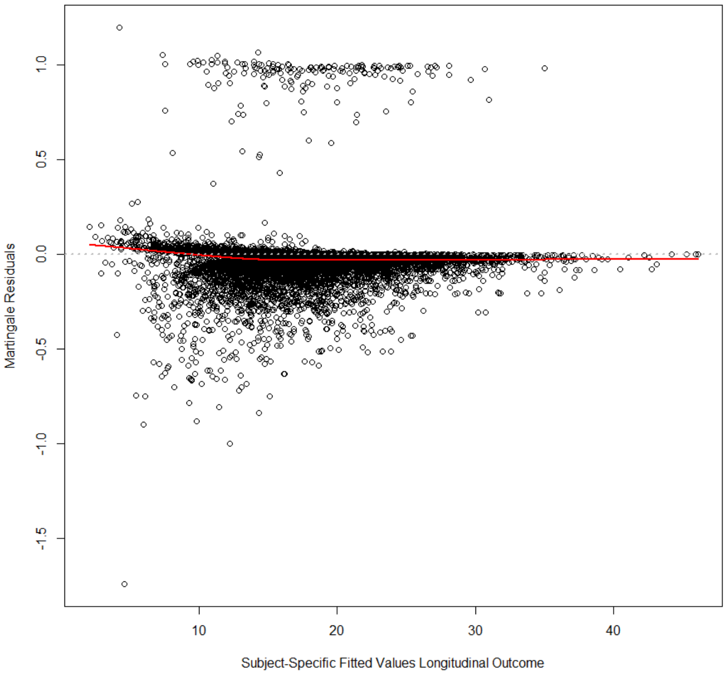


Figure 6.6 – Martingale residuals versus the subject-specific fitted values of the longitudinal outcome for the CAT dataset. The red solid line denotes the fit of the loess smoother

Figure 6.6 shows the scatter plot with a superimposed loess curve, we can observe that for small fitted values there's a slight deviation of the loess smoother from zero, this deviation is very small suggesting that the functional form for the CD4 count is appropriate, however it is advisable to additionally check for systemaic trend in the martingale residuals when we condition on other baseline covariates.

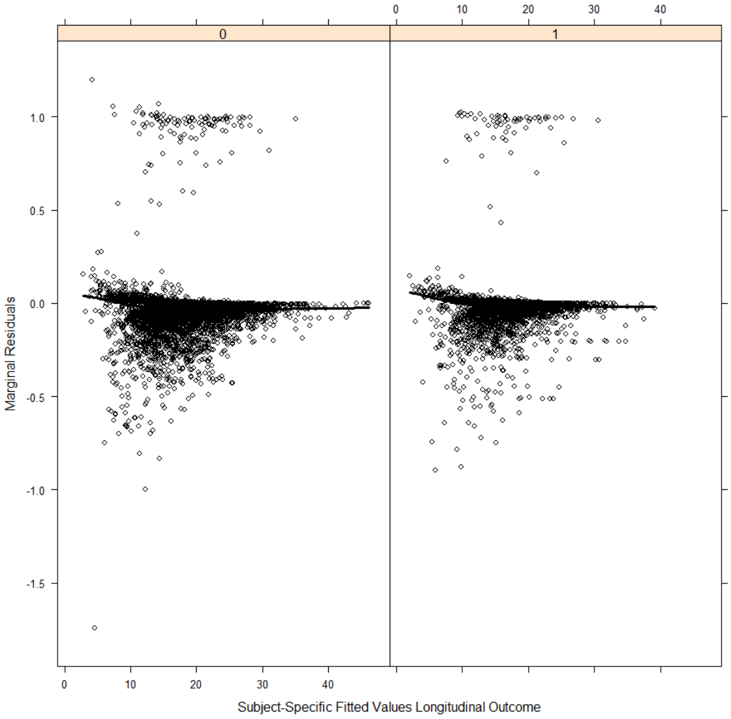


Figure 6.7 – Martingale residuals versus the subject-specific fitted values of the longitudinal outcome for the CAT dataset. The grey solid line denotes the fit of the loess smoother

Figure 6.7 again shows some small deviations from the null horizontal line for both sites. We proceed in our residual analysis for the survival outcome by assesing the overall fit of the survival submodel using the Cox-Snell residuals. Comparing the fit of the Kaplan-Meier estimate to the expected asymptotic distribution, we do not observe any discrepancies. As can be seen in Figure 6.8 the survival function of the unit exponential distribution lies within the 95% pointwise confidence intervals of the Kaplan-Meier estimate. To further scrutinize the fit of the model, we stratify the residuals by site, and we plot survival function estimates. ”When the model fits the data well, we expect the survival function estimates for each strata to hover around the unit exponential distribution” (Rizopoulos, 2012). Figure 6.9 shows no lack of fit for residuals from the two sites.

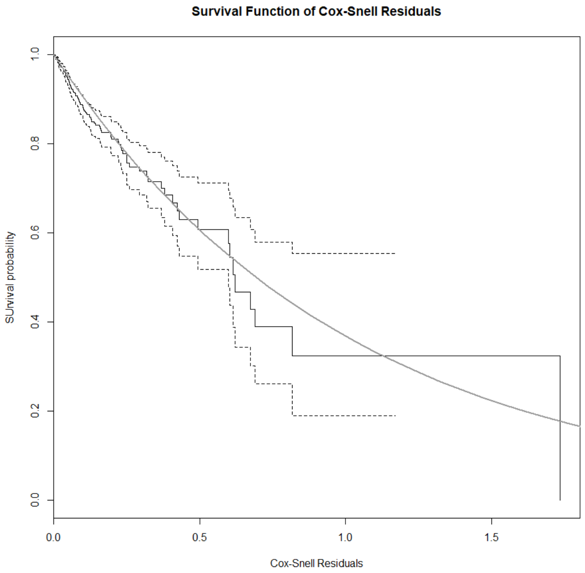


Figure 6.8 – Cox-Snell residuals for the CAT dataset. The black solid lines denote the Kaplan-Meier estimates of the survival function of the residuals (with the dashed lines corresponding to the 95% pointwise confidence intervals), and the grey solid line, the survival function of the unit exponential distribution

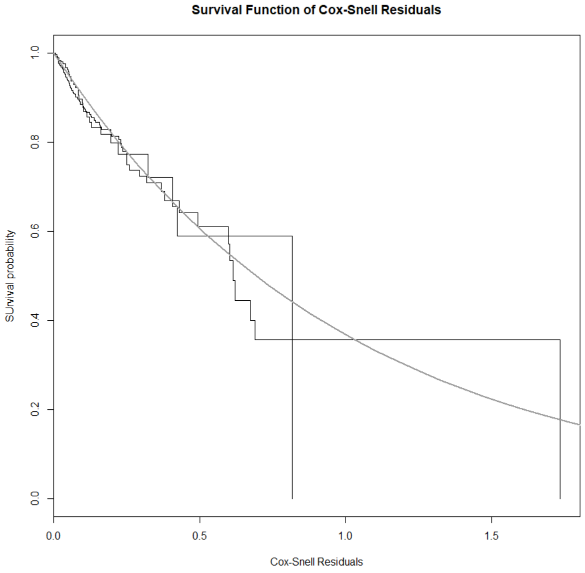


Figure 6.9 – Cox-Snell residuals for the CAT dataset. The black solid lines denote the Kaplan-Meier estimates of the survival functions of the Cox-Snell residuals for the two sites, and the grey solid line, the survival function of the unit exponential distribution

Thus in summary we can conclude that our joint model fitted quite well with fairly good diagnostic attributes.

6.6.1 Summary

We fitted different types of linear mixed models using SAS procedure MIXED, namely; marginal and random effects models under univariable and multivariable models. After comparing the models, the random effects model was found to be the best model for our data under univarible and multivariable using the AIC. Moreover, the best covariance structures to model the between and within subject variation were the unstructured and the spatial spherical covariance structure under both ML and REML. Results from the un-adjusted and adjusted hazards regression both found CD4:CD8 ratio, viral load, gender and age of patients to be significant predictors of mortality. The joint model indicated that CD4 count change due to HAART and mortality had been influenced jointly by gender, age, baseline viral load, baseline CD8 count, time (in years) , CD4:CD8 ratio and by the interaction effects of time (in years) with TB status, baseline viral load and baseline CD8 cell count. Model diagnostics showed that the joint model was the best fit to our data.

Chapter 7

Discussion and concluding remarks

The main aim of this research project was to use joint modelling data analysis techniques for longitudinal and time-to-event data to study the effect of CD4 count on mortality in patients initiated on HAART from rural and urban KwaZulu-Natal. In addition to determine if the patients baseline BMI, baseline age, gender, baseline viral load, baseline CD8 count, TB status and clinic site, influences the rate of change in CD4 count over time .

An in depth literature review on joint models for longitudinal and time-to-event data was done and discussed in chapter 1.

Chapter 2 explored the patients baseline characteristics and distributional properties of the biomarkers. CD4 count, CD8 count and CD4:CD8 ratio violated the normality assumption and had to be square root transformed based on previous literature, after the transformation a better approximation to the normal distribution was observed. The viral load was transformed using a logarithm approximation. The spaghetti plots indicated some within and between patient variation which suggested that a model with both random intercepts and slopes could be plausible. The mean plots suggested an increase in the evolution of CD4 count over time after patients had been initiated on HAART. The mean plots of CD4 count by site, gender and TB status showed that female patients, those patients from the EThekwini site and patients with prevalent TB had a higher rate of change in CD4 count compared to their counterparts.

The linear mixed model was examined in Chapter 3. Two estimation methods (Maximum likelihood and Restricted maximum likelihood) were discussed including their advantages and draw backs. The linear regression that is used for a continuous outcome assumes that the observations being used are independent and identically distributed. However, longitudinal data may be correlated. Thus modeling correlated data without factoring in the lack of independence either over estimates or underestimates the standard errors; which consequently also affects the p-values and confidence intervals. As such, one either over-estimates or under-estimates the effect of the covariates on the outcome. Capturing the best covariance structure of the repeated measurements within subjects is of great importance in longitudinal analysis. This is because misspecification of the covariance structure for repeated measures in longitudinal analysis may lead to biased estimates of the model parameters. Different types of covariance structures were briefly described and the best structure can be selected by choosing the model with a structure that gives the lowest Akaike Information Criteria (AIC). Different types of residuals for model diagnostics were discussed

Chapter 4 examined the theory of survival analysis, the main focus was on the Cox proportional hazards model. Different resdiuals for checking the PH assumptions were briefly described. Strategies for dealing with non proportionality were also briefly described including an alternative method for modelling survival data when the PH assumption fails.

In chapter 5, the joint modelling approach of Rizopoulos (2012) was examined. Estimation methods such as the two stage methods and likelihood methods were discussed. Advantages for joint models over separate analyses were briefly discussed and residuals for model diagnostics were briefly discussed.

In chapter 6, the methods developed in Chapters 3, 4 and 5 were applied to data on HIV positive individuals initiated on HAART in the CAT study. This data had a lot of missing values thus mixed model were very powerful in handling this missingness because the model is valid under MAR (missing at random) which is a less restrictive assumption than the MCAR (missing completely at random) assumption. Furthermore linear mixed models allowed us to explicitly model individual change across time, and presented a very flexible specification of the covariance structure among repeated measure. Often longitudinally measured data and time-to-event or survival data are associated in some ways. In this research project patients infected with HIV were monitored until they developed AIDS or died, and they were regularly measured on intermittent visit for the condition of the immune system using markers such as the CD4 count and CD8 count or the estimated viral load. Thus introducing the association between time to event and the longitudinal trajectories.

Separate analyses of longitudinal data and survival data are not applicable in this case because they may lead to inefficient or biased results. Joint models, on the other hand, provide valid and efficient inferences by optimally incorporating all available information.

Results from the linear mixed models, showed that the random effects model both under uni-variable and multi-variable proved to be a better fit for our data compared to the marginal model because it had the smaller AIC. The best covariance structures to model the between and within subject variation were the unstructured and the spatial spherical covariance structure under both ML and REML. The results from the multi-variable random effects model in particular showed no statistical difference between the eThekwini and Vulindlela sites in terms of the CD4 count improvement over time, with patients from the EThekwini site having a higher rate of change. This finding reaffirms the results obtained by Yende (2010). Men and older people on average had a significantly lower mean rate of change in CD4 count. These results support those obtained by Maskew et al. (2013) who found that men gained fewer CD4 cells after treatment initiation compared to women. Patients presenting without TB at ART initiation started HAART with high mean CD4 count compared to those with prevalent TB but their rate of change in CD4 count was significantly less compared to those with prevalent TB . (Yende (2010); Maskew et al. (2013); Seyoum & Temesgen (2017) and Prins et al. (1999)) also found that CD4 count change was affected by covariates such as age, weight, gender and visiting times, thereby affirming our results. Previous studies using the CAPRISA CAT data and similar datasets in South Africa focused on separately modelling mortality and the longitudinal HIV biomarkers such as CD4 counts. In this work, we consider joint modelling as the best approach for with dealing the two outcomes.

Results from the Cox proportional hazards regression analysis showed that patients from the EThekwini site (urban) had a higher survival prognosis compared to those from the Vulindlela site (rural), however this was not significant at 5% level of significance. Patients presenting without TB at ART initiation had an elevated risk of dying compared to those with prevalent TB. These results reaffirms the results obtained by (Dawood et al., 2018). Prevalent TB also previously showed to be associated with a low mortality, maybe related to TB care being an access point to earlier ART initiation (Dawood et al., 2018). Published literature has cited that undiagnosed TB is higher among patients accessing ART than in the general population; with the majority of incident TB diagnosed in the early weeks of ART initiation being TB prevalent but missed at baseline screening (Etard et al., 2006). In addition male patients, older people or those with a higher mean Log_{10} viral load had a significantly higher risk of death (refer to Table 6.28). The finding is in consonance with previous research which showed that men and older patients were at an increased risk of mortality (Maskew et al. (2013); Dawood et al. (2018) and Prins et al. (1999)). Furthermore, patients with a CD4:CD8 ratio greater than or equal to 0.05 had a significantly lower risk of death compared to those with CD4:CD8 ratio less than 0.05.

Joint models were advantageous for answering multivariate questions at the same time (in our case CD4 count and mortality). One of the most important tools in joint models is its ability to capture or take into consideration the association between the survival time and repeated measurement of a risk factor variable (Rizopoulos, 2012). The joint model found a significantly strong association between the square root CD4 count and the risk for death, implying that CD4 count is a really good predictor of mortality. The joint model also helped assess the correlation between the two response variables and gave ample opportunity to see predictors of the two response variables jointly. The result in this study indicated that CD4 count change due to to HAART and mortality had been influenced jointly by some of the covariates like gender, age, baseline viral load, baseline CD8 count, time (in years) , CD4:CD8 ratio and by the interaction effects of time (in years) with TB status, baseline viral load and baseline CD8 count (refer to Table 6.29). Research findings from a longitudinal study by Guo & Carlin (2004) also proved that CD4 count change was affected by many of these covariates. The joint models were fitted using the JM package in R. Model residuals were calculated and the traditional approach of inspection of residual plots was used to check model assumptions (Rizopoulos et al., 2010).

All the objectives of this study were met. We found that after ART initiation the CD4 count increases and is influenced by measured covariates such as age, gender and TB status. Furthermore, gender, baseline viral load and CD4:CD8 ratio were found to be significant predictors of mortality due to HIV/AIDS. The joint model found a strong association between CD4 count and mortality which means that CD4 is a predictor of mortality. These results are consonant with previous research.

The application of linear mixed models revealed that residual correlation is present. The joint model, however, makes the somewhat restrictive assumption in the longitudinal component that, conditional on the random effects, the residuals are uncorrelated. However, as discussed by Rizopoulos (2012), extending a linear mixed model by including a more elaborate random effects structure is computationally simpler to implement and can produce almost indistinguishable fits to the data when compared with a model that includes a serial correlation term.

In biomedical research where measurements of various outcomes are taken over a time period in an attempt to understand patients health or the risk of an event occurring, the joint modelling approach will be the most useful tool to consider in an effort to link the longitudinal and survival outcomes. Though joint modelling maybe the most suitable approach, it has a low convergence rate mainly because there are a very large number of parameters that need to be estimated when considered under the Markov Chain Monte Carlo method (Guure et al., 2017).

An area for future work would be to jointly model multiple longitudinal outcomes (CD4:CD8 ratio counts and viral loads), mortality and TB infection status adjusting for possible informative drop-out (i.e. departures from the MAR assumption).

References

Aalen, O. (1980). A model for nonparametric regression analysis of counting processes. In *Mathematical statistics and probability theory*, (pp. 1–25). Springer.

Allen, D. M. (1974). The relationship between variable selection and data agumentation and a method for prediction. *Technometrics*, *16*(1), 125–127.

Andrinopoulou, E.-R., Rizopoulos, D., Takkenberg, J. J., & Lesaffre, E. (2014). Joint modeling of two longitudinal outcomes and competing risk data. *Statistics in medicine*, *33*(18), 3167–3178.

Antonio, K., & Beirlant, J. (2007). Actuarial statistics with generalized linear mixed models. *Insurance: Mathematics and Economics*, *40*(1), 58–76.

Asar, Ö., Ritchie, J., Kalra, P. A., & Diggle, P. J. (2015). Joint modelling of repeated measurement and time-to-event data: an introductory tutorial. *International journal of epidemiology*, *44*(1), 334–344.

Beckman, R. J., Nachtsheim, C. J., & Cook, R. D. (1987). Diagnostics for mixed–model analysis of variance. *Technometrics*, *29*(4), 413–426.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). Regression diagnostics. j.

Bennett, D. E., Bertagnolio, S., Sutherland, D., Gilks, C. F., et al. (2008). The World Health Organization's global strategy for prevention and assessment of HIV drug resistance. *Antiviral therapy*, *13*, 1.

Britton, T. (1997). Tests to detect clustering of infected individuals within families. *Biometrics*, (pp. 98–109).

Chen, Q., May, R. C., Ibrahim, J. G., Chu, H., & Cole, S. R. (2014). Joint modeling of longitudinal and survival data with missing and left-censored time-varying covariates. *Statistics in medicine*, *33*(26), 4560–4576.

Cohen, J., Cohen, P., West, S. G., Aiken, L. S., et al. (1983). Applied multiple regression/correlation analysis for the behavioral sciences.

Collett, D. (2015). *Modelling survival data in medical research*. Chapman and Hall/CRC.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, *19*(1), 15–18.

Cox, D. R. (1972). Models and life-tables regression. *JR Stat. Soc. Ser. B*, *34*, 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, *62*(2), 269–276.

Cox, D. R. (1983). Some remarks on overdispersion. *Biometrika*, *70*(1), 269–274.

Dawood, H., Hassan-Moosa, R., Zuma, N.-Y., & Naidoo, K. (2018). Mortality and treatment response amongst HIV-infected patients 50 years and older accessing antiretroviral services in South Africa. *BMC infectious diseases*, *18*(1), 168.

De Beaudrap, P., Etard, J.-F., Diouf, A., Ndiaye, I., Guèye, N. F., Guèye, P. M., Sow, P. S., Mboup, S., Ndoye, I., Ecochard, R., et al. (2009). Modeling CD4+ Cell Count Increase Over a Six-Year Period in HIV-1 Infected Patients on Highly Active Antiretroviral Therapy in Senegal. *The American journal of tropical medicine and hygiene*, *80*(6), 1047–1053.

De Gruttola, V., & Tu, X. M. (1994). Modelling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics*, (pp. 1003–1014).

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, (pp. 1–38).

Ding, J., & Wang, J.-L. (2008). Modeling longitudinal data with nonparametric multiplicative random effects jointly with survival data. *Biometrics*, *64*(2), 546–556.

Dobson, A., & Henderson, R. (2003). Diagnostics for joint longitudinal and dropout time modeling. *Biometrics*, *59*(4), 741–751.

Etard, J.-F., Ndiaye, I., Thierry-Mieg, M., Guèye, N. F. N., Guèye, P. M., Laniece, I., Dieng, A. B., Diouf, A., Laurent, C., Mboup, S., et al. (2006). Mortality and causes of death in adults receiving highly active antiretroviral therapy in Senegal: a 7-year cohort study. *Aids*, *20*(8), 1181–1189.

Faucett, C. L., Schenker, N., & Taylor, J. M. (2002). Survival analysis using auxiliary variables via multiple imputation, with application to AIDS clinical trial data. *Biometrics*, *58*(1), 37–47.

Faucett, C. L., & Thomas, D. C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in medicine*, *15*(15), 1663–1685.

Fox, J. (2002). Cox proportional-hazards regression for survival data. *An R and S-PLUS companion to applied regression, 2002*.

Gandhi, N. R., Moll, A., Sturm, A. W., Pawinski, R., Govender, T., Lalloo, U., Zeller, K., Andrews, J., & Friedland, G. (2006). Extensively drug-resistant tuberculosis as a cause of death in patients co-infected with tuberculosis and HIV in a rural area of South Africa. *The Lancet*, *368*(9547), 1575–1580.

Gianola, D., & Foulley, J. L. (1990). Variance estimation from integrated likelihoods (VEIL). *Genetics Selection Evolution*, *22*(4), 403.

Gilmour, A. R., Thompson, R., & Cullis, B. R. (1995). Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, (pp. 1440–1450).

Grambsch, P. M., & Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, *81*(3), 515–526.

Guo, X., & Carlin, B. P. (2004). Separate and joint modeling of longitudinal and event time data using standard computer packages. *The American Statistician*, *58*(1), 16–24.

Guure, C. B., Ibrahim, N. A., Adam, M. B., & Said, S. M. (2017). Joint modelling of longitudinal 3MS scores and the risk of mortality among cognitively impaired individuals. *PloS one*, *12*(8), e0182873.

Hallahan, C. (2003). Longitudinal Data analysis with Discrete and Continuous responses using Proc Mixed. *Maintained at: http://www. cpcug. org/user/sigstat/PowerPointSlides*.

Hanagal, D. D. (2011). *Modeling survival data using frailty models*. Chapman and Hall/CRC.

Hartley, H. O., & Rao, J. N. (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika*, *54*(1-2), 93–108.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, *72*(358), 320–338.

Hemmerle, W. J., & Hartley, H. O. (1973). Computing maximum likelihood estimates for the mixed AOV model using the W transformation. *Technometrics*, *15*(4), 819–831.

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics*, *9*(2), 226–252.

Henderson, R., Diggle, P., & Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, *1*(4), 465–480.

Hosmer, D. W., Royston, P., et al. (2002). Using aalens linear hazards model to investigate time-varying effects in the proportional hazards regression model. *Stata J*, *2*(4), 331–350.

Hsieh, F., Tseng, Y.-K., & Wang, J.-L. (2006). Joint modeling of survival and longitudinal data: likelihood approach revisited. *Biometrics*, *62*(4), 1037–1043.

Hunter, D. R., & Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician*, *58*(1), 30–37.

Jacqmin-Gadda, H., & Commenges, D. (1995). Tests of homogeneity for generalized linear models. *Journal of the American Statistical Association*, *90*(432), 1237–1246.

Karim, S. A., & Karim, Q. A. (2010). *HIV/AIDS in South Africa*. Cambridge University Press.

Karim, S. A., Karim, Q. A., & Baxter, C. (2005). Overview of the book. In *HIV/AIDS in South Africa*, (pp. 37–47). Cambridge University Press.

Kincaid, C. (2005). Guidelines for selecting the covariance structure in mixed model analysis. In *Proceedings of the Thirtieth Annual SAS Users Group International Conference*, 198-30. SAS Institute Inc Cary NC.

Kleinbaum, D. G., & Klein, M. (2005). Competing risks survival analysis. *Survival Analysis: A self-learning text*, (pp. 391–461).

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, (pp. 963–974).

Lee, E. T., & Wang, J. (2003). *Statistical methods for survival data analysis*, vol. 476. John Wiley & Sons.

Liang, K.-Y. (1987). A locally most powerful test for homogeneity with many strata. *Biometrika*, *74*(2), 259–264.

Lin, D. Y., Wei, L.-J., & Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika*, *80*(3), 557–572.

Littell, R. C., Pendergast, J., & Natarajan, R. (2000). Modelling covariance structure in the analysis of repeated measures data. *Statistics in medicine*, *19*(13), 1793–1819.

Lo, J., Abbara, S., Shturman, L., Soni, A., Wei, J., Rocha-Filho, J. A., Nasir, K., & Grinspoon, S. K. (2010). Increased prevalence of subclinical coronary atherosclerosis detected by coronary computed tomography angiography in HIV-infected men. *AIDS (London, England)*, *24*(2), 243.

Lu, W.-S. (1997). Score tests for overdispersion in poisson regression models. *Journal of Statistical Computation and Simulation*, *56*(3), 213–228.

Margolick, J. B., Muñoz, A., Donnenberg, A. D., Park, L. P., Galai, N., Giorgi, J. V., O'Gorman, M. R., & Ferbas, J. (1995). Failure of T-cell homeostasis preceding AIDS in HIV-1 infection. *Nature medicine*, *1*(7), 674.

Martinussen, T., & Scheike, T. H. (2007). *Dynamic regression models for survival data*. Springer Science & Business Media.

Maskew, M., Brennan, A. T., Westreich, D., McNamara, L., MacPhail, A. P., & Fox, M. P. (2013). Gender differences in mortality and CD4 count response among virally suppressed HIV-positive patients. *Journal of women's health*, *22*(2), 113–120.

Meyer, K. (1991). Estimating variances and covariances for multivariate animal models by restricted maximum likelihood. *Genetics Selection Evolution*, *23*(1), 67.

Miller, J. J. (1973). Asymptotic properties and computation of maximum likelihood estimates in the mixed model of the analysis of variance. Tech. rep., STANFORD UNIV CA DEPT OF STATISTICS.

Nelder, J. (1954). The interpretation of negative components of variance. *Biometrika*, *41*(3/4), 544–548.

O'Brien, L. M., & Fitzmaurice, G. M. (2004). Analysis of longitudinal multiple-source binary data using generalized estimating equations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *53*(1), 177–193.

Ohara Hines, R. (1997). A comparison of tests for overdispersion in generalized linear models. *Journal of Statistical Computation and Simulation*, *58*(4), 323–342.

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, *58*(3), 545–554.

Pawitan, Y., & Self, S. (1993). Modeling disease marker processes in AIDS. *Journal of the American Statistical Association*, *88*(423), 719–726.

Prins, M., Robertson, J. R., Brettle, R. P., Aguado, I. H., Broers, B., Boufassa, F., Goldberg, D. J., Zangerle, R., Coutinho, R. A., & van den Hoek, A. (1999). Do gender differences in CD4 cell counts matter? *Aids*, *13*(17), 2361–2364.

RA Fisher, M. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. R. Soc. Lond. A*, *222*(594-604), 309–368.

Ramroop, S. (2010). Analysis of longitudinal binary data: an application to a disease process, Ph.D thesis, University of KwaZulu-Natal.

Reda, A. A., Biadgilign, S., Deribew, A., Gebre, B., & Deribe, K. (2013). Predictors of change in CD4 lymphocyte count and weight among HIV infected patients on anti-retroviral treatment in Ethiopia: a retrospective longitudinal study. *PLoS One*, *8*(4), e58595.

Reddy, T., Molenberghs, G., Njagi, E. N., & Aerts, M. (2016). A novel approach to estimation of the time to biomarker threshold: applications to HIV. *Pharmaceutical statistics*, *15*(6), 541–549.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. Chapman and Hall/CRC.

Rizopoulos, D., Verbeke, G., & Lesaffre, E. (2009). Fully exponential laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*(3), 637–654.

Rizopoulos, D., Verbeke, G., & Molenberghs, G. (2010). Multiple-imputation-based residuals and diagnostic plots for joint models of longitudinal and survival outcomes. *Biometrics*, *66*(1), 20–29.

Sainz, T., Serrano-Villar, S., Díaz, L., Tomé, M. I. G., Gurbindo, M. D., de José, M. I., Mellado, M. J., Ramos, J. T., Zamora, J., Moreno, S., et al. (2013). The CD4/CD8 ratio as a marker T-cell activation, senescence and activation/exhaustion in treated HIV-infected children and young adults. *Aids*, *27*(9), 1513–1516.

Schabenberger, O. (2005). Mixed model influence diagnostics. In *SUGI*, vol. 29, (pp. 189–29). Citeseer.

Scheike, T. H. (2006). *Dynamic Regression Models for Survival Data. Statistics for Biology and Health*. Springer.

Schemper, M. (1992). Cox analysis of survival data with non-proportional hazard functions. *The Statistician*, (pp. 455–465).

Searle, S., Casella, G., & McCulloch, C. (2008). Maximum likelihood (ML) and restricted maximum likelihood (REML). *Variance components. John Wiley & Sons, Inc., Hoboken, NJ*, (pp. 232–257).

Self, S., & Pawitan, Y. (1992). Modeling a marker of disease progression and onset of disease. In *AIDS epidemiology*, (pp. 231–255). Springer.

Seyoum, A., & Temesgen, Z. (2017). Joint longitudinal data analysis in detecting determinants of CD4 cell count change and adherence to highly active antiretroviral therapy at Felege Hiwot Teaching and Specialized hospital, North-west Ethiopia (Amhara Region). *AIDS research and therapy*, *14*(1), 14.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3/4), 591–611.

Smith, P. J., & Heitjan, D. F. (1993). Testing and adjusting for departures from nominal dispersion in generalized linear models. *Applied Statistics*, (pp. 31–41).

Taylor, J., Fahey, J. L., Detels, R., & Giorgi, J. V. (1989). CD4 percentage, CD4 number, and CD4: CD8 ratio in HIV infection: which to choose and how to use. *Journal of acquired immune deficiency syndromes*, *2*(2), 114–124.

Therneau, T. M., & Grambsch, P. M. (2000). Testing proportional hazards. In *Modeling survival data: extending the Cox model*, (pp. 127–152). Springer.

Therneau, T. M., & Grambsch, P. M. (2013). *Modeling survival data: extending the Cox model*. Springer Science & Business Media.

Therneau, T. M., Grambsch, P. M., & Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika*, *77*(1), 147–160.

Thompson Jr, W. (1962). Estimation of dispersion parameters. *J. Res. Natl. Bur. Standards Sec. B*, *66*, 161–164.

Tsiatis, A., Degruttola, V., & Wulfsohn, M. (1995). Modeling the relationship of survival to longitudinal data measured with error. applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*, *90*(429), 27–37.

UNAIDS (2016). Fact sheet: latest statistics on the status of the AIDS epidemic.

UNAIDS, W. (2017). Fact sheet: World AIDS day 2017. *Global HIV statistics*.

Van Walraven, C., & Hart, R. G. (2008). Leave em alone–why continuous variables should be analyzed as such. *Neuroepidemiology*, *30*(3), 138–139.

Verbeke, G. (1997). Linear mixed models for longitudinal data. In *Linear mixed models in practice*, (pp. 63–153). Springer.

Verbeke, G., & Molenberghs, G. (2000). Linear mixed models for longitudinal data.

Wandeler, G., Gsponer, T., Mulenga, L., Garone, D., Wood, R., Maskew, M., Prozesky, H., Hoffmann, C., Ehmer, J., Dickinson, D., et al. (2013). Zidovudine impairs immunological recovery on first-line antiretroviral therapy: collaborative analysis of cohort studies in southern Africa. *AIDS (London, England)*, *27*(14), 2225–2232.

Werner, L. (2010). Modelling acute HIV infection using longitudinally measured biomarker data including informative drop-out, MSc thesis, University of KwaZulu-Natal.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, (pp. 1–25).

Wu, L., Liu, W., Yi, G. Y., & Huang, Y. (2012). Analysis of longitudinal and survival data: joint modeling, inference methods, and issues. *Journal of Probability and Statistics*, *2012*.

Wulfsohn, M. S., & Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, (pp. 330–339).

Xie, X., Strickler, H. D., & Xue, X. (2013). Additive hazard regression models: an application to the natural history of human papillomavirus. *Computational and mathematical methods in medicine*, *2013*.

Ye, W., Lin, X., & Taylor, J. M. (2008). Semiparametric modeling of longitudinal measurements and time-to-event data–a two-stage regression calibration approach. *Biometrics*, *64*(4), 1238–1246.

Yende, N. (2010). Modelling CD4+ count over time in HIV positive patients initiated on HAART in South Africa using linear mixed models. MSc thesis, University of KwaZulu-Natal.

Zhang, G., & Chen, J. J. (2013). Adaptive fitting of linear mixed-effects models with correlated random effects. *Journal of statistical computation and simulation*, *83*(12), 2291–2314.

Appendix A

A.1 LINEAR MIXED MODEL FOR SQUARE ROOT CD4 COUNT

The SAS codes for the final models used in the analysis of the CAT data are given below:

```
/****************************** FINAL MARGINAL MULTIVARIATE MODEL *****************************/

proc mixed data=nobuhle.forjoint method=reml covtest noclprint empirical;
class pid gender1 site prev times;
model sqrtcd4=site gender1 prev age1 sqrtcd8 logviral times_years times_years*site
times_years*prev times_years*age1 times_years*logviral times_years*sqrtcd8 /s;
repeated times/type=sp(SPH)(times) local subject=pid ;
run;


/************************** FINAL RANDOM EFFECTS MULTIVARIATE MODEL *************************/

proc mixed data=nobuhle.forjoint method=reml covtest noclprint empirical;
class pid gender1 site prev times;
model sqrtcd4=site gender1 prev age1 logviral sqrtcd8 times_years times_years*prev
times_years*logviral times_years*sqrtcd8 /s;
repeated times/type=sp(SPH)(times_years) local subject=pid ;
random intercept times_years/ subject=pid type=un ;
run;
```

Appendix B

B.1 COX PH MODEL

```
res.cox ← coxph(Surv(timetodeath, death) ~ factor(gender1) +factor(prev)+factor(site)+
logviral+CD4_CD8_ratio+age1, data = long.id)
summary(res.cox)
```

B.2 JOINT MODEL

```
lmeFit.long ← lme(sqrtcd4 ~ times_years +age1+logviral+factor(gender1)+times_years:factor(prev)+
times_years:logviral+sqrtcd8+times_years:sqrtcd8, random =  times_years | pid, data=long)

survFit ← coxph(Surv(timetodeath, death) ~ factor(gender1)+CD4_CD8_ratio+logviral+age1,
data = long.id, x = TRUE)

jointFit.p1 ← jointModel(lmeFit.long, survFit, timeVar = ”times_years”, method =
”piecewise-PH-aGH”)

summary(jointFit.p1)
exp(confint(jointFit.p1,parm = ”Event”))
```

Appendix C

Table 7.1: Covariance Parameter Estimates by ML under the marginal multivariable model

Cov Parm	Estimate	Standard Error	Z Value	Pr > Z
Variance	22.3900	0.7152	31.30	$< .0001$
SP(SPH)	52.4357	1.9016	27.57	$< .0001$
Residual	2.4140	0.1664	14.51	$< .0001$

Table 7.2: Solution for Fixed Effects by ML under the marginal multivariable model

Effect		Estimate	Standard Error	DF	t Value	Pr > $\|t\|$
Intercept		12.0075	1.1281	1638	10.64	$< .0001$
site	Esite	-0.9664	0.3082	1638	-3.14	0.0017
gender1	Men	-1.9279	0.2741	1638	-7.03	$< .0001$
prev	No TB	0.5807	0.3458	1638	1.68	**0.0933**
age1		-0.00106	0.01357	1638	-0.08	**0.9380**
logviral		-0.7176	0.1423	1638	-5.04	$< .0001$
sqrtcd8		0.2026	0.01408	1638	14.39	$< .0001$
bmi		0.01934	0.01741	1638	1.11	**0.2667**
times_years		3.1422	0.6356	6179	4.94	$< .0001$
times_years*site	Esite	0.5905	0.1872	6179	3.15	0.0016
times_years*gender1	Men	0.1467	0.1530	6179	0.96	**0.3376**
times_years*prev	No TB	-0.7718	0.2488	6179	-3.10	0.0019
times_years*age1		-0.01693	0.007595	6179	-2.23	0.0258
times_years*logviral		0.2976	0.08282	6179	3.59	0.0003
times_years*sqrtcd8		-0.06253	0.006749	6179	-9.27	$< .0001$
times_years*bmi		-0.00232	0.007293	6179	-0.32	**0.7506**

Table 7.3: Fit statistics for different covariance structures by ML and REML under the random effects model

Covariance structure	-2 Log Likelihood	AIC	BIC
Maximum Likelihood (ML)			
SP(SPH)	41488.2	**41532.2**	41651.1
SP(POW)	41488.4	41532.4	41651.3
SP(EXP)	41488.4	41532.4	41651.3
SP(GAU)	41553.0	41597.0	41715.9
SP(LIN)	41623.7	41667.7	41786.7
CS	42061.3	42103.3	42216.9
Restricted Maximum Likelihood (REML)			
SP(SPH)	41557.4	**41569.4**	41601.8
SP(POW)	41557.5	41569.5	41602.0
SP(EXP)	41557.5	41569.5	41602.0
SP(GAU)	41621.8	41633.8	41666.2
SP(LIN)	41692.5	41704.5	41737.0
CS	42129.4	42139.4	42166.4

Table 7.4: Covariance Parameter Estimates by ML under the random effects multivariable model

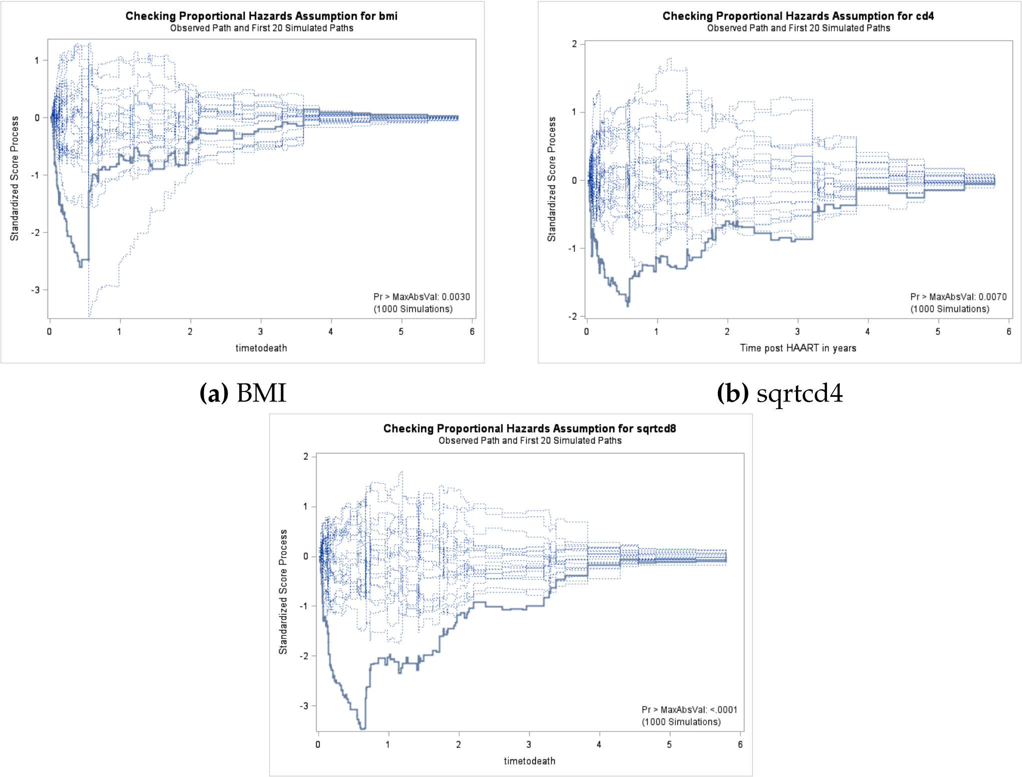
Cov Parm	Estimate	Standard Error	Z Value	Pr > Z
UN(1,1)	8.1372	1.1608	7.01	< .0001
UN(2,1)	-1.8472	0.2677	-6.90	< .0001
UN(2,2)	1.8731	0.2047	9.15	< .0001
Variance	11.3461	1.2279	9.24	< .0001
SP(SPH)	33.4508	3.3528	9.98	< .0001
Residual	2.4942	0.1681	14.84	< .0001

Table 7.5: Solution for Fixed Effects by ML under the random effects multivariable model

Effect		Estimate	Standard Error	DF	t Value	Pr > $\|t\|$
Intercept		11.3982	0.9704	1633	11.75	$<.0001$
site	Esite	-0.5513	0.2711	4698	-2.03	0.0420
gender1	Men	-0.5541	0.2337	4698	-6.65	$<.0001$
prev	No TB	0.5114	0.2931	4698	1.74	**0.0811**
age1		-0.01795	0.01168	4698	-1.54	**0.1246**
logviral		-0.6234	0.1243	4698	-5.02	$<.0001$
sqrtcd8		0.1977	0.01287	4698	15.37	$<.0001$
bmi		0.01941	0.01545	4698	1.26	**0.2090**
times_years		4.8063	0.5377	1486	8.94	$<.0001$
times_years*site	Esite	0.2897	0.1655	4698	1.75	**0.0801**
times_years*gender1	Men	-0.09629	0.1310	4698	-0.73	**0.4624**
times_years*prev	No TB	-1.0345	0.1869	4698	-5.54	$<.0001$
times_years*age1		-0.00848	0.006646	4698	-1.28	**0.2022**
times_years*logviral		0.2150	0.06937	4698	3.10	0.0019
times_years*sqrtcd8		-0.07244	0.006183	4698	-11.72	$<.0001$
times_years*bmi		-0.00423	0.006346	4698	-0.67	**0.5056**

Table 7.6: Test of Equality over Strata

Test	Chi-Square	DF	P-value
TB status			
Log-Rank	0.3242	1	0.5691
Wilcoxon	0.2075	1	0.6487
-2Log(LR)	0.00029	1	0.9881
Site			
Log-Rank	0.3242	1	0.5691
Wilcoxon	0.2075	1	0.6487
-2Log(LR)	0.00029	1	0.9881
Gender			
Log-Rank	0.3242	1	0.5691
Wilcoxon	0.2075	1	0.6487
-2Log(LR)	0.00029	1	0.9881

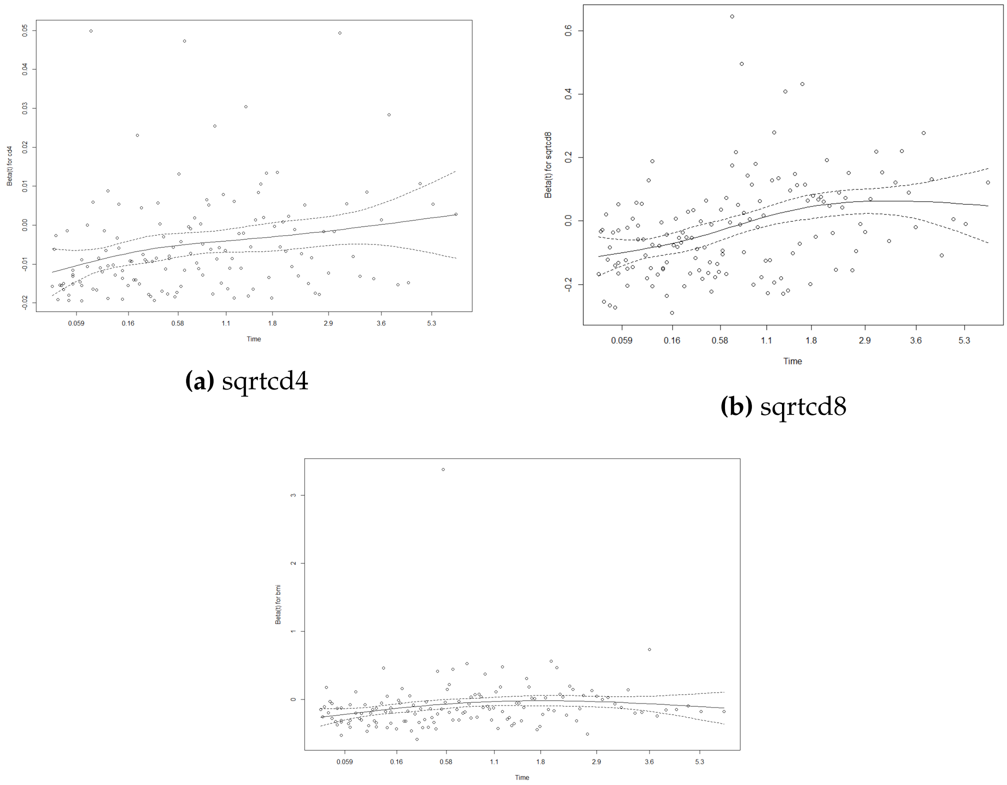


(a) BMI

(b) sqrtcd4

(c) sqrtcd8

Figure 7.1 – Score process for variables that violate the Cox PH assumption



(a) sqrtcd4

(b) sqrtcd8

(c) BMI

Figure 7.2 – Time trend of the hazard ratio for variables that violate the Cox PH assumption