Modelling CD4+ Count Over Time in HIV Positive Patients Initiated on HAART in South Africa Using Linear Mixed Models

A thesis presented to

The University of KwaZulu Natal

in fulfillment of the requirement for the degree

of

Master of Science in Statistics

by

Nonhlanhla Yende

School of Statistics and Actuarial Science

December 2009

Abstract

HIV is among the highly infectious and pathogenic diseases with a high mortality rate. The spread of HIV is influenced by several individual based epidemiological factors such as age, gender, mobility, sexual partner profile and the presence of sexually transmitted infections (STI). CD4+ count over time provided the first surrogate marker of HIV disease progression and is currently used for clinical management of HIV-positive patients. The CD4+ count as a key disease marker is repeatedly measured among those individuals who test HIV positive to monitor the progression of the disease since it is known that HIV/AIDS is a long wave event. This gives rise to what is commonly known as longitudinal data.

The aim of this project is to determine if the patients' weight, baseline age, sex, viral load and clinic site, influences the rate of change in CD4+ count over time. We will 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 September 2006, including two years of follow-up for each patient. Analysis was done using linear mixed models methods for longitudinal data. The results showed that larger increase in CD4+ count over time was observed in females and individuals who were younger. However, upon fitting baseline log viral load in the model instead of the log viral at all visits was that, larger increase in CD4+ count was observed in females, individuals who were younger, had higher baseline log viral load and lower weight.

Declaration

The research work is the original work done by the author (Nonhlanhla Yende) and it is not a duplicate of some of the research work done by other authors. All the references that were used to refer to are duly acknowledged.

Ms. Nonhlanhla Yende (201502966)	Date
Prof. Henry Mwambi (Supervisor)	Date

Notes

One paper has been drafted from this thesis.

1. Factors associated with the rate of increase in CD4+ count over the first two years in patients initiated on HAART in KwaZulu Natal, South Africa. *Presented at the Center for the AIDS Programme of Research in South Africa (CAPRISA) Academic Day, Durban, South Africa, 28 August 2009.* The presentation of the paper won the first prize.

Acknowledgements

I wish to thank the Center for the AIDS Programme of Research in South Africa (CAPRISA) for allowing me to use their data. I also thank my supervisor, Prof. Henry Mwambi for his assistance. To my fiancé, Zamani Zuma thank you for all your courage and support.

"Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write." H.G.Wells

Contents

Chapter 1

Introduction

Since the emerging of the Human Immunodeficiency Virus (HIV), South Africa has experienced an unprecedented HIV prevalence. In South Africa, some 5.5 million [4.9 million-6.1 million] people, including 240 000 [93 000-500 000] children younger than 15 years, were living with HIV in 2005 (Joint United Nations Programme on HIV/AIDS, 2006). Acquired Immunodeficiency Syndrome (AIDS) was first reported in South Africa in 1983 and as in most of Africa, AIDS first became apparent as an urban phenomenon in South Africa but it spread rapidly into rural areas (Abdool Karim and Abdool Karim, 2005). HIV is among the highly infectious and pathogenic diseases with a high mortality rate.

The spread of HIV is influenced by several individual based epidemiological factors such as age, gender, mobility, sexual partner profile and the presence of sexually transmitted infections (STI). There is a continuing, rising trend nationally in HIV infection levels among pregnant women attending public antenatal clinics. National HIV prevalence figures based on antenatal care data ranged from 22.4% in 1999 to 30.2% in 2005 (Department of Health, 2006). HIV prevalence estimates give the number and proportion of people who are living with HIV at a given point in time. However, HIV prevalence data cannot tell us what proportion of HIV positive people are in the later stages of HIV infection and at risk of progressing to fully blown AIDS.

In order to analyze the prognosis of patients infected with HIV, we use the CD4+ count. CD4+ count is the measure of the number of helper T cells per cubic millimeter of blood. T helper cells are a sub-group of lymphocytes (a type of white blood cell or leukocyte) that play an important role in establishing and maximizing the capabilities of the immune system. CD4+ cells are a vital component of the immune system and also a prime target of HIV infection and HIV infection is characterised by continuous loss of CD4+ cells (O'Brien et al., 1996). CD4+ count provided the first reliable marker of disease progression (Abdool Karim and Abdool Karim, 2005) as compared

to other possible markers and it is one of the markers most closely correlated with the stage of HIV infection (Prins et al., 1999). After few weeks post infection the viral load peaks and CD4+ count declines dramatically. However, after within few weeks the immune system responds to HIV resulting in a decline in viral load and CD4+ count return to near normal values.

The CD4+ count as a key disease marker is repeatedly measured among those individuals who test HIV positive to monitor the progression of the disease since it is known that HIV/AIDS is a long wave event. This gives rise to what is commonly known as longitudinal data. It follows from this that longitudinal studies are needed to establish the effects on individuals over time. The response to the AIDS epidemic in South Africa developed slowly at first. In the 1990s it suffered huge crises of credibility, faltering seriously in the late 1990s despite some momentum in the period just after the dawn of democracy in 1994, but in the present decade it has been gathering momentum once again with the announcement by the government in 2003 that it would make free antiretroviral treatment available in the public health service (Abdool Karim and Abdool Karim, 2005).

CD4+ count is used to make a decision as to when to commence highly active antiretroviral therapy (HAART). Against this background the government's Department of Health has adopted the provision of ART tables in its guidelines (Department of Health, 2006) to help in decision making as to when to initiate HAART to an infected patient. The decision to provide HAART is complex. The criteria to provide HAART seem to differ between the developed and developing countries. According to South Africa's ART programme, an individual who is HIV positive can only be initiated on HAART once his/her CD4+ count is less than or equal to 200 cells/μL or if a patient presents with certain clinical symptoms and also if that particular patient is ready and understands the importance of adherence very well. On the other hand, in the United States HAART is initiated when the CD4+ count is less than 350 cells/μL.

In the absence of any antiretroviral therapy, the median time to AIDS from the point of HIV infection is 8 to 10 years, at least in the USA and Europe where there is generally good access to health care (Abdool Karim and Abdool Karim, 2005). It is always a good idea not to start HAART at early stages of infection because ARV drugs do not make any difference to a person's health. The longer a person's immune system is exposed to HAART, the more likely the chance that the HIV will develop resistance to treatment and become no longer beneficial to the patient. Starting HAART too early also means that serious side effects associated with HAART are allowed to set in unnecessarily.

Thus, for planning purposes, efficiency, high acceptance and adherence, it is important to understand the rates at which CD4+ count decrease to reach the minimum required number to initiate

HAART and also the rate at which CD4+ count increases after HAART initiation to reach an acceptable level. An acceptable or normal CD4+ count for an adult person is between 500 and 1200 cells/μL. On the other hand Diggle et al. (1994, 2002) report that an uninfected individual has around 1100 cells/μL of blood. The factors affecting the rate of change are critical. These factors can be grouped broadly into socio-demographic and biomedical factors. The purpose of the current study is to understand factors that influence increase or gain in CD4+ count for patients on HAART. In other words, the patients in this study were all under HAART and we wish to study the factors that influence the success of HAART on them.

Although, worldwide, there are as many women as there are men with HIV infection, this averaged figure conceals marked geographical differences in gender distribution of the disease (Abdool Karim and Abdool Karim, 2005). Strategies to prevent HIV/AIDS should include the education to promote delayed onset of sexual activity since the HIV/AIDS is quite prevalent amongst women as compared to men in the age group 15-24 years. For social, cultural and economic reasons men are usually in a stronger position in their relationships with women and this gives them more control in deciding when to have sex as well as whether or not to use the condom. This phenomenon is particularly more apparent in developing countries. This issue of promoting delayed onset of sexual activities is just but one of the protective measures about HIV/AIDS among other things.

Furthermore, adherence to HAART needs further understanding. Adherence is when a person initiated on HAART take medications as prescribed by a healthcare provider, in the exact dose (number of pills/tablets/capsules) and at the right times. If the patient is not taking medications as prescribed, the medication may not provide the required benefit intended for. Poor adherence to therapy may also allow HIV to develop resistance to anti-HIV medications. When this happens, viral load goes up and CD4+ count drops, signalling treatment failure. Naturally therefore we expect CD4+ count and viral load to be negatively correlated. For someone who is on HAART or on other immune boosting medication we expect their CD4+ count to increase and their viral load to decrease.

However, individual responses are quite variable and the correlation between CD4+ count response and viral load in some individual is very weak (Abdool Karim and Abdool Karim, 2005). CD4+ count is a measure of strength of the immune system. Higher CD4+ count imply a strong immune system while low CD4+ count implies a weak immune system. The CD4+ count does not always reflect how someone with HIV feels and functions. This means that there could be other latent factors which influence the dynamics of the disease. It should, however, be remembered that surrogate markers do not precisely reflect clinical outcomes (Abdool Karim and Abdool Karim, 2005).

Surrogate endpoints are collected in a shorter time period and are proposed based on biological considerations within a progression model of disease. One example is CD4+ count levels in AIDS; the CD4+ count can potentially serve as a surrogate endpoint for death (Ghosh, 2008). Modelling surrogate endpoints has been the focus of much recent statistical research (Burzykowski, Molenberghs and Buyse, 2005). Studies have shown that HAART reduces both mortality and morbidity in people infected with HIV. As viral replication falls, the CD4+ count increases, but whether the CD4+ count returns to the level seen in HIV negative people is still unknown (Mocroft et al., 2007).

1.1 Data description

This research project will use data collected at two sites, eThekwini (Durban) and Vulindlela (near Howick) 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 eThekwini site is situated in an urban area while the latter is in a rural area. 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 April 2005 and started enrolling again in November 2005. 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 Mafakathini 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).

The reason for recruiting patients infected with tuberculosis (TB) is that, HIV greatly increases the risk of active tuberculosis disease and about 80% of patients presenting with active tuberculosis in the province of KwaZulu Natal, South Africa, are co-infected with HIV (Gandhi et al., 2006). The rising incidence of TB has been attributed to HIV co-infection especially in developing countries. Recruited people receive Voluntary Counselling and Testing (VCT) from trained counsellors. In developed countries with epidemics in high risk core groups, high-quality VCT has been shown to substantially reduce the incidence of sexually transmitted diseases (STD) especially if supplemented with increased condom use (Sherr et al., 2007).

In this study patients who are HIV positive get screened to check if they are eligible for the CAPRISA AIDS Treatment Project. Considering the delay in South African Department of

Health's HAART roll out, the CAPRISA AIDS Treatment Project helps by rolling out HAART to people who are HIV positive. The eligibility is to have CD4+ count of less than or equal to 200 cells/μL and be at least 14 years of age, but if the CD4+ count is greater than 200 cells/μL and a patient is very sick he/she is still initiated on HAART for ethical reasons. The CD4+ count and viral load are measured at baseline and at every six months interval thereafter.

Some patients come 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. This raises a lot of questions such as whether HIV prevalence or incidence of HIV is higher for women than for men. Or are women sensitive to better care of their lives and therefore get HIV tested whenever they are not feeling well hence accessing ARVs as soon as possible. Maybe it is because the clinics are primarily antenatal centres and hence more women are expected to attend.

1.2 What drives young women into sex at young age?

We know that South Africa is one of the developing countries where poverty prevalence is the most critical challenges that the government is facing. Women are increasingly becoming infected with HIV and deaths due to HIV/AIDS have left a large number of children as orphans. In Sub-Saharan Africa alone, the epidemic has orphaned nearly 12 million children less than 18 years (Joint United Nations Programme on HIV/AIDS, 2008). This is definitely counter-productive to the government's efforts to eradicate poverty. Some young women are the victims of HIV/AIDS in a sense that they are the ones looking after their younger siblings because their parents were killed by HIV/AIDS. Some younger and older women are compelled to become prostitutes for income reasons. In addition young females tend to have relationships with older and sometimes married men in exchange for food, money, shelter and warm clothes. They are also promised jobs and promotions at work in exchange for sex.

Younger women and children are most often victims of rape and that increases the risk of them being infected with HIV. In addition women are at risk of being infected with HIV at a younger age than younger men because they on average have partners five years older than themselves and these partners are more likely to be already HIV infected. Young women think that rich older men are an avenue to a better life. One should not forget that some young women become prostitutes because of material wealth not because the circumstances force them to do that. Age on the other hand puts young people at risk, in terms of inexperience and inability to negotiate the terms of relationships specifically the use of condom. In general, while both men and women are

vulnerable, adolescent women represent one of the most vulnerable population groups in relation to HIV/AIDS.

1.3 Objectives of the study

The aim of this project is to use longitudinal data analysis techniques to study the evolution of CD4+ count in patients on HAART in rural and urban KwaZulu Natal. We will assess whether the evolution of CD4+ count for individuals on HAART is dependent on other factors associated with the individual. The predictor or covariates that are going to be modelled are age, sex, site, weight and log viral load. In observational studies, subjects may be very heterogeneous at baseline such that longitudinal changes need to be studied after correction for potential confounders such as age, sex, geographical location and others. It has been noted that in any population, there is considerable heterogeneity in the individual rate and magnitude of CD4+ cell reconstitution (Battegay et al., 2006). Thus the specific objectives of the study are:

- To determine if the patients' weight, baseline age, sex, viral load and clinic site, influences the rate of change in CD4+ count over time
- To construct a longitudinal data analysis model for CD4+ count in patients initiated on HAART
- Account for within and between individual variability in the evolution of CD4+ count post HAART initiation
- Estimate the predictive effects of measured covariates using empirical Bayes methods
- Discuss the problem of missing data and future studies

It should be noted that for modelling purposes viral load was $\log_{10}$ transformed since the viral load is very right skewed. The analysis will be done using the statistical software called SAS (version 9.1.; SAS Institute Inc., Cary, NC, USA).

Chapter 2

Exploratory data analysis

2.1 Introduction

In this chapter a detailed exploratory data analysis to the CD4+ count data for the CAT project is carried out. The aim of this process is to understand the data structure and determine the relevant modelling approaches suitable for it.

2.2 Baseline characteristics

First we start with understanding the baseline characteristics of individuals enrolled in the study. There were 1176 patients aged 14-69 were enrolled, 409 (34.8%) from the eThekwini site and 767 (65.2%) from the Vulindlela site. Out of the 1176, 365 (31.0%) were males and 811 (69.0%) were females. All patients had a mean weight of 60.4 kg at enrolment or baseline. Table 2.1 shows a cross distribution of patients according to sex and age for each site. The mean age and weight in each site and sex are tabulated in Table 2.2. Tests of no association or relationship between site and the variables namely gender and age group was performed using the Chi-square test of independence of factors. The tests were performed at 5% level of significance.

Table 2.1: Baseline characteristics

Characteristic	Vulindlela (n=767)	eThekwini (n=409)	p-value
Sex			
Male	231 (30.1%)	134 (32.8%)	
Female	536 (69.9%)	275 (67.2%)	0.3503
Age(years)			
≤ 24	74 (9.7%)	40 (9.8%)	
25-30	228 (29.7%)	108 (26.4%)	
31-36	218 (28.4%)	130 (31.8%)	
37-42	125 (16.3%)	64 (15.7%)	
≥ 43	117 (15.3%)	62 (15.2%)	0.714
missing	5 (0.7%)	5(1.2%)	

Table 2.1 shows that the percentage distribution of males and females across the two sites is almost the same. Also the percentage distribution of age groups across the sites is almost the same. The p-values in Table 2.1 are not statistically significant, and therefore we fail to reject the null hypothesis that there is no association between site and the variables gender and age group. This observation is evident from the percentage distribution of these variables across sites in Table 2.1.

Table 2.2: Distribution of patient's baseline characteristics

Characteristic	Vulindlela	eThekwini
Age(years),mean (std)		
Sex		
Male	36 (9)	36 (9)
Female	33 (8)	33 (8)
Weight(kg),mean (std)		
Sex		
Male	58.7 (10.4)	61.1 (9.2)
Female	59.8 (13.6)	62.4 (14.0)

Table 2.2 shows that women on average are younger than men. The difference in age for males and females from eThekwini and Vulindlela is statistically significant with an independent sample t-test p-value of 0.0031 and 0.0023 respectively. The eThekwini site has higher mean weight at baseline than Vulindlela for both males and females. However, the mean weight for males and

females within each site was almost the same.

2.3 Distributional properties of CD4+ count

A histogram plot of CD4+ count data is presented in Figure 2.1. The figure shows that the distribution is skewed to the right hence may not satisfy the normality assumption. To normalize the data a square root transformation was carried out to the data and the histogram re-plotted as shown in Figure 2.2. The skewness for the histogram in Figure 2.1 and Figure 2.2 was 1.15 and 0.04 respectively. It might be more plausible to use logarithmic transformation to CD4+ counts but we will use the commonly used square root transformation just like in many studies. The transformed square root distribution now looks much more bell shaped and hence the normality assumption can hold on the square root scale.

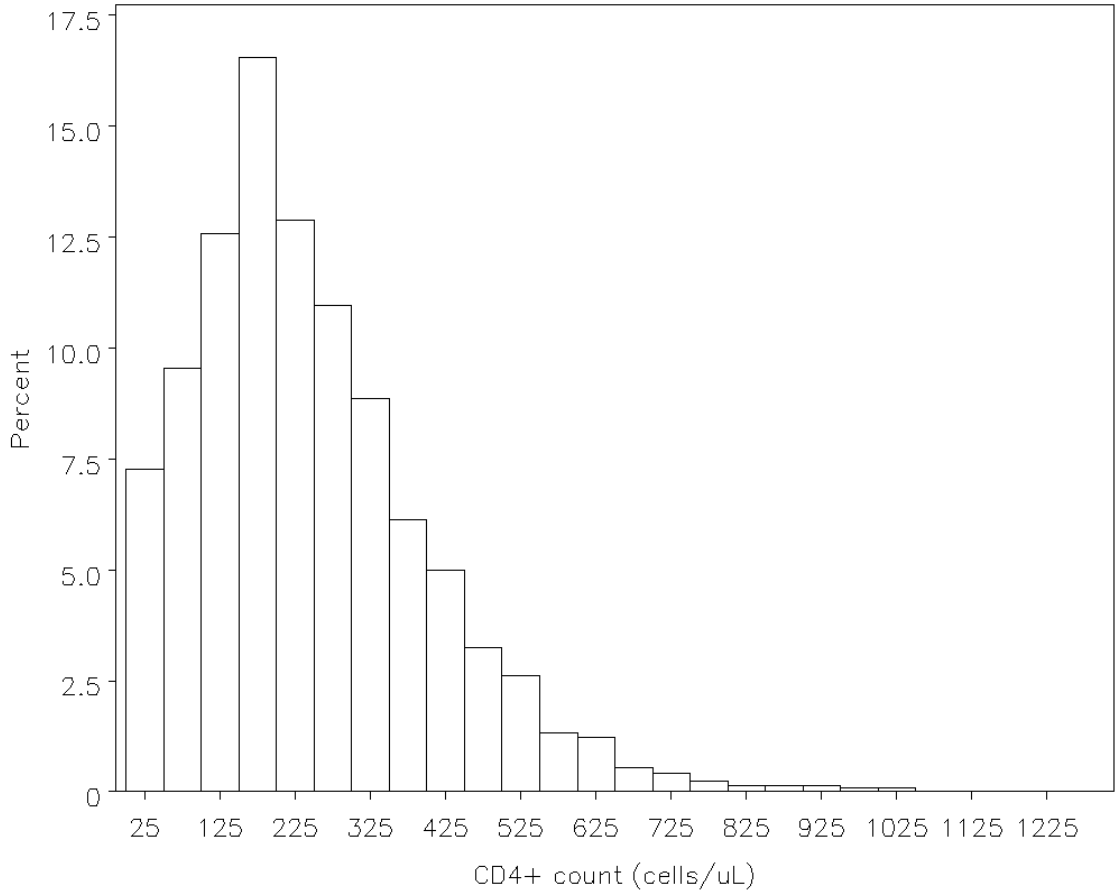

Figure 2.1 Histogram for CD4+ count

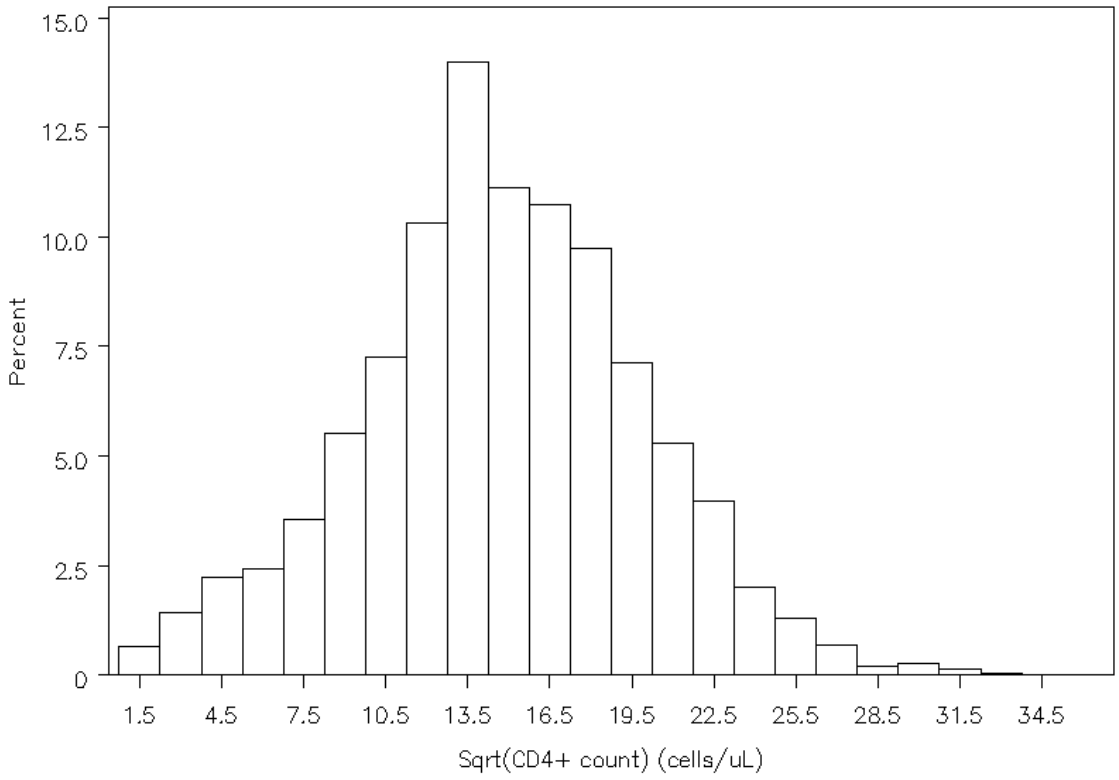

Figure 2.2 Histogram for square root CD4+ count

Both actual CD4+ count and the square root transformed values were used in the exploratory data analysis. The data set was unbalanced in the sense that the number of repeated observations per individual was not the same for all patients but the measurements for all subjects were taken at fixed time points of six monthly visits. The maximum number of observations per subject is 5.

The mean CD4+ count at baseline for the eThekwini and Vulindlela site were 105 and 106 cells/μL respectively. Further exploration of the data shows that women at both Vulindlela and eThekwini sites started with mean CD4+ count of 108 cells/μL respectively. However, men from both sites started with lower CD4+ count as compared to women. Men from Vulindlela and EThekwini started with mean CD4+ count of 100 and 99 cells/μL respectively.

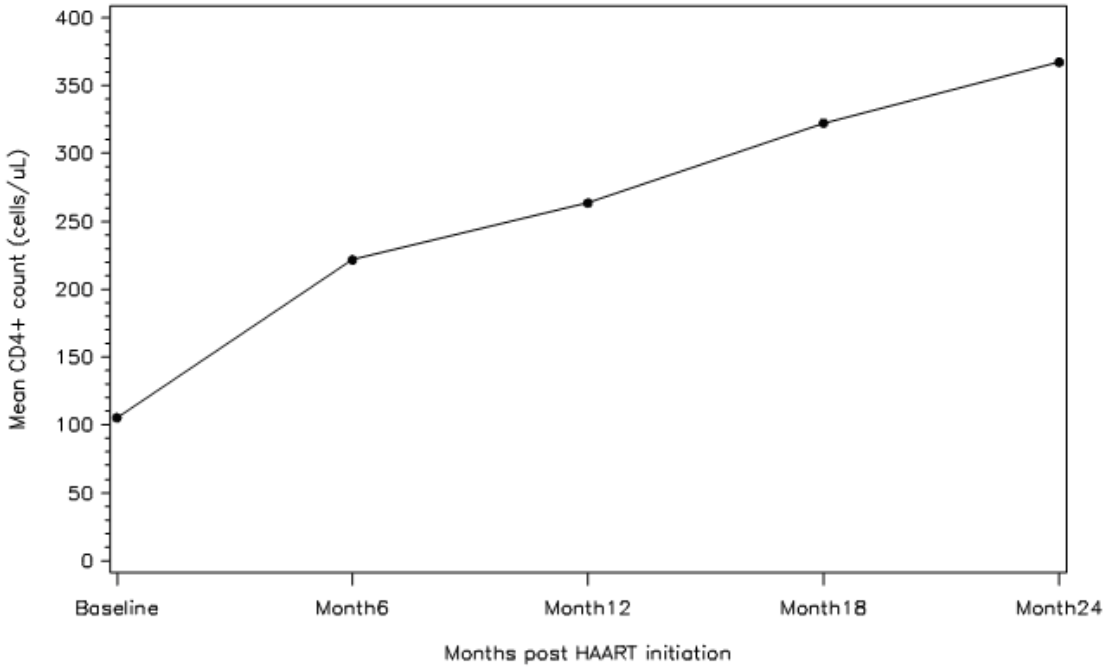

Figure 2.3 Mean CD4+ count for combined data from eThekwini and Vulindlela

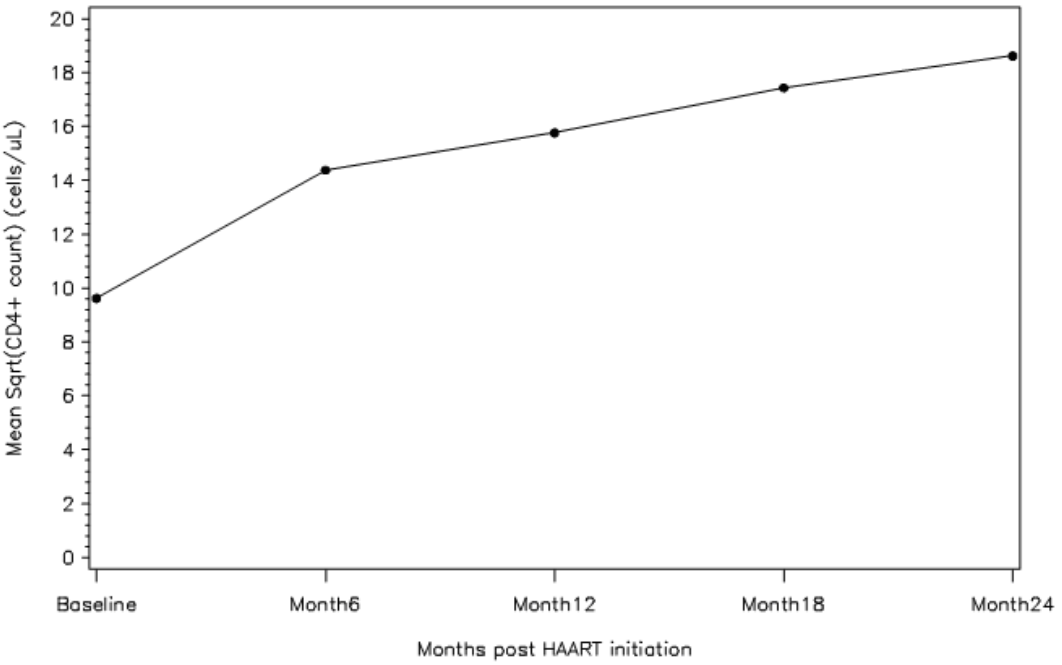

Figure 2.4 Mean square root CD4+ count for all patients from both sites

From Figure 2.3 and 2.4 it is evident that the overall mean CD4+ count increases with time. The CD4+ count tend to increase rapidly following the initiation of antiretroviral drug therapy which is a reflection of the extent of suppression of viral replication, but it should be noted these plots are mean plots which can possibly be different from individual plots because they may show some patients responding better than others. This initial increase relies on a reduction in T-cell activation and primarily consists of a release of memory CD4+ cells trapped in the lymphoid tissue (Bucy et al., 1999). This observation is in line with what is expected biologically that CD4+ count should increase rapidly following antiretroviral drug therapy (Abdool Karim and Abdool Karim, 2005).

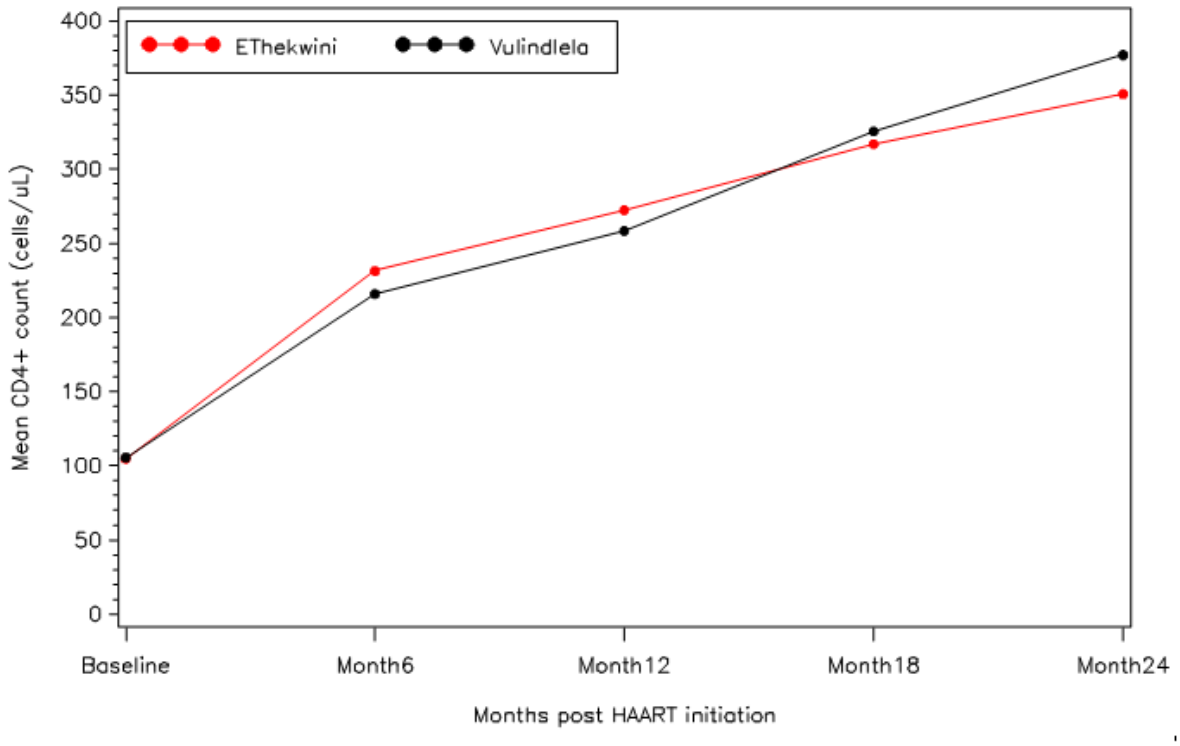

Figure 2.5 Mean CD4+ count for eThekwini and Vulindlela

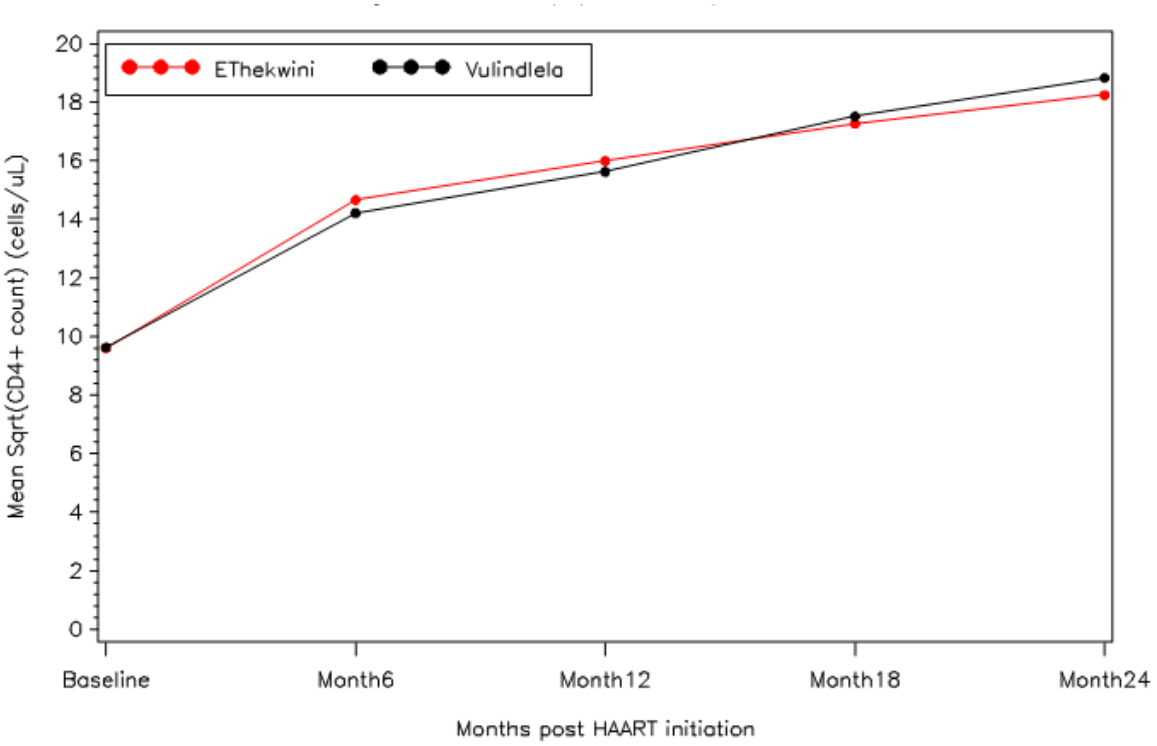

Figure 2.6 Mean square root CD4+ count for both sites

Figure 2.5 and 2.6 are raw CD4+ count and square root CD4+ count for each site. One can see that the mean CD4+ count for the two sites behave differently from baseline to month 24. At baseline the means are almost the same. As we move from baseline to just before month 15, the means for eThekwini site are greater than that of Vulindlela. However, just after month 15 the means for eThekwini site drop. Some patients were terminated as early as month 6 and therefore did not contribute any data thereafter. It is possible that issues related to drop out may be the cause of this difference in behaviour in the site specific curves. This is a manifestation of the complexity introduced by missing data in carrying out comparisons between groups.

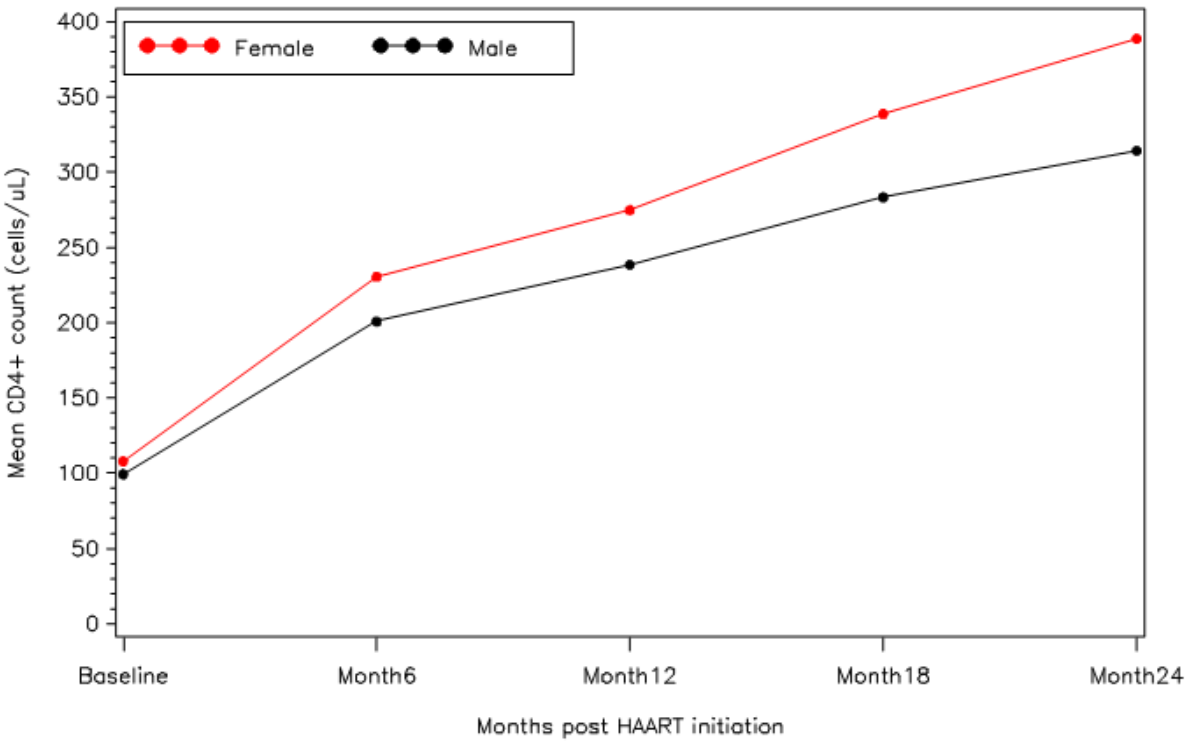

Figure 2.7 Mean CD4+ count for males and female using combined data

From Figure 2.7 one can see that the mean CD4+ count for women is greater than that of men from baseline to month 24. Relative to their counterparts, both HIV negative and HIV positive women tend to have higher CD4+ count (Tollerud et al., 1989). This difference can be due to adherence problems or due to biomedical factors. However, Nattrass (2008) reported that male South African HAART patients, whether in the public or private sectors, have on average lower CD4+ count than women when starting HAART.

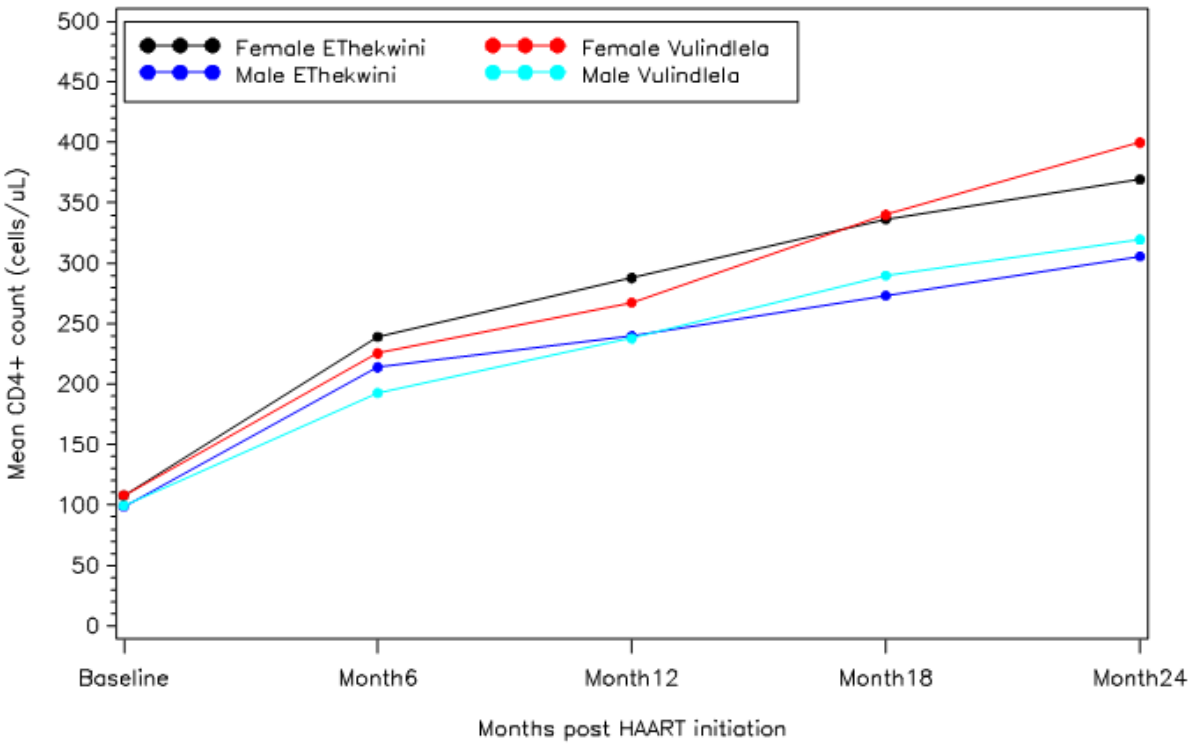

Figure 2.8 Mean CD4+ count for sex across sites

Figure 2.8 gives site specific mean CD4+ count for males and female over time in one plot. One can notice that from baseline to around month 12, the mean CD4+ count for males in the eThekwini site is greater than that for males in Vulindlela site. For females we observe the same trend around month 18.

In Figure 2.9 the mean CD4+ count for different age groups increases with time, but one can see that the age group $\leq$ 24 years is doing better than the other age groups. Moreover, as age increases CD4+ count decreases as one can see for the relative positions of age specific curves in Figure 2.9. Older groups have lowest mean CD4+ count over time as compared to younger groups. This might be attributed to the ability to re-produce CD4+ cells in younger than in older individuals. Younger patients might have higher capacity to re-produce at higher rate than older patients. The relationship between age and the immunological response supports the concept of an age related decline in thymic function and probably other regenerative mechanisms such as the peripheral expansion of CD4 T lymphocytes (Douek et al,. 1998).

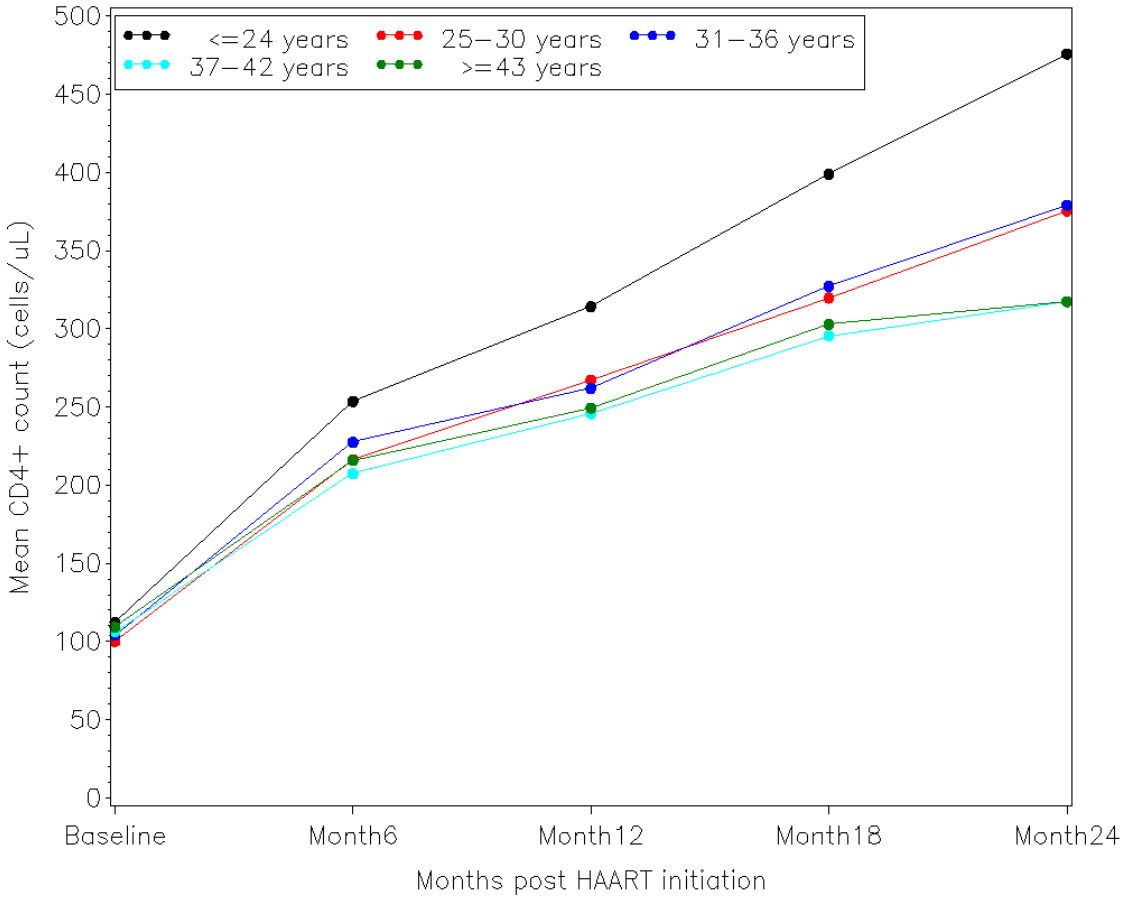

Figure 2.9 Mean CD4+ count over time for different age groups (years)

2.4 Profile plots for a random sample of patients from each site

A random sample of 50 patients from each site was selected and a plot of CD4+ count over time constructed for the 50 patients on the same graph. The aim of such a plot was to assess if there is an indication of subject to subject variability in the evolution of CD4+ count as well as within subject variability over time. A similar plot was constructed on a square root CD4+ count scale for both sites. It is important to note the presence of incomplete profiles in the data. Males and females of different age groups had an equal chance of being randomly selected.

Figure 2.10 shows evidence of individual to individual variability as well as within individual variability in the evolution of CD4+ count. The presence of incomplete profiles for patients who did not reach month 24 is also evident. Most individual plots suggest a steep increase between the baseline and month 6 measurements. These are features that will be incorporated in the analysis

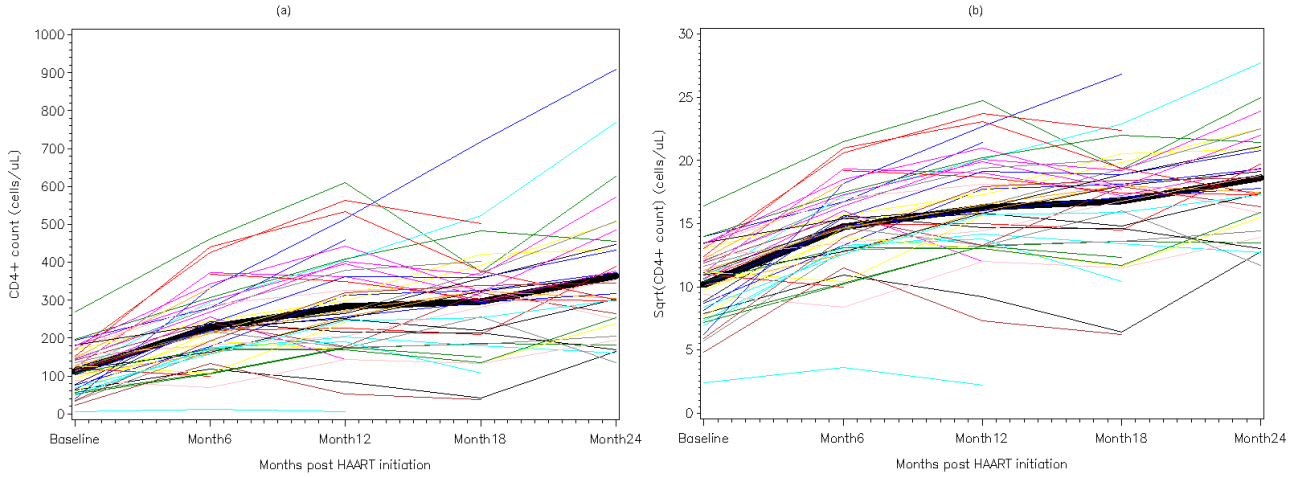

Figure 2.10 CD4+ count and square root CD4+ count for a random sample from the eThekwini site

in order to better understand the relationship of CD4+ count and the measured covariates. The two plots on original and square root scale in Figure 2.10 portray the same qualitative feature.

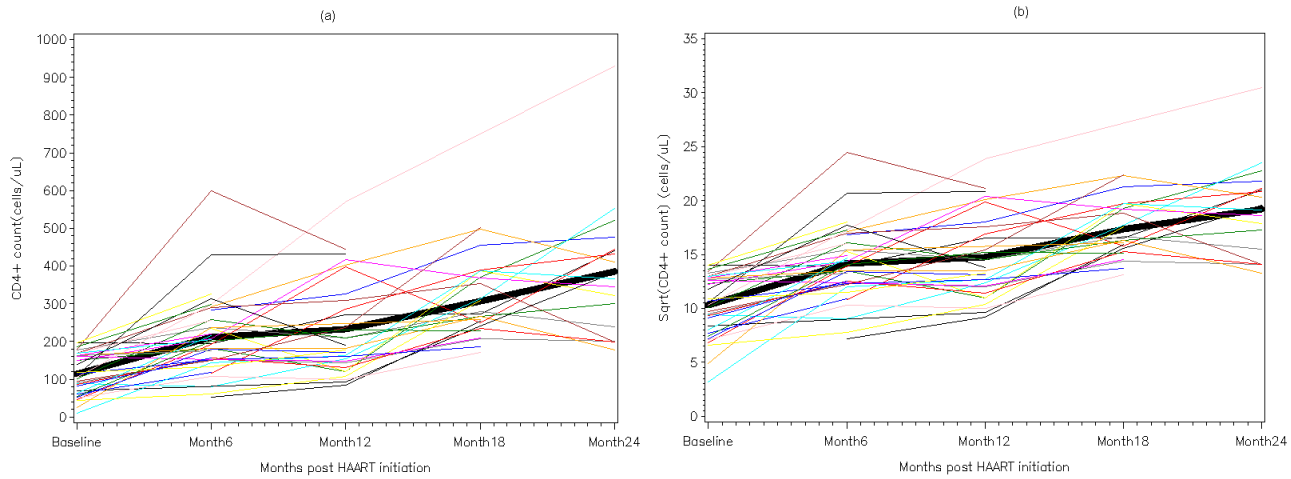

Figure 2.11 CD4+ count and square root CD4+ count for a random sample from the Vulindlela site

Figure 2.11 for Vulindlela site show similar features as those observed in the eThekwini site.

2.5 Scatter plots for CD4+ count against covariates

Several relationships between the response variable versus covariates were investigated. Covariates such as weight and log viral load are plotted against CD4+ count on the square root and original scale to establish if there is any relationship between these measured predictor variables. These plots were done at different visits and also overall visits to see if they follow the same trend.

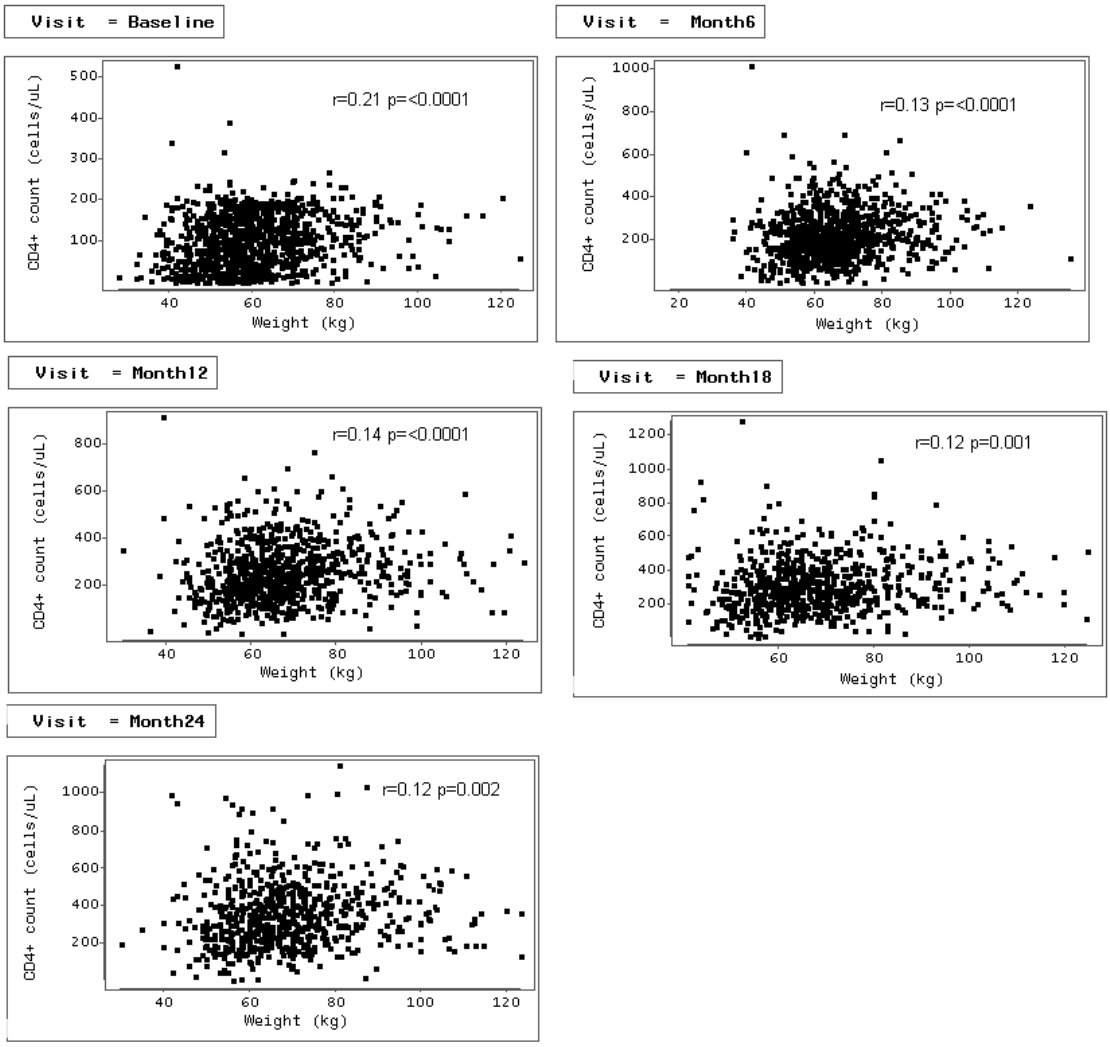

Figure 2.12 Scatter plots for CD4+ count vs. weight at different visits for combined data

Figure 2.12 indicates or suggests a positive correlation between CD4+ count and weight at all the five measurement occasions (0, 6, 12, 18 and 24 months) with possibly varying degrees of correlation. But at this point we cannot conclude whether this is a statistically significant result or not until a formal significance analysis is carried out in chapter 4.

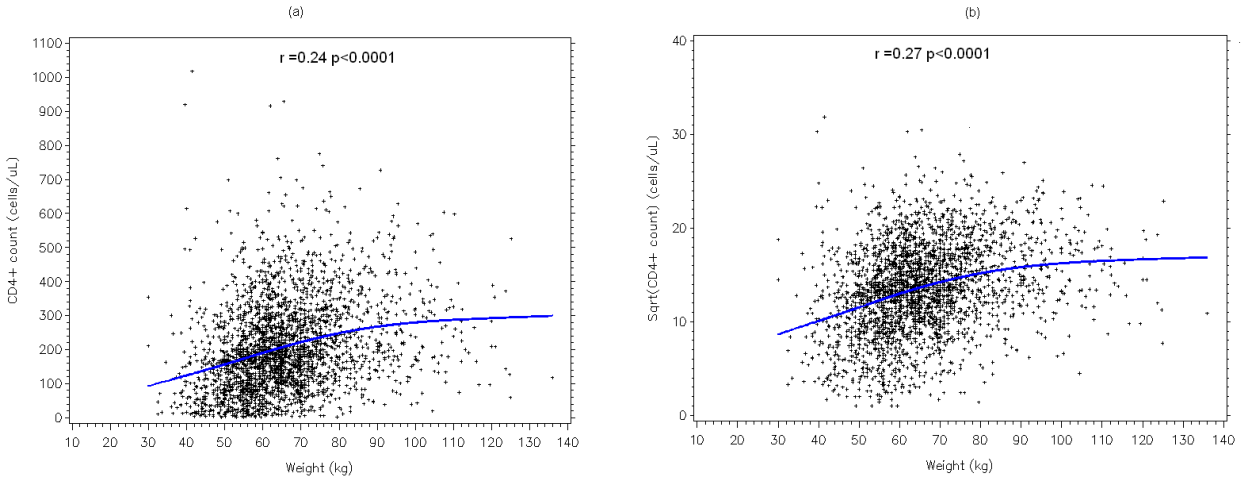

Figure 2.13 Scatter plots for CD4+ count vs. weight for combined data

Figure 2.13 shows a scatter plot of CD4+ count on the original scale (Figure 2.13(a)) and square root scale (Figure 2.13(b)) using the entire data with a loess curve fitted across. Both Figure 2.13(a) and 2.13(b) suggest a positive correlation between CD4+ count and weight.

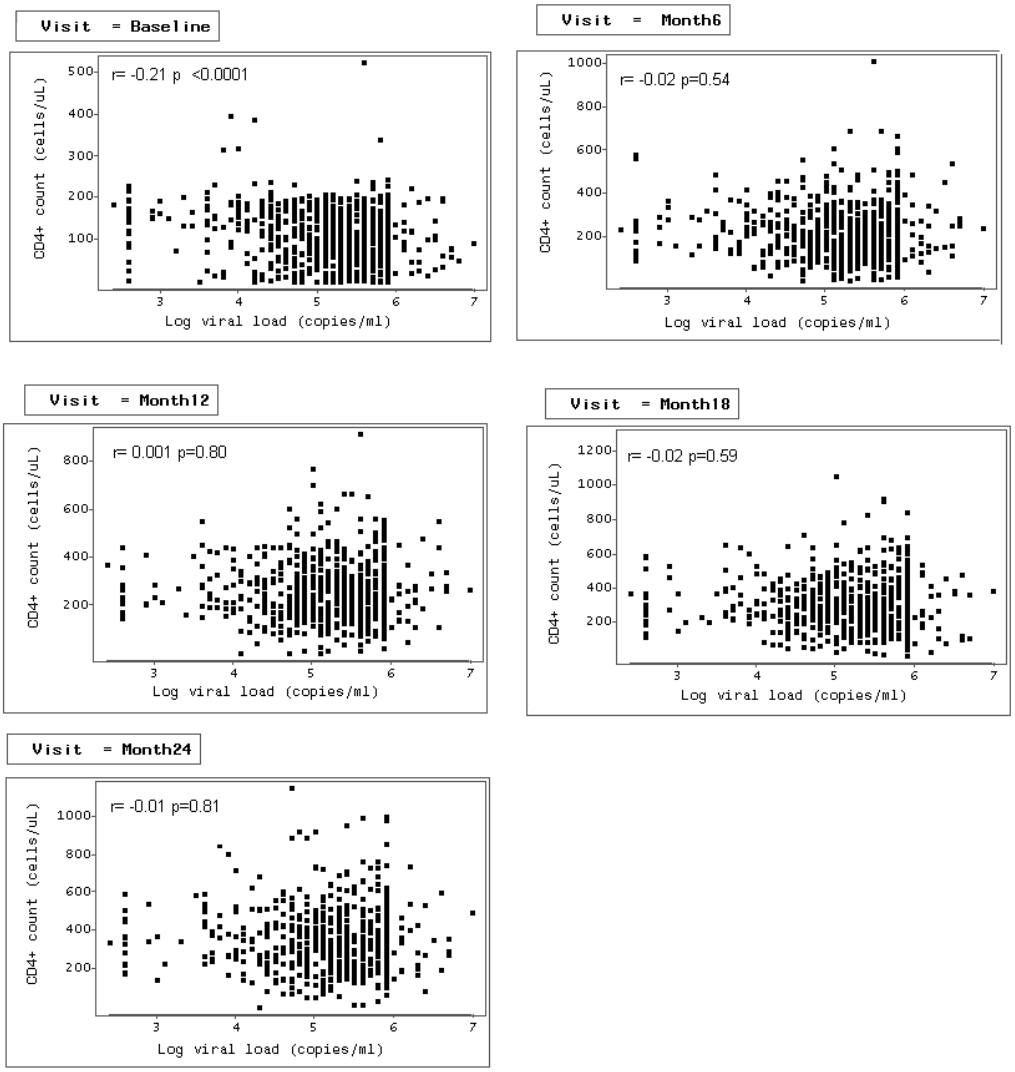

Figure 2.14 Scatter plots for CD4+ count vs. log viral load at different visits for combined data

Figure 2.14 shows a plot of CD4+ count at all visits versus log viral load at baseline. There is a statistical evidence of a weak negative correlation between CD4+ count and baseline log viral load except for month 12 measurement.

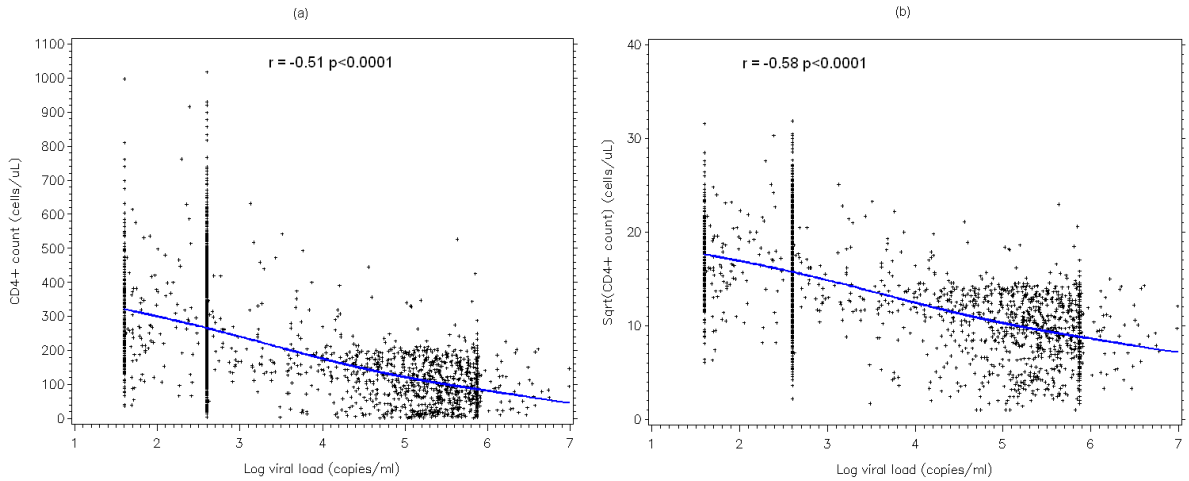

Figure 2.15 Scatter plots for CD4+ count vs. log viral

Figure 2.15(a) and (b) is a scatter plot of CD4+ count and square root CD4+ count versus log viral load using all the data are presented respectively. The plots suggest a strong negative correlation between the two markers as expected.

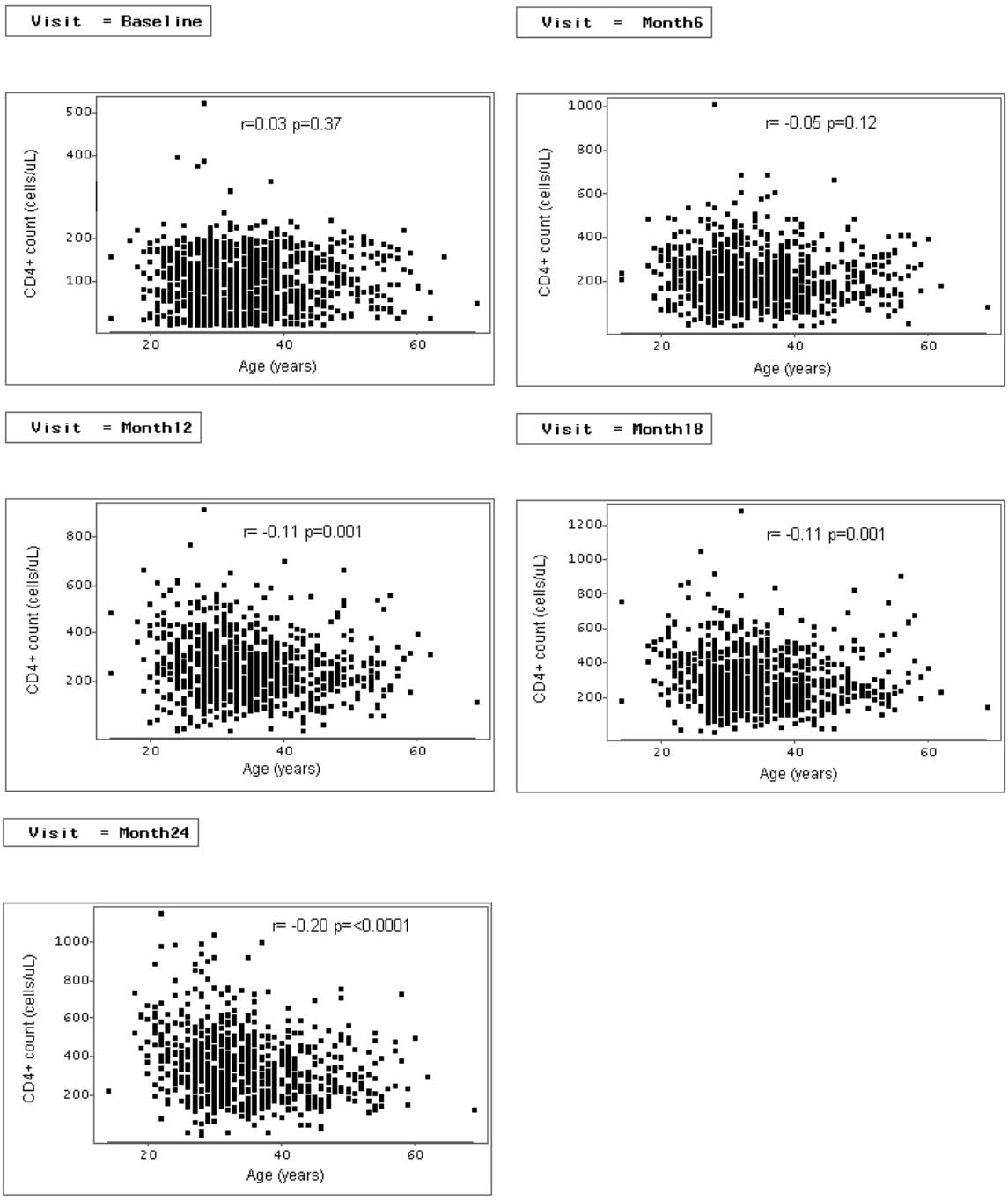

Figure 2.16 Scatter plots for CD4+ count vs. age at different visits for combined data

Figure 2.16 shows a plot of CD4+ count versus baseline age for all five different visits. There is a negative correlation between CD4+ count and age at different visits except for baseline.

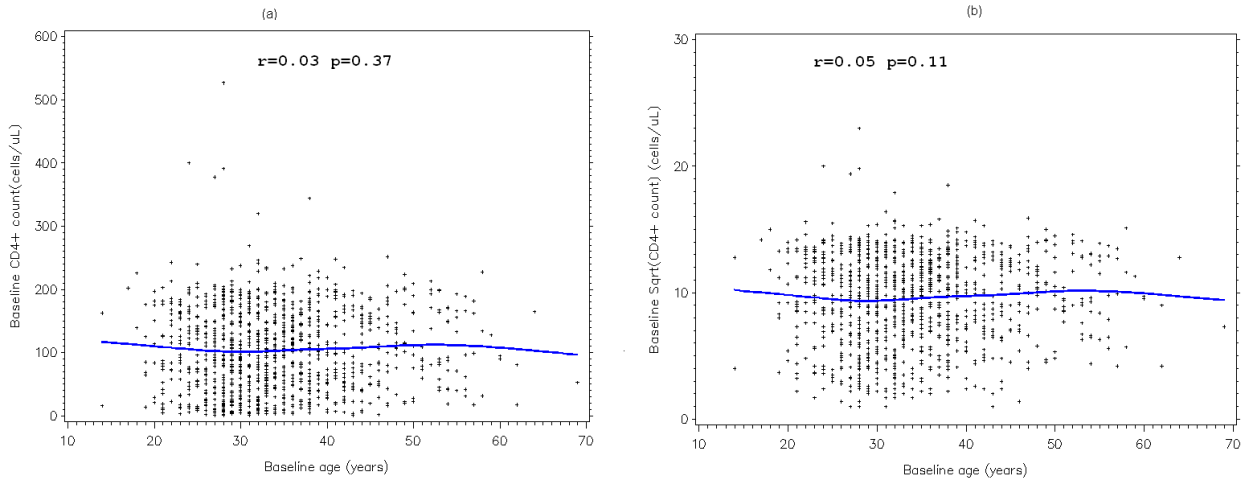

Figure 2.17 Scatter plots for CD4+ count vs. age

Figure 2.17 shows a scatter plot of baseline CD4+ count versus baseline age. The plot shows a weak positive correlation between the two variables.

2.5.1 Scatter plot and correlation matrix

The scatter plot and correlation matrix can be used for exploring the correlation between the repeated measurements. For modelling purposes this information is necessary in order to be able to capture the correct correlation structure between observations over time.

The scatter plot matrix in Figure 2.18 shows a positive correlation between any two repeated measures. However there is a clear emerging pattern of correlations as expected. Two adjacent measurements are more correlated than measurements which are distant apart and that is confirmed by Figure 2.19. The highest correlation was observed between month 18 and month 24. One of the aims of the current project will be to find a simple less parametric correlation structure that best describes the data. This problem is addressed in detail under chapter 4.

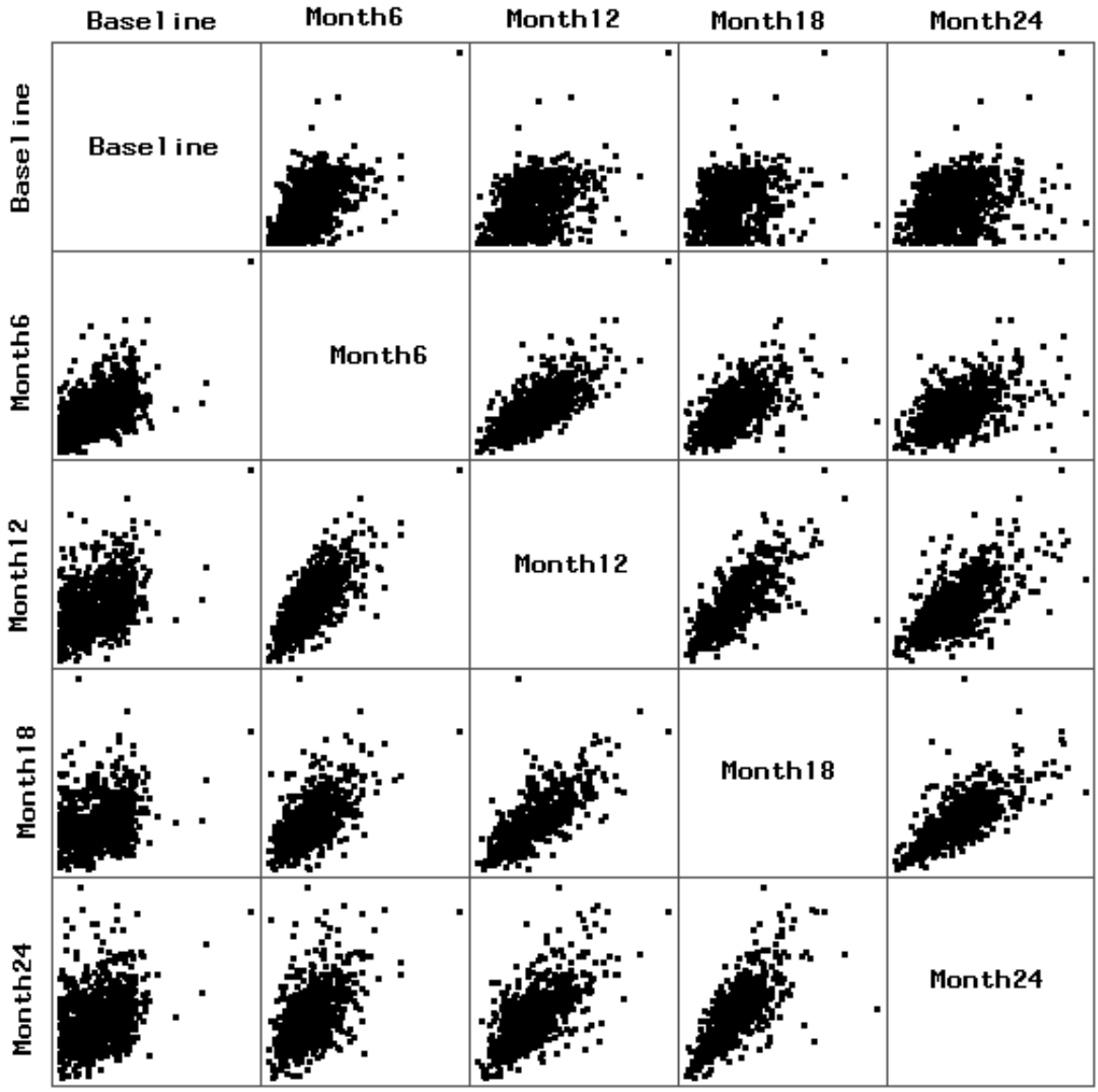

Figure 2.18 Scatter plot matrix

Correlation Matrix					
	Baseline	Month6	Month12	Month18	Month24
Baseline	1.0000	0.5472	0.4120	0.3057	0.2782
Month6	0.5472	1.0000	0.7653	0.6578	0.5756
Month12	0.4120	0.7653	1.0000	0.7658	0.6915
Month18	0.3057	0.6578	0.7658	1.0000	0.7721
Month24	0.2782	0.5756	0.6915	0.7721	1.0000

Figure 2.19 Correlation matrix

2.6 Testing equality of mean CD4+ count for different variables

First, a non-parametric Wilcoxon Mann-Whitney test was used to compare mean CD4+ count between two sites over the five measurement occasions. Furthermore, the test was used to compare mean CD4+ count of males and females. The results of this test at α=0.05 are displayed in Table 2.3 to 2.8. Mean CD4+ count are significantly different between the two sites at month 6 and 24 with p-values 0.0259 and 0.0385 respectively.

Table 2.3: Mean CD4+ count by site

Visit	eThekwini	Vulindlela	p-value
Baseline	105	106	0.9263
Month 6	231	216	0.0259
Month 12	272	258	0.1363
Month 18	317	326	0.4842
Month 24	351	377	0.0385

The p-values in the Table 2.4 show that mean CD4+ count for males and females are statistically different at all visits with females having higher mean CD4+ over all the five visits. This analysis was based on combined data over the two sites.

Table 2.4: Both sites-Mean CD4+ count by sex

Visit	Male	Female	p-value
Baseline	99	108	0.0273
Month 6	201	231	<0.0001
Month 12	239	275	<0.0001
Month 18	283	339	<0.0001
Month 24	314	389	<0.0001

Table 2.5: eThekwini site: Mean CD4+ count by sex

Visit	Male	Female	p-value
Baseline	99	108	0.0916
Month 6	214	239	0.0112
Month 12	240	288	0.0001
Month 18	273	337	0.0002
Month 24	306	370	0.0008

Table 2.5 and 2.6 represents similar results as those in Table 2.4 except that they are now site specific. Again the p-values in Table 2.5 shows that mean CD4+ count for males and females in eThekwini are statistically significantly different from month 6 to month 24 with females having the highest CD4+ count at all visits.

Table 2.6: Vulindlela site: Mean CD4+ count by sex

Visit	Male	Female	p-value
Baseline	100	108	0.1390
Month 6	193	226	0.0001
Month 12	238	267	0.0039
Month 18	290	340	0.0001
Month 24	320	400	<0.0001

Table 2.7: Mean CD4+ count for males between sites

Visit	Vulindlela	eThekwini	p-value
Baseline	100	99	0.6871
Month 6	193	214	0.1502
Month 12	238	240	0.7707
Month 18	290	273	0.4491
Month 24	320	306	0.6330

Table 2.7 shows that the mean CD4+ count for males between the two sites at all visits are not significantly different (p-values >0.05).

Table 2.8: Mean CD4+ count for females between sites

Visit	Vulindlela	eThekwini	p-value
Baseline	108	107	0.8044
Month 6	226	239	0.0747
Month 12	267	288	0.0368
Month 18	340	337	0.8333
Month 24	400	370	0.0381

Table 2.8 shows that mean CD4+ count for females between the two sites at different visit occasions are not statistically different except for month 12 and month 24 (p-value <0.05). The results in Tables 2.7 and 2.8 tells us that the improvement of CD4+ count is the same whether you are a male from urban or rural area. However, for females there seems to be significant site difference at months 12 and 24. One reason that may explain this is adherence to ARV drugs and drop out effects, but this needs further investigation. It is possible that individual(s) who dropped out between month 6 and 12 in Vulindlela were those with high CD4+ count hence biasing the mean CD4+ count in month 12 downwards, and likewise for individual(s) dropping between months 18 and 24 in eThekwini. But Table 2.4, 2.5 and 2.6 shows that females have on average higher CD4+ count than males at all visits post baseline.

2.7 Sample variogram and autocorrelation

Sources of variability in any data set can be determined by analyzing ordinary least squares (OLS) residuals with the semi-variogram. Alternatively one can describe the degree of association among repeated measurements using the full variogram instead of the semi-variogram (Diggle et al., 1994, 2002). In longitudinal sample data the counterpart of the variogram is called the sample vari-

ogram. There are three error sources that can be found in an individual's data sequence namely: subject-specific random effects (b_i), serial correlation ($\varepsilon_{2(i)}$) and measurement error ($\varepsilon_{1(i)}$), where here i denotes a typical individual in the sample.

Random effects (a measure of between-subject variability) reflect how much subject-specific profiles deviate from the average profile. When units are sampled at random from a population, various aspects of their behaviour may show stochastic variation between units. Perhaps the simplest example of this is when the general level of the response profile varies between units, that is, some units are intrinsically high responders, others low responders (Diggle et al., 1994, 2002). The response here is in terms of the response to HAART treatment. Serial correlation (a measure of within-subject variability) usually is a decreasing function of the time separation between measurements as demonstrated by the sample correlation matrix in Figure 2.19.

This type of stochastic variation results in a correlation between pairs of measurements on the same unit which depends on time separation between the pair of measurements (Diggle et al., 1994, 2002). This correlation becomes weaker as the time separation increases. Measurement error means that there may be a certain level of variation in the measurement itself. Diggle et al. (1994, 2002) came up with the following example to illustrate measurement error: Two samples taken simultaneously from a cow would have different measured protein contents, because the measurement process involves an assay technique which itself introduces a component of random variation. So, one need to establish which one amongst these three is the main source of variation namely: random effects, serial correlation and measurement error. Let Y_{ij} denote a repeated measurement observation taken at time t_{ij} where $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, n_i$. Thus the set of repeated observations from unit i are Y_i=$(Y_{i1}, \ldots, Y_{in_i})$. For balanced data t_{ij}=t_j and n_i=n. Let the mean as a function of time be $E(Y_{ij})$=$\mu(t_{ij})$. Based on mean function $\mu(t_{ij})$ the residuals

$$r_{ij} = y_{ij} - \mu(t_{ij}) \tag{2.1}$$

can be obtained and are assumed to follow the model

$$r_i = Z_i b_i + \varepsilon_{1(i)} + \varepsilon_{2(i)} \tag{2.2}$$

where the components in (2.2) respectively account for between subject variability, measurement error and serial correlation. The semi-variogram assumes constant variance, which implies that the only random effects in the model will at most be due to random intercept effects hence $Z_i = (1, 1, \ldots, 1)'$. However, some of the limitations of the semi-variogram have been relaxed for example in the work by Verbeke et al. (1998) and Serroyen et al. (2009). We will denote the variance of the random intercepts by v^2. The covariance matrix is then of the form

$$V_i = Var(Y_i) = Var(r_i) = v^2 Z_i Z_i' + \sigma^2 I_{ni} + \tau^2 H_i \tag{2.3}$$

where H_i is a matrix with elements $g(|t_{ij} - t_{ik}|)$. Thus residuals r_{ij} have constant variance $v^2 + \sigma^2 + \tau^2$. It follows that the correlation between any two residuals r_{ij} and r_{ik} from the same subject i is given by

$$\rho(|t_{ij} - t_{ik}|) = \frac{v^2 + \tau^2 g(|t_{ij} - t_{ik}|)}{v^2 + \sigma^2 + \tau^2}. \tag{2.4}$$

One can easily show that for i=1, . . . ,N and $j \neq k$,

$$\frac{1}{2} E(r_{ij} - r_{ik})^2 = \sigma^2 + \tau^2 (1 - g(|t_{ij} - t_{ik}|)) = v(u_{ijk}). \tag{2.5}$$

The function $v(u)$ is called the semi-variogram, and it only depends on the time points t_{ij} through the time lags $u_{ijk} = |t_{ij} - t_{ik}|$. This means that decreasing serial correlation functions $g(.)$ yields increasing semi-variogram $v(u)$, with $v(0) = \sigma^2$, which converges to $\sigma^2 + \tau^2$ as u grows to infinity. The sample-variogram consists of a smooth curve fitted to a scatter plot. The points along the x-axis correspond to the lags $u_{ijk} = |t_{ij} - t_{ik}|$, the time points for the cross-sectional unit i, and the points on the y-axis are

$$v_{ijk} = \frac{1}{2}(r_{ij} - r_{ik})^2 \tag{2.6}$$

for j<k, where r_{ij} is the OLS residual for cross-sectional unit i at time period t_{ij}. The total variability or process variance is calculated as

$$\hat{\sigma}^2 = \sum_{iljk} \frac{1}{2} \frac{(r_{ij} - r_{lk})^2}{count} \tag{2.7}$$

where $i \neq l$ and count is the total number of terms in the sum. $\hat{\sigma}^2$ measures the cross-sectional residual variability among all subjects over all time periods, but does not include the residual variability within subjects over time. A smooth loess curve was fitted on a scatter plot of square root CD4+ count versus time shown in Figure 2.20. The aim of this is to get a smooth function that attempts to capture important patterns in the data while leaving out noise.

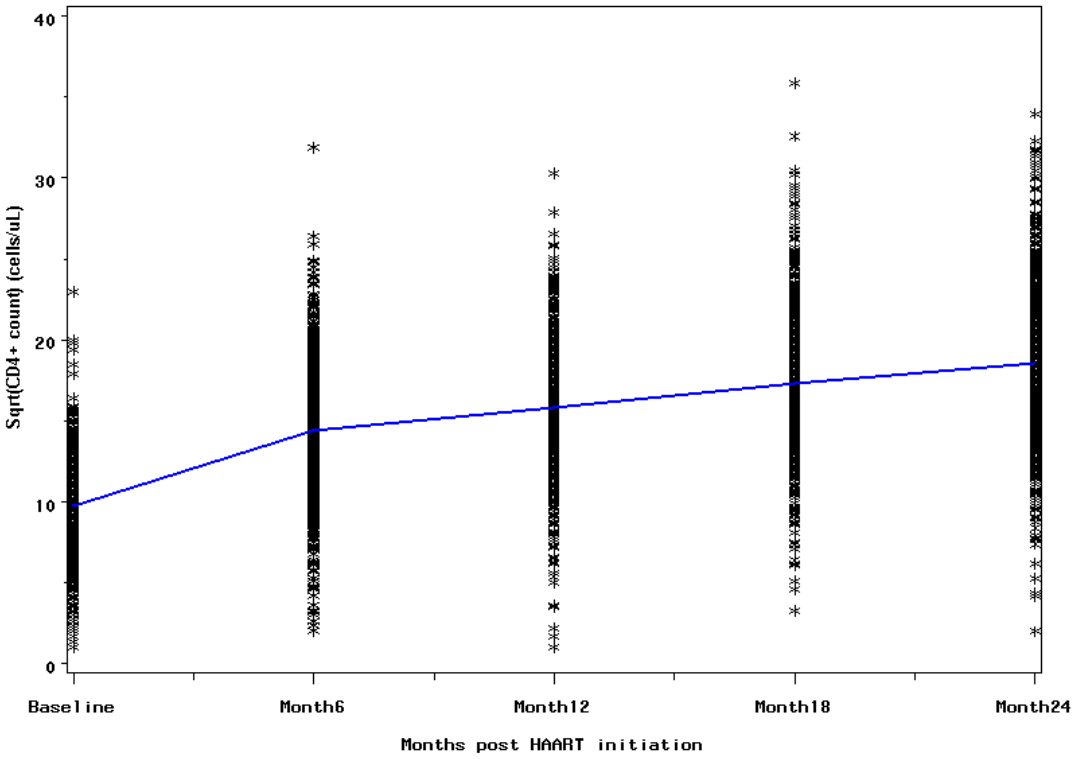

Figure 2.20 Loess smoothing

The overall smooth plot in Figure 2.20 shows a sharp increase in square root CD4 count between baseline and month 6 and then a steady increase thereafter.

In Figure 2.21 the horizontal line is the process variance. The variogram based estimate of the process variance was found to be 13.88. Since the variogram line does not begin at zero there is evidence that there is measurement error. Since the slope of the line is not zero this may imply the presence of serial correlation.

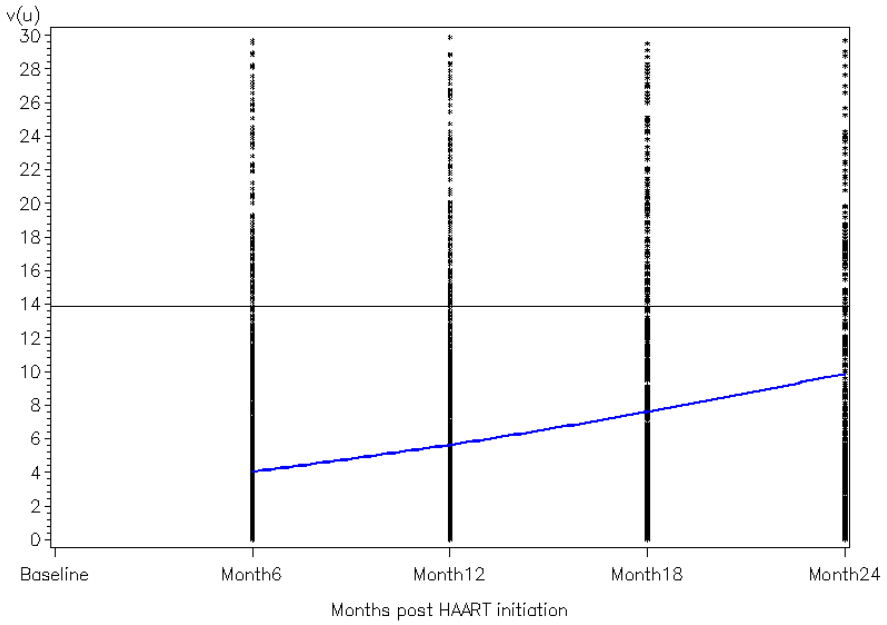

Figure 2.21 Sample-variogram

We also noted that the line does not approach the process variance, therefore suggesting the presence of individual to individual variability or random effects. These findings are similar to one reported in Diggle et al. (1994, 2002). Now that we have the variogram plot we can then get the autocorrelation plot.

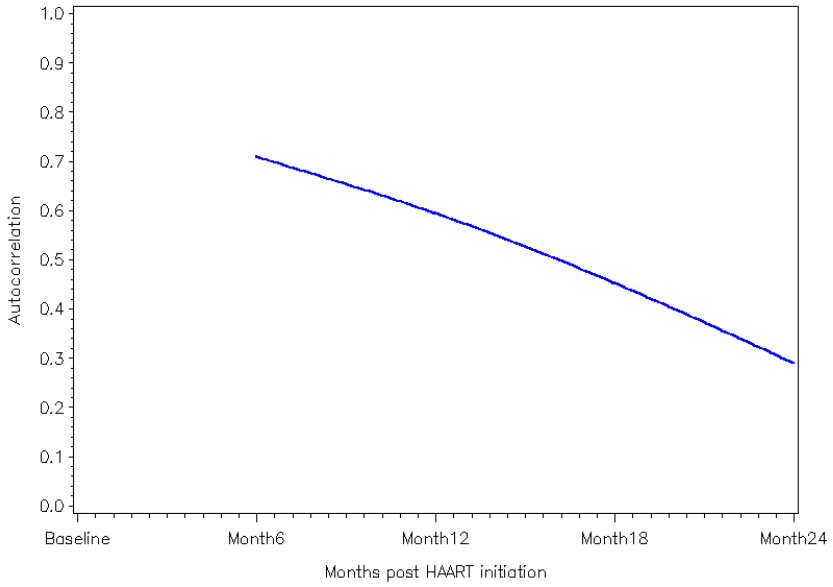

Figure 2.22 Autocorrelation plot

The autocorrelation plot is given by

$$\hat{\rho}(u) = \frac{1 - \hat{\gamma}(u)}{\hat{\sigma}^2} \tag{2.8}$$

where $\hat{\gamma}(u)$ is the average of the observed half-squared differences between residuals corresponding to that particular value of u, i.e., the average of all terms $\frac{1}{2}(r_{ij} - r_{ik})^2$ where $|t_{ij} - t_{il}| = u$ and $\hat{\sigma}^2$ is the within subject variation (Hallahan, 2003). The autocorrelation plot in Figure 2.22 shows a decreasing correlation within subject from about 0.72 to 0.28 over the range of the data. Hallahan (2003) stated that this means that a covariance structure that accounts for serial correlation is necessary.

Chapter 3

Linear Mixed Model

3.1 Introduction

In this chapter the theory of linear mixed models for longitudinal data will be developed and discussed. In particular the 2 stage fitting of the linear mixed models for longitudinal data will be the building block of the full model. One of the topics that will be discussed includes the estimation and inferences for fixed effects. The need for random effects to model extra variability will necessitate the estimation and inference for variance components. The problem of estimating the random effects including how to make inference will also be studied. The approach by Verbeke and Molenberghs (2000) is used extensively in the development of the theory.

3.2 Theory of the linear mixed model

In the entire thesis the response variable will be the square root CD4+ count. Mixed models provide a flexible and powerful tool for the analysis of data with complex covariance structure, such as longitudinal correlated data. A mixed model has two types of components, the systematic or fixed, or the mean model component and the random component. The fixed component is a sub-model representing the contribution by fixed effects and the random component represents the contribution by random effects. A fixed effect is an effect where all levels of the variable are contained in the data and the effect is universal to the entire target population. Linear mixed effects models for repeated measures data formalize the idea that an individual's pattern of responses is likely to depend on many characteristics of that individual, including some that are unobserved (Der and Everitt, 2006).

These unobserved effects are then included in the model as random variables, or equivalently called, random effects (Der and Everitt, 2006). A random effects model means that the levels of the factor

variable in the data being modelled comprise a random sample of levels in the target population. A fixed effect is considered to be a constant which we wish to estimate, but the random effect is considered as just an effect coming from a population of effects (McCulloch et al., 2008). To emphasize on this distinction we use the term *prediction* of random effects rather than estimation (McCulloch et al., 2008). Deciding whether a factor is random or fixed is not always easy and can be controversial (Littell et al., 2006). The collection of such class of linear models is referred to as linear mixed models. In linear mixed models, fixed effects are used for modelling the mean of y while random effects govern the variance-covariance structure of y (McCulloch et al., 2008).

A repeated measures model is a special case of the general linear mixed model. The distinguishing feature of the repeated measures model lies in the specification of the covariance structure of the repeated measures. The choice of fixed and random effects is not always determined by the structure of the experiment, but may depend on the information required. The model for repeated measurements from the same individual implies an underlying correlation structure between measurements on the same subject which constitute a cluster of observations. Before stating the linear mixed model, the general multivariate model is briefly discussed. The model is given by

$$Y_i = X_i\beta + \varepsilon_i \tag{3.1}$$

where Y_i=$(Y_{i1} \dots Y_{in})$ is the vector of n repeated measurements from the ith subject and $\varepsilon_i \sim \mathrm{N}(0, \Sigma)$. The distribution for Y_i is $Y_i \sim \mathrm{N}(X_i\beta, \Sigma)$. Assuming independence across individuals, β and the parameters in Σ_i can be estimated by optimizing the likelihood given by

$$L_{ML} = \prod_{i=1}^{N} \left\{ (2\pi)^{-n/2} |\Sigma_i|^{-1/2} exp\Big(-\frac{1}{2}(Y_i - X_i\beta)' \Sigma_i^{-1} (Y_i - X_i\beta) \Big) \right\}. \tag{3.2}$$

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

3.3 Linear mixed model for longitudinal data

Many longitudinal studies are designed to investigate change over time in a characteristic which is measured repeatedly for each patient (Laird and Ware, 1982). Analyses of multiple observations measured on the same individual over time are different from observations measured on different people. The advantage of a longitudinal study is its effectiveness for studying change. There are natural changes like peoples' change in height and there are also changes caused by interventions or treatment which we are often interested in. Often, we cannot fully control for the circumstances under which measurements are taken, and there may be considerable variation among individuals in the number and timing of observations (Laird and Ware, 1982). Investigators gather repeated

measures or longitudinal data in order to study change in a response variable over time as well as to relate these changes in explanatory variables over time (McCulloch et al., 2008). Responses measured repeatedly on the same unit or individual are correlated because they contain a common contribution from that unit. To estimate treatment effect, it is important to adequately model the covariance structure of the repeated measurements. Moreover, measurements on the same individual close in time tend to be more correlated than measures far apart in time. Therefore it is important to try and model the correct correlation structure and this will yield more precise estimators of interest.

In addition, modelling the true correlation structure becomes significant in the presence of missing values and when the number of observations per subject is not large. There are two types of covariates in longitudinal studies in general. There are time invariant or baseline covariates (e.g. gender) and time varying covariates (e.g. weight). The Linear Mixed Model (LMM) has become the most commonly used tool for analyzing continuous repeated measures data from a sample of individuals in agriculture, biomedical, economical, and social applications. Thus the term 'individual' will have different interpretation or meaning for different areas of application. A special case of a linear mixed model is when there are no fixed effects leading to what is called a random effects model (McCulloch et al., 2008). Here are some of the examples to show how measurements may be taken repeatedly on the same unit .

- The units may be trees in a forest. For each tree, measurements of the diameter of the tree are made at several points along the trunk of the tree. Thus, the tree is measured repeatedly over positions along the trunk.
- The units may be pregnant female rats. Each rat gives birth to a litter of pups, and birth weight of each pup is recorded. Thus, the rat is measured repeatedly over each of her pups.
- The units may be patients in a longitudinal study where measurements of biological laboratory markers such as CD4+ count and viral load are taken at every six monthly visits. Thus the patient is measured repeatedly giving rise to a cluster of observations from each patient

Repeated measurements over time are a special case of clustered data. In the first case the observations are clustered within a tree trunk and in the second case the observations are clustered within a female rat. Thus the clusters are the tree trunk and female rat respectively. In the third example observations are clustered within an individual. Longitudinal data can be collected either prospectively following subjects forward in time, or retrospectively, by extracting multiple measurements on each person form historical records (Diggle et al., 1994, 2002). However the latter approach can lead to biased information if proper validation of records is not undertaken. Time can be measured in variety of scales such as days, months, years, seasons and so on. The research

design can be experimental or observational. Often, subject-specific longitudinal profiles can well be approximated by linear regression functions. One can think of a two-stage development of such functions. First one fits a linear regression model for each subject then next fit a model to regress subject-specific regression co-efficients from stage one to known population based covariates.

In the current data we envisage that there is much variability between patients and little variability within patients since measurements taken on the same patient are correlated. There is a considerable variability across individuals due to influence of unmeasured characteristics such as genetic make-up, environmental exposures, personal habits, and so on (Diggle et al., 1994, 2002). Also the within subject CD4+ count are subject to biologic variation and measurement error, both of which can be considerable (Malone et al., 1990).

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

3.3.1 Two-staged fitting of the linear mixed model for longitudinal data

Modelling longitudinal or repeated data can be thought of as a two-stage process (Verbeke and Molenberghs, 2000). The two-stage approach is presented for purposes of insight and didactics otherwise with SAS procedure MIXED we can almost always fit the model in an integrated manner. First a linear regression model is specified for every subject separately, modelling the outcome variable as a function of time. Afterwards, in the second stage, multivariate linear models are used to relate the subject-specific regression parameters from the first-stage model to subject characteristics such as age, gender, weight, etc. The two-stage liner mixed model formulation is derived as follows. We consider time as the only covariate of interest. Thus the stage 1 model can be written as

$$Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + \varepsilon_{ij} \tag{3.3}$$

for i=1, . . . , N assuming there are N individuals and n_i observations for the ith individual measured at time t_{ij}, j=1, . . . ,n_i. Here β_{0i} and β_{1i} are the subject-specific intercept and slope respectively which can further be expanded in the stage 2 development as

$$\beta_{0i} = \beta_0 + b_{0i}$$
$$\beta_{1i} = \beta_1 + b_{1i} \tag{3.4}$$

where b_{0i} and b_{1i} are the individual specific random intercept and slope measuring the deviations from the population mean intercept and slope β_0 and β_1 respectively. Combining equation (3.3) and (3.4) we get

$$\begin{aligned} Y_{ii} &= \beta_0 + b_{0i} + (\beta_1 + b_{1i})t_{ij} + \varepsilon_{ij} \\ &= \beta_0 + \beta_1 t_{ij} + b_{0i} + b_{1i}t_{ij} + \varepsilon_{ij}. \end{aligned} \tag{3.5}$$

The first two terms in equation (3.5) constitute the fixed effects component or sub-model and the remaining portion gives the random effects components. Note that $\beta_0 + b_{0i}$ measures the average response level for the ith subject when time or visit is zero while $\beta_1 + b_{1i}$ measures change in the response over time specific to the ith subject. The inclusion of the measurement errors ε_{ij}, allows the response at any occasion to vary randomly above and below the subject-specific trajectories (Fitzmaurice et al., 2004). So the individual intercept and slope are explained by an average part and a random effect. The joint distribution of the random effects (intercept and slope) is assumed to be bivariate normal such that $(b_{0i}, b_{1i})'$ is distributed as

$$\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} g_0^2 & g_{01} \\ g_{10} & g_1^2 \end{pmatrix} \right].$$

Non-zero values of g_{01} indicate that subject specific rates of change are associated with subject specific average response levels. In matrix notation the stage 1 model is written as

$$Y_i = Z_i\beta_i + \varepsilon_i \tag{3.6}$$

with

$$Z_i = \begin{pmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ \vdots & \vdots \\ 1 & t_{ini} \end{pmatrix}.$$

The vector β_i includes all subject specific effects and Z_i is the corresponding design matrix at this stage. The vector ε_i is the error term of dimensions $N \times 1$. Here Z_i is of the form $N \times q$ and β_i is of dimensions $q \times 1$. The stage 2 model is designed to relate β_i to subject specific covariates through population based parameter β as

$$\beta_i = K_i\beta + b_i. \tag{3.7}$$

- K_i is a $(q \times p)$ matrix of known covariates
- β is a p-dimensional vector of unknown regression parameters

This model indicates that individual i's initial level is determined by the population parameter β describing average trends plus a unique contribution for that individual b_i. b_i describes how the evolution of the ith subject deviates from the average evolution in the population (Molenberghs and Verbeke, 2001). This tells us that there are unobserved factors represented by the $b_i's$ that are common to all responses for a given patient but which vary across each patient, thus inducing an inherent correlation between observations within the same individual.

Combining stages 1 and 2 using equation 3.6 and 3.7 gives us the following model:

$$Y_i = Z_i(K_i\beta + b_i) + \varepsilon_i = Z_iK_i\beta + Z_ib_i + \varepsilon_i \tag{3.8}$$

where Z_iK_i=X_i and the final model becomes

$$Y_i = X_i\beta + Z_ib_i + \varepsilon_i \tag{3.9}$$

where

- Y_i is the $n_i \times 1$ response vector for ith subject: $Y_i = (Y_{i1}, Y_{i2}, ..., Y_{in_i})'$
- Z_i is a $n_i \times q$ matrix of known covariates
- X_i is a $n_i \times p$ design matrix for the fixed effects
- β is a p-dimensional vector of subject specific regression coefficients
- b_i is q-dimensional vector of unknown random effects
- ε_i is $n_i \times 1$ error vector$\sim$ N(0, Σ_i), often $\Sigma_i = \sigma^2 I_{n_i}$
- $b_i \sim N(0, G)$

Model 3.9 assumes that we have both fixed and random effects in the model. If $\Sigma_i = \sigma^2 I_{n_i}$ where I_{n_i} denotes an identity matrix, then we call the model the 'conditional-independence model' since it implies the n_i responses on individual i are independent, conditional on b_i and β (Laird and Ware, 1982). Furthermore, $b_1, ..., b_N$ and $\varepsilon_i, ..., \varepsilon_N$ are assumed to be independent. The elements for the variance components are in the matrices G and Σ_i. In the so-called error component models, ε_i can be decomposed into two components representing both subject-specific variation and variation over time, that is serial correlation (Hallahan, 2003). Therefore $\varepsilon_i = \varepsilon_{1(i)} + \varepsilon_{2(i)}$ where $\varepsilon_{1(i)}$ is the measurement error associated with i^{th} subject and $\varepsilon_{2(i)}$ is associated with serial correlation for i^{th} subject. $\varepsilon_{1(i)} \sim N(0, \sigma^2 I_{ni})$ and $\varepsilon_{2(i)} \sim N(0, \tau^2 H_i)$.

The correlation matrix H is assumed to have its (j,k)th element given by $h_{ijk} = g(|t_{ij} - t_{ik}|)$ for some decreasing function $g(.)$ with $g(0) = 1$. This means that the correlation between $\varepsilon_{2(ij)}$ and $\varepsilon_{2(ik)}$ only depends on the time interval between the measurements y_{ij} and y_{ik}, and decreases if the length of this interval increases (Verbeke and Molenberghs, 2000). $\varepsilon_{2(i)}$ represents the belief that part of an individual's observed profile is a response to time varying stochastic processes operating within that individual (Verbeke and Molenberghs, 2000). The model given by equation 3.9 is restated more explicitly in matrix notation as

$$\begin{bmatrix} Y_{1.} \\ \vdots \\ Y_{N.} \end{bmatrix} = \begin{bmatrix} X_{1.} \\ \vdots \\ X_{N.} \end{bmatrix} \beta + \begin{bmatrix} Z_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & Z_N \end{bmatrix} \begin{bmatrix} b_{1.} \\ \vdots \\ b_{N.} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1.} \\ \vdots \\ \varepsilon_{N.} \end{bmatrix}. \tag{3.10}$$

The marginal mean and variance of Y_i is given by

$$E(Y_i) = X_i\beta \tag{3.11}$$

and

$$V(Y_i) = V(Z_i b_i + \varepsilon_i) = Z_i G Z_i' + \Sigma_i = V_i. \tag{3.12}$$

This gives us the between and within subject variation contained in the first and second components of V_i. Both G and Σ_i can be estimated by either the method of Maximum Likelihood (ML) or Restricted Maximum Likelihood(REML) estimation. But Diggle et al. (1994, 2002) report that ML estimation produces biased estimators. Assuming we are sampling from a Gaussian distribution then marginally

$$Y_i \sim N(X_i\beta, Z_i G Z_i' + \Sigma_i). \tag{3.13}$$

On the other hand Hallahan (2003) said that REML estimators are less sensitive to outliers than ML estimators. If one is interested in the fixed effects (population averages) then one should focus more on the marginal model. In this case random effects are considered nuisance parameters since the method does not use the Z matrix to predict the response Y, rather the role of random effects enter into the var(Y) under the components of variance in G.

The conditional mean and variance of Y_i is given by

$$E(Y_i \mid b_i) = X_i\beta + Z_i b_i \tag{3.14}$$

and

$$V(Y_i \mid b_i) = \Sigma_i. \tag{3.15}$$

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

$$Y_i \mid b_i \sim N(X_i\beta + Z_i b_i, \Sigma_i). \tag{3.16}$$

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

3.4 Estimation of fixed effects

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

$$L_{ML}(\theta) = \prod_{i=1}^{N} \left\{ (2\pi)^{-n_i/2} |V_i(\alpha)|^{-1/2} exp\Big(-\frac{1}{2}(Y_i - X_i\beta)' V_i^{-1}(\alpha)(Y_i - X_i\beta) \Big) \right\} \tag{3.17}$$

where α is the vector of all variance components in G and Σ_i and θ=$(\beta', \alpha')'$ is the vector of all parameters in marginal model. The log likelihood function for subject i is

$$l_i = logL_i = -\frac{n_i}{2} log 2\pi - \frac{1}{2} log|V_i| - \frac{1}{2}(y_i - X_i\beta)' V_i^{-1}(y_i - X_i\beta) \tag{3.18}$$

and

$$\frac{\partial l_i}{\partial \beta} = -X_i V_i^{-1} X_i \beta + X_i' V_i^{-1} \tag{3.19}$$

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

$$\hat{\beta}(\alpha) = \Big(\sum_{i=1}^{N} X_i' V^{-1} X_i \Big)^{-1} \sum_{i=1}^{N} X_i' V^{-1} y_i. \tag{3.20}$$

In most cases α is not known and needs to be estimated as say $\hat{\alpha}$, then V_i^{-1} should subsequently be replaced by $V_i(\hat{\alpha})^{-1}$. The two frequently used methods to estimate α are maximum likelihood and restricted maximum likelihood. The method of restricted maximum likelihood was introduced by Patterson and Thompson (1971). It was developed in order to avoid biased variance component estimates that are produced by ordinary maximum likelihood estimation. This is because maximum likelihood estimates of variance components takes no account of the degrees of freedom used in estimating fixed effects. This means that ML estimates of variance component have

a downwards bias which increases with the number of fixed effects in the model. This in turn leads to underestimates of standard errors for fixed effects, hence leading to unnecessarily more conservative confidence intervals than expected.

3.4.1 Maximum likelihood estimation (ML)

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

3.4.2 Restricted maximum likelihood estimation (REML)

To explain the concept of REML estimation more clearly consider a sample of N observations $Y_1, ..., Y_N$ from $N(\mu, \sigma^2)$. For known μ, the MLE of σ^2 is given by

$$\hat{\sigma}_u^2 = \sum_i (Y_i - \mu)^2 / N \tag{3.21}$$

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

$$\hat{\sigma}_b^2 = \sum_i (Y_i - \bar{Y})^2 / N. \tag{3.22}$$

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

$$E(\hat{\sigma}_b^2) = \frac{N-1}{N}\sigma^2. \tag{3.23}$$

The biased expectation of $\hat{\sigma}_b^2$ implies that an unbiased estimate of σ^2 should be

$$S_u^2 = \sum_i \frac{(Y_i - \bar{Y})^2}{N-1} = \frac{N}{N-1}\hat{\sigma^2}_b. \tag{3.24}$$

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

Let U= $\begin{pmatrix} Y_1 - Y_2 \\ Y_2 - Y_3 \\ \vdots \\ Y_{N-2} - Y_{N-1} \\ Y_{N-1} - Y_N \end{pmatrix} = A'Y \sim (0, \sigma^2 A'A).$

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

$$S_u^2 = \sum_i \frac{(Y_i - \bar{Y})^2}{N-1} \tag{3.25}$$

where A defines the set of $N-1$ linearly independent error contrasts of the vector $Y = (Y_1, \ldots, Y_N)'$. Therefore S_u^2 is called the REML estimate of σ^2, which is independent of A. This estimation can be extended to linear regression models. To do this consider a sample of N observations $Y_1, \ldots, Y_N$ from a linear regression model where Y $\sim$ N$(X\beta, \sigma^2 I)$. The MLE of σ^2 is

$$\hat{\sigma_b^2} = (Y - X\hat{\beta})'(Y - X\hat{\beta})/N \tag{3.26}$$

which is biased for σ^2 since

$$E(\hat{\sigma_b^2}) = \frac{N-p}{N}\sigma^2. \tag{3.27}$$

From the biased expression one can derive an unbiased estimate Mean Square Error(MSE) given by

$$\sigma_u^2 = \frac{(Y - X\hat{\beta})'(Y - X\hat{\beta})}{N-p}. \tag{3.28}$$

The MSE can also be obtained from transforming the data orthogonal to X like in the simple example above. The MSE is called the REML estimate of σ^2. So far we have done the estimation of variance for the normal population and also in the linear regression model. Now we consider a similar estimate in the case of linear mixed model. Consider models where $Y_i \sim$ N$(X\beta, V_i)$ where V_i was defined in equation (3.12) under the formulation of the linear mixed models. If we combine the subject-specific sub-models into one model we get Y$\sim$ N$(X\beta, V(\alpha))$.

Where $V(\alpha)$= $\begin{pmatrix} V_1 \ldots 0 \\ \vdots \ddots \vdots \\ 0 \ldots V_n \end{pmatrix}$

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

$$L_{REML}(\theta) = \Big| \sum_{i=1}^{N} X_i' V_i^{-1}(\alpha) X_i \Big|^{-\frac{1}{2}} L_{ML}(\theta) \tag{3.29}$$

with respect to θ, that is, with respect to α and β simultaneously. $L_{REML}(\theta)$ can also be seen as a penalized likelihood where the penalty term is given by $\Big| \sum_{i=1}^{N} X_i' V_i^{-1}(\alpha) X_i \Big|^{-\frac{1}{2}}$.

3.5 Inference for the marginal model

This section is devoted to the estimation of fixed effects, variance components and inference methods suitable for the ensuing estimates.

3.5.1 Inference for the fixed effects

One should remember that $\hat{\beta}(\alpha)$ is multivariate normal with mean β and covariance var$(\hat{\beta})$ where

$$
\begin{aligned}
E\Big[\hat{\beta}(\alpha)\Big] &= \left(\sum_{i=1}^{N} X_i' V_i^{-1} X_i\right)^{-1} \sum_{i=1}^{N} X_i' V_i^{-1} E(Y_i) \\
&= \left(\sum_{i=1}^{N} X_i' V_i^{-1} X_i\right)^{-1} \sum_{i=1}^{N} X_i' V_i^{-1} X_i \beta \\
&= \beta
\end{aligned}
$$

provided that $E(Y_i) = X_i\beta$ and assuming α is known. In order for $\hat{\beta}$ to be unbiased, it is sufficient that the mean of the response variable is correctly specified.

$$
\begin{aligned}
\text{var}(\hat{\beta}) &= \left(\sum_{i=1}^{N} X_i' V_i^{-1} X_i\right)^{-1} \left(\sum_{i=1}^{N} X_i' V_i^{-1}(\alpha) Var(Y_i) V_i^{-1}(\alpha) X_i\right) \left(\sum_{i=1}^{N} X_i' V_i^{-1}(\alpha) X_i\right)^{-1} \\
&= \left(\sum_{i=1}^{N} X_i' V_i^{-1} X_i\right)^{-1} \qquad (3.30)
\end{aligned}
$$

The standard errors based on the above expression are valid, only if E(Y_i) and var(Y_i) are correctly modelled as $X_i\beta$ and $V_i = Z_i G Z_i' + \Sigma_i$ respectively. This covariance specification method is therefore often called the naive estimate. In practice it is often difficult to assess correct specification of the covariance structure, therefore one often prefers standard errors to be based on var$(\hat{\beta})$, obtained by replacing var(Y_i) in (3.30) by $r_i r_i'$ where $r_i = y_i - X_i\hat{\beta}$ and therefore var(Y_i) is estimated by $\left[Y_i - X_i\hat{\beta}\right]\left[Y_i - X_i\hat{\beta}\right]'$ rather than V_i. This leads to robust or empirical standard errors which are still consistent, as long as the mean is correctly specified. This suggests that as long as interest is only in inferences for the mean structure, little effort should be spent in modelling the covariance structure if the data set is sufficiently large. A common problem in statistical analysis is that of testing hypothesis about group comparison or contrasts. For any known matrix L, consider testing the hypothesis

$$H_o : L\beta = 0 \qquad (3.31)$$

versus

$$H_A : L\beta \neq 0. \tag{3.32}$$

The Wald test statistic to test a such hypothesis is given by

$$G_s = \hat{\beta}' L' \left[L \left(\sum_{i=1}^{N} X_i' V_i^{-1}(\hat{\alpha}) X_i \right)^{-1} L' \right]^{-1} L\hat{\beta}. \tag{3.33}$$

An asymptotic null distribution of G_s is χ^2 with degrees of freedom given by the rank of L. However it should be noted that the variability introduced from replacing α by some estimate is not taken into account in Wald tests. Therefore, Wald tests will only provide valid inferences in sufficiently large samples. This is often resolved by replacing the χ^2 distribution by an appropriate F distribution for testing hypotheses about β. In other words, an F-statistics is an alternative test statistic to the Wald test. For the null hypothesis H_0 stated above the corresponding F-test statistic is given by:

$$F_s = \frac{\hat{\beta}' L' \left[L \left(\sum_{i=1}^{N} X_i' V_i^{-1}(\hat{\alpha}) X_i \right)^{-1} L' \right]^{-1} L\hat{\beta}}{\text{rank}(L)} \tag{3.34}$$

and approximate null distribution of the statistic above is F with numerator degrees of freedom equal to rank(L), and denominator degrees of freedom are estimated from the data. There are a number of methods that one can use to estimate denominator degrees of freedom such as the Containment method, Satterthwaite approximation, Kenward and Roger approximation just to mention a few. As in most longitudinal data analysis applications, we assume different individuals contribute independent information, which results in the numbers of denominator degrees of freedom which are typically large enough, such that whatever estimation method is used for estimation of degrees of freedom we end up with similar p-values (Molenberghs and Verbeke, 2005).

Only for very small samples, or when linear models are used outside the context of longitudinal data analysis, should different estimation methods for estimation of degrees of freedom lead to severe differences in the resulting p-values. For univariate hypotheses rank(L)=1 and the F-test reduces to a t-test. Having dealt with Wald, t- and F-tests one needs to consider the likelihood-ratio test (LR) as it is an alternative test that can be used to test hypotheses about model parameters. A likelihood-ratio test is a statistical test for goodness of fit between two nested models. A complex model is compared to a relatively simpler model to see if it fits the data significantly better. It is a statistical test for making a decision between two hypotheses based on the value of the ratio of the likelihood under a restricted parameter space compared to an unrestricted one.

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

$$-2log\lambda = -2logL(\hat{\beta}_0) + 2logL(\hat{\beta}) \tag{3.35}$$

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

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

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

3.5.2 Inference for the variance components

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

3.5.3 Information criteria

Likelihood ratio test is best applicable to compare nested models and not non-nested models. The general idea behind LR test for comparing model A to a more extensive model B is to select model

A if the increase in likelihood under model B is small compared to increase in complexity. A similar argument can be used to compare non-nested models. Akaike's information criterion (AIC) is a measure of goodness of fit of an estimated statistical model. It is not a test on the model in the sense of hypothesis testing, rather it is a tool for model selection.

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

3.6 Inference for the random effects

Although, one is usually primarily interested in estimating parameters in the marginal model, it is often useful under the conditional model assumptions to calculate estimates for the random effects b_i as well because they reflect how much the subject-specific profiles deviate from the overall average profile $X_i\beta$. Since $E(Y_i|b_i) = X_i\beta + Z_ib_i$ such estimates can then be interpreted as residuals which may be helpful for detecting outlying individuals who are behaving differently over time. Also, estimates for the random effects are needed whenever interest is in the prediction of subject-specific evolutions or trajectories (Molenberghs and Verbeke, 2005).

Obviously, it is then no longer sufficient to assume that the data can be described well by the marginal model $\mathrm{N}(X_i\beta, V_i)$ (Molenberghs and Verbeke, 2005). This is only meaningful under the conditional model interpretation since $Y_i|b_i \sim \mathrm{N}(X_i\beta + Z_ib_i, \Sigma_i)$. Because the subject-specific parameters b_i are assumed random, it is most natural to estimate them using similar approaches as the Bayesian techniques (Molenberghs and Verbeke, 2005). To explore the inference for the random effects we will start with the posterior density

$$
\begin{aligned}
f(b_i|y_i) &\equiv f(b_i|Y_i = y_i) \\
&= \tfrac{f(y_i|b_i)f(b_i)}{\int f(y_i|b_i)f(b_i)db_i} \\
&\propto f(y_i|b_i)f(b_i) \\
&\propto exp\left\{ \tfrac{-1}{2}\Big(b_i - GZ_i'V_i^{-1}(y_i - X_i\beta)\Big)'\Lambda_i^{-1}\Big(b_i - GZ_i'V_i^{-1}(y_i - X_i\beta)\Big)\right\}
\end{aligned}
$$

for some positive definite Λ_i. Therefore the posterior distribution is

$$b_i|y_i \sim N(GZ_i'V_i^{-1}(y_i - X_i\beta), \Lambda_i). \tag{3.36}$$

The posterior mean as an estimate for b_i is

$$\begin{aligned} \hat{b_i}(\theta) &= E[b_i|Y_i = y_i] \\ &= \int b_i f(b_i|y_i) db_i \\ &= GZ_i'V_i^{-1}(\alpha)(y_i - X_i\beta) \end{aligned} \tag{3.37}$$

$\hat{b_i}(\theta)$ is normally distributed with zero mean and covariance matrix

$$\text{var}(\hat{b_i}(\theta)) = GZ_i'\left\{V_i^{-1} - V_i^{-1}X_i\left(\sum_{i=1}^{N} X_i'V_i^{-1}X_i\right)^{-1} X_i'V_i^{-1}\right\}Z_iG \tag{3.38}$$

Nonetheless inferences for b_i should account for the variability in b_i. Therefore the inference for b_i is usually based on

$$\text{var}(\hat{b_i}(\theta) - b_i) = G - \text{var}(\hat{b_i}(\theta)) \tag{3.39}$$

and this takes into account the variation of b_i. Parameters in θ can be replaced by ML or REML estimates obtained from fitting the marginal model. $\hat{b_i}(\hat{\theta})$ is called Empirical Bayes estimate for b_i. Similar to fixed effects, inference is often based on approximate t-tests or F-tests rather than on Wald test. It immediately follows that for any linear combination λb_i of the random effects, the following inequality holds

$$\text{var}(\lambda'\hat{b_i}) \leq \text{var}(\lambda' b_i) \tag{3.40}$$

indicating that the Empirical Bayes estimates show less variability than actually present in the random-effects population (Molenberghs and Verbeke, 2005). This is often called shrinkage estimation. Often parameters of interest are linear combinations of fixed effects in β and random effects in b_i. For example, the subject-specific slope is the sum of the average slope for subjects with the same covariate values, and the subject-specific random slope for that subject. Thus suppose

$$\mu = \lambda_\beta'\beta + \lambda_b' b_i \tag{3.41}$$

is of interest, then

$$\hat{\mu} = \lambda_\beta'\hat{\beta} + \lambda_b'\hat{b_i} \tag{3.42}$$

is the best linear unbiased predictor (BLUP) in a sense that it is linear in the observations Y_i, unbiased for μ and has minimum variance among all unbiased linear estimators.

Now, consider the prediction of the evolution of the ith subject, then

$$\begin{aligned}
\hat{Y_i} &\equiv X_i\hat{\beta} + Z_i\hat{b_i} \\
&= X_i\hat{\beta} + Z_iGZ_i'V_i^{-1}(y_i - X_i\hat{\beta}) \\
&= (I_{n_i} - Z_iGZ_i'V_i^{-1})X_i\hat{\beta} + Z_iGZ_i'V_i^{-1}y_i \\
&= \Sigma_iV_i^{-1}X_i\hat{\beta} + (I_{n_i} - \Sigma_iV_i^{-1})y_i. \qquad (3.43)
\end{aligned}$$

Y_i is a weighted mean of the population-averaged profile $X_i\hat{\beta}$ and the observed data y_i, with weights $\hat{\Sigma_i}\hat{V_i^{-1}}$ and $I_{n_i} - \hat{\Sigma_i}\hat{V_i^{-1}}$ respectively. The regression coefficients in a random effect model have a subject-specific interpretation. The importance of (3.43) is that a bigger weight goes to the overall population mean if the within subject variability is high, while if between subject variability is large then more weight is given to y_i.

To illustrate the shrinkage process, consider an example in McCulloch et al. (2008) page 170: "Suppose the kth bull has a daughter with average milk yield $\bar{y}_k$. It is perfectly reasonable to think that in the population of bulls there will be bulls other than the kth that nevertheless have (or could have) the same daughter average, namely $\bar{y}_k$. Despite this, these bulls will not necessarily all have the same genetic values, let alone all the same as that of bull k. Therefore, since $\bar{y}_k$ is our data, and if a is the random effect representing bull genetic values, the best we can do for estimating bull k's genetic value is the conditional mean $E[a|\bar{y}_k]$. Not surprisingly, since the predictors calculated as $E[a_i|y]$ are 'best', they have smaller mean squared error than would estimates based on assuming the random effects were fixed effects. They also have less variability and are sometimes called *shrinkage estimators*. This is because

$$\begin{aligned}
\text{var}(a) &= \text{var}(E[a|y]) + E[\text{var}(a|y)] \\
&= \text{var}(\tilde{a}) + \text{a positive value},
\end{aligned}$$

where $\tilde{a} = E[a|y]$ is the predictor. Thus $\text{var}(\tilde{a}) \leq \text{var}(a)$ and so $\tilde{a}$ is said to be a *shrinkage estimator*". Relating McCullloch's example to the CD4 count data for this analysis it is reasonable to think that in the population of patients on HAART, there will be other patients on HAART other than those included in the current analysis that have the same average CD4+ count. These patients will however not have the same genetic values, let alone social and behavioural characteristics.

Chapter 4

Application

4.1 Introduction

In this chapter a series of models will be fitted to the square root CD4+ count with the aim of coming up with a model or models which best describe the longitudinal CD4+ count data. 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. The next stage of the process will be an attempt to capture any unobserved heterogeneity by means of allowing for possible subject-specific random effects and correlation structure between the random effects if any. The models will be built in increasing order of complexity and will be shown in subsequent sections and sub-sections.

4.2 Univariate models

4.2.1 Marginal models

We will start by fitting marginal models and then deal with subject specific models afterwards. In univariate models the relationship between square root CD4+ count and each of the explanatory variables will be explored. Our explanatory variables are site, sex, age, weight, log viral load and time post HAART initiation. Site and sex are categorical variables both with two levels, while age, weight and log viral load are continuous variables. The first model is where we fit time as the predictor variable and square root CD4+ count as the response to see how the square root CD4+ count change over time. Thus the first model is given by

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \varepsilon_{ij} \tag{4.1}$$

where Y_{ij} is the response vector for ith subject measured at time t_{ij} for i=1, . . . , N and j=1,

$\ldots, n_i$. This model gives us the average intercept and slope respectively of all the subjects. In this model individuals do not vary in their baseline level of response as well as in their change in the mean response over time. The SAS program for this model is given below.

proc mixed data=newdata method=reml covtest noclprint empirical;

class pid visits;

model sqrtcd4=visit/solution;

repeated visits/subject=pid;

run;

This is a brief explanation of what each statement or piece of the program does within this particular SAS software procedure. PROC MIXED statement calls SAS software procedure mixed. One may ask why PROC MIXED is used and not PROC GLM. PROC MIXED uses all available data not only the complete cases but also the incomplete cases. In the CD4+ count data that is used for this project, PROC GLM would have only used the data for patients with all 5 measurements available. 'Method' specifies the estimation method used for the analysis. The 'covtest' gives p-values for all variance and covariance estimates. The 'empirical' option gives us empirical standard errors. This method yields a consistent estimator of precision, even if the covariance is misspecified (Verbeke and Molenberghs, 2000).

The CLASS statement defines categorical variables in the model. The MODEL statement specifies the fixed model, it also includes 'intercept' by default. This statement allows for time-varying, time-invariant and cross level variables all together. REPEATED statement gives the ordering of measurements within subjects. The effect(s) specified must be categorical. 'Type' gives the type of residual covariance matrix Σ_i. If omitted, we get the default $\Sigma_i = \sigma^2 I_{ni}$. The variable visit is a continuous variable denoting the measurement times post HAART initiation and visits is the categorical version of visit. The variable pid contains unique numbers associated with each patient. It is also categorical in nature.

Table 4.1: Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error	Z value	Pr Z
Residual	16.8623	0.3468	48.62	<0.0001

Table 4.1 shows that there is a significant within subject variation shown by the residuals. This may also indicate some lack of fit because of some extra variability that ought to be accounted for. The results in Table 4.2 give us an average intercept and slope over time. β_0=10.7887 is the average intercept across patients. In other words, the average square root CD4+ count at baseline is 10.7887 and this is an estimate of the patients from the CAPRISA-CAT study only. β_1=2.1792 is the average slope across patients. Hence the average person with a square root CD4+ count of

Table 4.2: Solution for Fixed Effects

Effect	Estimate	Standard error	DF	t Value	Pr>|t|
Intercept	10.7887	0.1026	1175	105.19	<0.0001
Visit	2.1792	0.0423	3553	51.56	<0.0001

10.7887 gained square root CD4+ count of 2.1792 per visit regardless of age, sex or site. This is in line with Figure 2.4 in chapter 2. The next model is where the relationship between CD4+ count and sex will be explored. The model is given by:

$$Y_{ij} = \beta_0 + \beta_1 G_i + \beta_2 t_{ij} + \beta_3 G_i t_{ij} + \varepsilon_{ij} \tag{4.2}$$

where G_i=1 if i^{th} patient is female, and G_i=0 otherwise. In this model, the mean rate of change for males and females is given by β_2 and (β_2+β_3) respectively. The SAS program for this model is shown below:

proc mixed data=newdata method=reml covtest noclprint ;
class pid visits sex;
*model sqrtcd4= sex visit visit*sex/solution ddfm=kr;*
repeated visits/subject=pid;
run;

The results after fitting model (4.2) are shown in Table 4.3 and 4.4.

Table 4.3: Covariance Parameter Estimates

Cov parm	Estimate	Standard error	Z value	Pr Z
Residual	16.5144	0.3397	48.61	<0.0001

Table 4.4: Solution for Fixed Effects

Effect	Sex	Estimate	Standard error	DF	t Value	Pr>|t|
Intercept		10.3480	0.1728	4726	59.87	<0.0001
Sex	Female	0.6447	0.2077	4726	3.10	0.0019
Visit		1.9503	0.0760	4726	25.67	<0.0001
Visit*Sex	Female	0.3182	0.0908	4726	3.51	0.0005

The covariance parameter estimates in Table 4.3 are similar to Table 4.1. But of note is that the residual estimate has decreased slightly. Results in Table 4.4 show that female and male patients

have mean square root CD4+ count of 10.9927 (10.3480+0.6447) and 10.3480 respectively. They also, on average gain square root CD4+ count at the rate of 2.2685 (1.9503+0.3182) and 1.9503 respectively. The intercepts and slopes 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 2.7. The next sub-model fitted is to assess whether the mean rate of change in square root CD4+ count is the same for patients in the two sites. The model for the two sites is given by:

$$Y_{ij} = \beta_0 + +\beta_1 S_i + \beta_2 t_{ij} + \beta_3 S_i t_{ij} + \varepsilon_{ij} \tag{4.3}$$

where S_i=1 if i^{th} patient is from the eThekwini site, and S_i=0 otherwise. In this model, the mean rate of increase for individuals from Vulindlela and eThekwini sites is given by β_2 and (β_2+β_3) respectively. Similar to the model on sex, the SAS program for this model is given by:

proc mixed data=newdata method=reml covtest noclprint empirical;
class pid visits site;
*model sqrtcd4= site visit visit*site/solution ;*
repeated visits/subject=pid;
run;
The results for this model are shown in Table 4.5 and 4.6.

Table 4.5: Covariance Parameter Estimates

Cov parm	Estimate	Standard error	Z value	Pr Z
Residual	16.8532	0.3467	48.61	<0.0001

Table 4.6: Solution for Fixed Effects

Effect	Sex	Estimate	Standard error	DF	t Value	Pr>|t|
Intercept		10.6561	0.1247	1174	85.46	<0.0001
Site	eThekwini	0.3719	0.2179	1174	1.71	0.0881
Visit		2.2462	0.0546	3552	41.11	<0.0001
Visit*Site	eThekwini	-0.1850	0.0857	3552	-2.16	0.0309

Results in Table 4.6 show that on average patients from eThekwini and Vulindlela started HAART with a mean square root CD4+ count of 11.028 (10.6561+0.3719) and 10.6561 respectively. Of note is that the two means are not statistically significantly different. On average the rate of increase in square root CD4+ count is 2.0612 (2.2462-0.1850) and 2.2462 for patients from eThekwini and Vulindlela respectively. One can see that even though the eThekwini patients start with high

mean square CD4+ count compared to those from Vulindlela but the rate of change in square root CD4+ count is less compared to those from Vulindlela. The next set of models to be explored are those relating square root CD4+ count to three continuous variables namely age, log viral load and weight separately. The model for age is given below:

$$Y_{ij} = \beta_0 + \beta_1 A_i + \beta_2 t_{ij} + \beta_3 A_i t_{ij} + \varepsilon_{ij} \tag{4.4}$$

where A_i= is the baseline age (in years) of the patient. The SAS program for this model is shown below:

proc mixed data=newdata method=reml covtest noclprint empirical;
class pid visits;
*model sqrtcd4=age visit visit*age/solution ;*
repeated visits/subject=pid;
run;

Table 4.7: Covariance Parameter Estimates

Cov parm	Estimate	Standard error	Z value	Pr Z
Residual	16.6740	0.3434	48.56	<0.0001

Table 4.8: Solution for Fixed Effects

Effect	Estimate	Standard error	DF	t Value	Pr>|t|
Intercept	10.1717	0.4407	1172	23.08	<0.0001
Age	0.0179	0.0123	3544	1.46	0.1453
Visit	3.1965	0.1875	3544	17.05	<0.0001
Visit*Age	-0.0297	0.0052	3544	-5.67	<0.0001

The estimate for age is not statistically significantly different from zero. This implies that the regression of square root CD4+ count on age is not statistically significant (p=0.1453) given time (visit) and visit*age are in the model. This means that younger and older patients started HAART with almost the same CD4+ count. However, the interaction visit*age is negative and significant. The interpretation of the interaction term for two continuous variables is tricky, and that is why continuous variables are frequently modified into categorical variables (van Walvaren and Hart, 2008). However, categorizing continuous variables can cause problems and the first one is information loss (van Walvaren and Hart, 2008). Taylor and Yu (2002) found that categorizing one

continuous variable can artificially make another variable appear associated with the outcome. Hallahan (2003) reported that it is useful to graphically interpret these interactions. However, for the output in Table 4.8 one can graphically interpret the interaction term by selecting percentiles of interest for age and plot them against time and the predicted square root CD4+ count.

Hallahan (2003) followed Tukey's suggestions and selected the 5%, 25%, 50%, 75% and 95% percentiles of age which according to the current data give the following age groups 23, 28, 32, 39 and 51. In order to plot this interaction we will use the estimates in Table 4.8 and replace A_i by each age group. For example, the predicted value $\hat{Y}_{ij}=\hat{\beta}_0+\hat{\beta}_1 A_i+\hat{\beta}_2 t_{ij}+\hat{\beta}_3 A_i t_{ij}$ at baseline (i.e. when time=0) and age=23 is given by 10.1717+0.0179(23)+3.1965(0)+(-0.0297(0)(23))=10.5834. This can be done for all the ages at baseline and at all the other visits. The graphical presentation of the interaction term is shown in Figure 4.1.

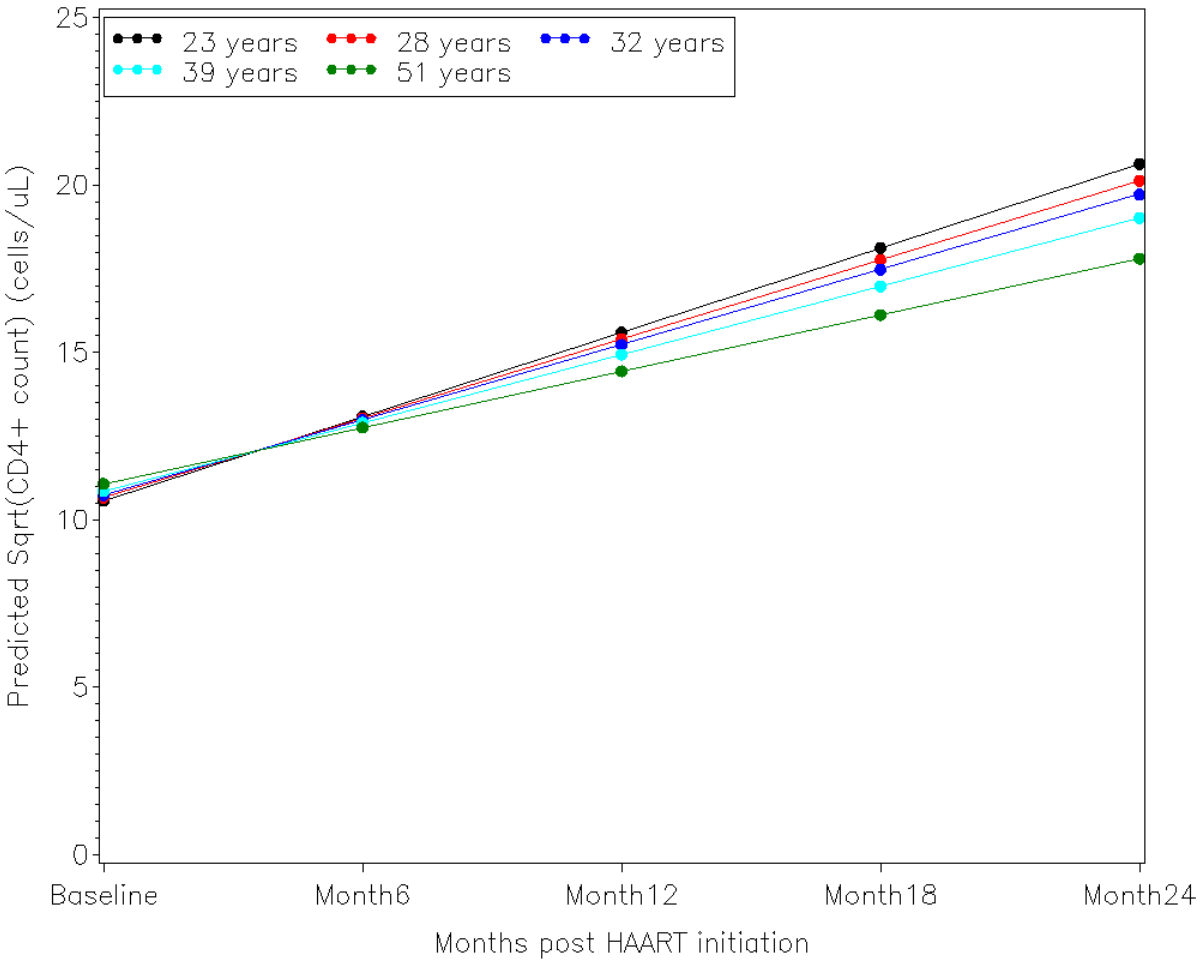

Figure 4.1 Interaction between age (years) and time

Figure 4.1 shows that, as baseline age increases the rate of increase of square root CD4+ count over time decreases. This basically tells us that the rate of increase in square CD4+ count over time is higher for younger than older individuals. This also means that on average younger patients are doing better than older patients. Even though every age group is gaining CD4+ count over time the rate of change for younger patients is greater than that for older patients. This was also evident in the exploratory analysis depicted in chapter 2, Figure 2.9. Next the effect of weight on square root CD4+ count was modelled and the results of this analysis are shown in Tables 4.9 and 4.10. However, it is noted that weight is a time varying covariate, but we will plot selected percentiles for baseline weight. The model is given by:

$$Y_{ij} = \beta_0 + \beta_1 W_{ij} + \beta_2 t_{ij} + \beta_3 W_{ij} t_{ij} + \varepsilon_{ij} \tag{4.5}$$

where W_{ij}= is the weight (in kg) of the patient.

Table 4.9: Covariance Parameter Estimates

Cov parm	Estimate	Standard error	Z value	Pr Z
Residual	16.0132	0.3464	46.22	<0.0001

Table 4.10: Solution for Fixed Effects

Effect	Estimate	Standard error	DF	t Value	Pr>|t|
Intercept	5.7231	0.5430	1158	10.54	<0.0001
Weight	0.0814	0.0084	3115	9.69	<0.0001
Visit	2.9364	0.2315	3115	12.69	<0.0001
Visit*Weight	-0.0137	0.0033	3115	-4.12	<0.0001

Even though the regression of mean square rot CD4+ count on age was not significant, results in Table 4.10 indicate that the effect of weight on mean square root CD4+ count is significant. Parameter estimate of the effect of weight on square root CD4+ count is 0.0814 and it is significant (p<0.0001) indicating that mean square root CD4+ count is correlated to the weight of that individual at that time. The graphical presentation of the interaction is shown in Figure 4.2.

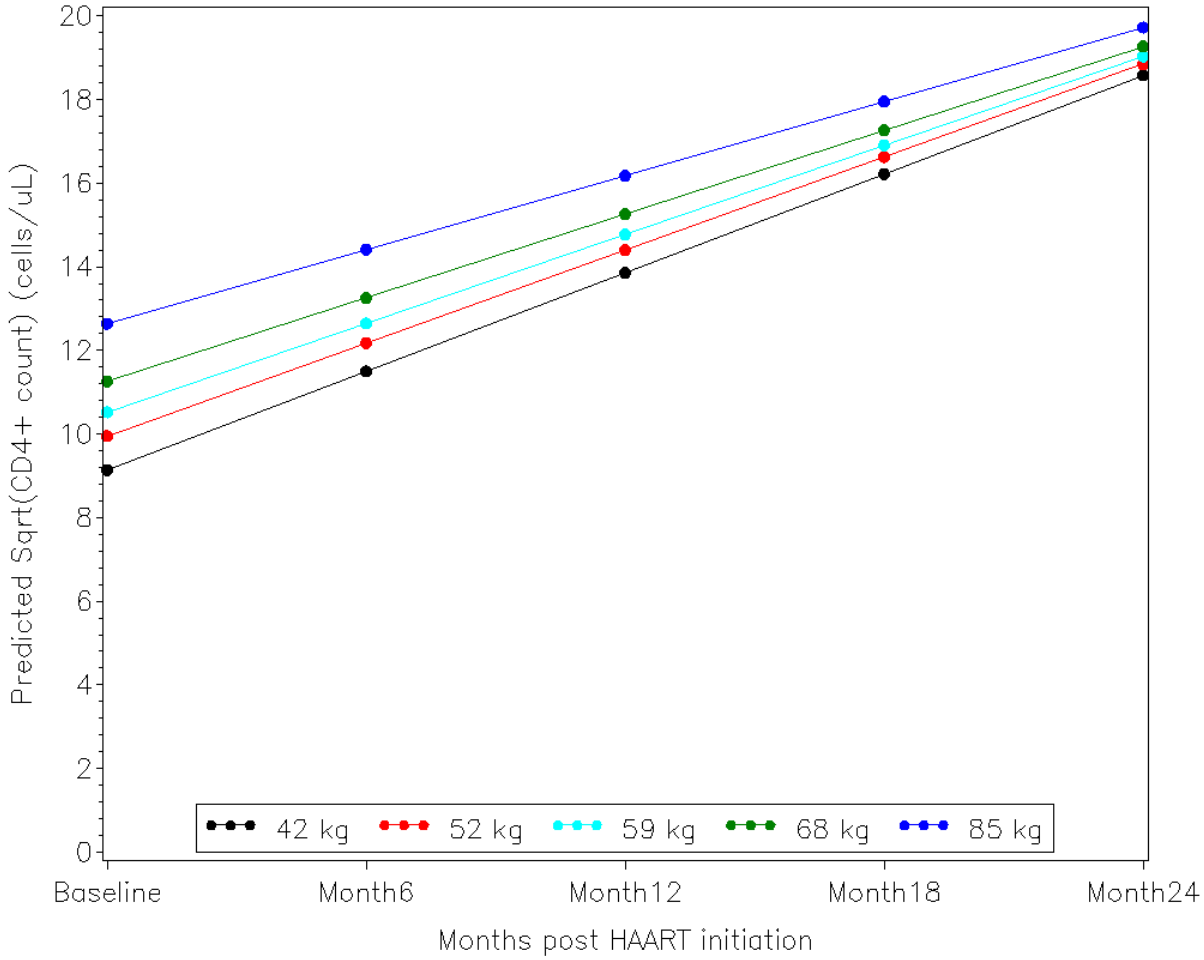

Figure 4.2 Interaction between weight and time

Figure 4.2 shows that as weight increases the rate of square root CD4+ count gain also increases, but the slopes seem not to be parallel over time. A careful assessment of Figure 4.2 shows that even though patients with low weight started with low CD4+ count, their rate of change in CD4+ ultimately converges to the slope for those who started HAART with higher weight. Now consider a model relating square root CD4+ count to log viral load. The model is given by:

$$Y_{ij} = \beta_0 + \beta_1 L_{ij} + \beta_2 t_{ij} + \beta_3 L_{ij} t_{ij} + \varepsilon_{ij} \tag{4.6}$$

where L_{ij}= is the log viral load (in copies/ml) of the patient. The results are shown in Tables 4.11 and 4.12.

Table 4.11: Covariance Parameter Estimates

Cov parm	Estimate	Standard error	Z value	Pr Z
Residual	15.0518	0.3243	46.41	<0.0001

Table 4.12: Solution for fixed effects

Effect	Estimate	Standard error	DF	t Value	Pr>|t|
Intercept	16.0394	0.2217	1124	72.35	<0.0001
Visit	1.2663	0.1073	976	11.80	<0.0001
Logv	-1.2078	0.0455	2208	-26.57	<0.0001
Visit*Logv	0.0410	0.0382	2208	1.07	0.2825

Table 4.12 shows that for every unit increase in log viral load square root CD4+ count decreases by 1.2078 subject to other effects held constant. Thus there is a negative correlation between CD4+ count and log viral load. The interaction term visit*logv is not statistically significant meaning that the rate of change is the same for everyone regardless of the level of the initial log viral load.

4.2.2 Random effects models

The set of models similar to the marginal models covered above will be fitted except that now we allow the intercept and slope to account for subject to subject heterogeneity through the subject specific random effects. In random intercept and slope effects models, individuals vary not only in their baseline level of response, that is, when time is zero, but vary also in terms of their change in the mean response over time. The notation in random effects models is the same as what was used in marginal models, except now b_{0i} and b_{1i} denote the random intercept and slope effects respectively. The first random effect model is given by:

$$Y_{ij} = \beta_0 + b_{0i} + (\beta_1 + b_{1i})t_{ij} + \varepsilon_{ij} \tag{4.7}$$

In SAS proc MIXED code the RANDOM statement defines the random effects in the model. It also specifies the G matrix developed in section 3.3.1 in the full linear mixed model. The 'subject' option under the repeated and random statements simply specify which unit is repeatedly observed or measured in the study. Independence across subjects is automatically assumed. 'Type' gives the type of random effects variance-covariance matrix G. If omitted, we get the default variance component structure $\sigma^2 I_{n_i}$. The results of the model specified in (4.7) under proc MIXED are displayed in Tables (4.13) and (4.14).

Table 4.13: Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error	Z value	Pr Z
UN(1,1)	8.3928	0.5260	15.96	<0.0001
UN(2,1)	-0.3836	0.1575	-2.44	0.0149
UN(2,2)	0.8612	0.0741	11.61	<0.0001
Residual	5.8883	0.1637	35.97	<0.0001

Table 4.13 gives parameter estimates for the unknown symmetric G matrix. The variance components for the G matrix are statistically significantly different from zero and therefore the random intercepts and slopes vary from individual to individual. The covariance between the random intercept and slope is negative and statistically significant. The remaining within subject variation shown by the residuals is also significant. This may mean that there is some extra variability which has not yet been accounted for by the model.

Table 4.14: Solution for Fixed Effects

Effect	Estimate	Standard error	DF	t Value	Pr>|t|
Intercept	10.7031	0.1025	1175	104.57	<0.0001
Visit	2.1241	0.0401	1006	53.02	<0.0001

The results in Table 4.14 give us an average intercept and slope over time where β_0=10.7031 is the average square root CD4+ count at baseline and this is an estimate of the population mean. β_1=2.1231 is also an estimate for the population rate of increase of square root CD4+ count. The intercept and slope for each patient is give by $(\beta_0 + b_{0i})$ and $(\beta_1 + b_{1i})$ respectively. The next model is where we consider the relationship between CD4+ count and sex will be explored. The model is given by:

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1 G_i + \beta_2 t_{ij} + \beta_3 G_i t_{ij} + b_{1i} t_{ij} + \varepsilon_{ij}. \tag{4.8}$$

In this model, the mean rate of change for males and females is given by β_2 and $(\beta_2+\beta_3)$ respectively. The SAS program for this model is shown below:

```
proc mixed data=newdata method=reml covtest noclprint empirical;
class pid visits sex;
model sqrtcd4=visit sex visit*sex/solution ;
repeated visits/subject=pid;
random intercept visit/subject=pid type=un;
run;
```

The results after fitting model (4.8) are shown in Table 4.15 and 4.16.

Table 4.15: Covariance Parameter Estimates

Cov parm	Estimate	Standard error	Z value	Pr Z
UN(1,1)	8.3150	0.5230	15.90	<0.0001
UN(2,1)	-0.4286	0.1560	-2.75	0.0060
UN(2,2)	0.8332	0.0729	11.43	<0.0001
Residual	5.8885	0.1636	35.98	<0.0001

Table 4.16: Solution for Fixed Effects

Effect	Sex	Estimate	Standard error	DF	t Value	Pr>|t|
Intercept		10.2613	0.1900	1175	54.00	<0.0001
Sex	Female	0.6422	0.2251	2546	2.85	0.0044
Visit		1.8732	0.0704	1005	26.60	<0.0001
Visit*Sex	Female	0.3687	0.0852	2546	4.21	<0.0001

Results in Table 4.16 are similar to results in Table 4.4. However, the estimates for the random effects model are smaller than those for the marginal model. Note in the Gaussian case it is straight forward to switch between the marginal and subject specific model but not that obvious in non-Gaussian models (Molenberghs and Verbeke, 2005). The next sub-model fitted is to assess whether the mean rate of change is the same in the two populations represented by the two sites. The model for the two sites is given by:

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1 S_i + \beta_2 t_{ij} + \beta_3 S_i t_{ij} + b_{1i} t_{ij} + \varepsilon_{ij} \tag{4.9}$$

The mean rate of change for Vulindlela and eThekwini sites is given by β_2 and (β_2+β_3) respectively. The SAS program for this model is shown below:

proc mixed data=newdata method=reml covtest noclprint empirical;
class pid visits site;
*model sqrtcd4=visit site visit*site/solution ;*
repeated visits/subject=pid;
random intercept visit/subject=pid type=un;
run;

The results for this model are shown in Table 4.17 and 4.18.

Table 4.17: Covariance Parameter Estimates

Cov parm	Estimate	Standard error	Z value	Pr Z
UN(1,1)	8.3568	0.5244	15.94	<0.0001
UN(2,1)	-0.3645	0.1568	-2.32	0.0201
UN(2,2)	0.8546	0.0738	11.58	<0.0001
Residual	5.8835	0.1635	35.99	<0.0001

Table 4.18: Solution for Fixed Effects

Effect	Sex	Estimate	Standard error	DF	t Value	Pr>$\|t\|$
Intercept		10.5453	0.1252	1174	84.21	<0.0001
Site	eThekwini	0.4475	0.2168	2547	2.06	0.0391
Visit		2.2059	0.0520	1005	42.45	<0.0001
Visit*Site	eThekwini	-0.2224	0.0808	2547	-2.78	0.0055

The results in Table 4.18 are slightly different from the results in Table 4.6 where we fitted the marginal model. The average intercepts for the two sites were not statistically significantly different in Table 4.6. The inclusion of random intercept and slope in the model allows the sites to differ significantly at 5% significance levels. Just like in section 4.2.1, the next set of models will explore the relationship between square root CD4+ count and the three continuous variables namely age, log viral load and weight separately. Thus the model for age is given by:

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1 A_i + \beta_2 t_{ij} + \beta_3 A_i t_{ij} + b_{1i} t_{ij} + \varepsilon_{ij} \tag{4.10}$$

The SAS program for this model is shown below:

proc mixed data=newdata method=reml covtest noclprint empirical;
class pid visits;
*model sqrtcd4=visit age visit*age/solution ;*
repeated visits/subject=pid;
random intercept visit/subject=pid type=un;
run;

The results of model formulation (4.10) are given in Tables (4.19) and (4.20).

Table 4.19: Covariance Parameter Estimates

Cov parm	Estimate	Standard error	Z value	Pr Z
UN(1,1)	8.4286	0.5296	15.92	<0.0001
UN(2,1)	-0.3863	0.1564	-2.47	0.0135
UN(2,2)	0.8235	0.0727	11.33	<0.0001
Residual	5.8955	0.1642	35.91	<0.0001

Table 4.20: Solution for Fixed Effects

Effect	Estimate	Standard error	DF	t Value	Pr>$\|t\|$
Intercept	10.0891	0.4364	1172	23.12	<0.0001
Age	0.0179	0.0121	2539	1.48	0.1400
Visit	2.9466	0.1846	1005	15.96	<0.0001
Visit*Age	-0.0240	0.0052	2539	-4.67	<0.0001

The results for the random effects model in Table 4.20 are similar to the results for the marginal model given in Table 4.8 except that now the standard errors are slightly smaller. The model dealing with weight is given by

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1 W_{ij} + \beta_2 t_{ij} + \beta_3 W_{ij} t_{ij} + b_{1i} t_{ij} + \varepsilon_{ij} \tag{4.11}$$

where W_{ij}= is the weight (in kg) of the patient. The corresponding results are given in Tables (4.21) and (4.22).

Table 4.21: Covariance Parameter Estimates

Cov parm	Estimate	Standard error	Z value	Pr Z
UN(1,1)	8.7776	0.5629	15.59	<0.0001
UN(2,1)	-0.4826	0.1666	-2.90	0.0038
UN(2,2)	0.8292	0.0758	10.95	<0.0001
Residual	5.2800	0.1615	32.69	<0.0001

Table 4.22: Solution for Fixed Effects

Effect	Estimate	Standard error	DF	t Value	Pr>|t|
Intercept	3.1223	0.5388	1158	5.79	<0.0001
Weight	0.1225	0.0085	2120	14.48	<0.0001
Visit	3.0920	0.2022	995	15.29	<0.0001
Visit*Weight	-0.0177	0.0029	2120	-6.01	<0.0001

The results show that the standard errors for fixed effects are now slightly reduced. Also the size of the fixed effects has increased under the random effects model except for the intercept, when compared to the marginal effects model in Table 4.10. The final model is where we fit the square root CD4+ count and log viral load. The model for log viral load is given by:

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1 L_{ij} + \beta_2 t_{ij} + \beta_3 L_{ij} t_{ij} + b_{1i} t_{ij} + \varepsilon_{ij} \tag{4.12}$$

where $L_{ij}=$ is the log viral load (in copies/ml) of the patient. The results are shown in tables (4.23) and (4.24).

Table 4.23: Covariance Parameter Estimates

Cov parm	Estimate	Standard error	Z value	Pr Z
UN(1,1)	7.8905	0.5329	16.68	<0.0001
UN(2,1)	-0.6794	0.1540	-4.41	<0.0001
UN(2,2)	0.9120	0.0718	12.71	<0.0001
Residual	4.1072	0.1230	33.40	<0.0001

Table 4.24: Solution for fixed effects

Effect	Estimate	Standard error	DF	t Value	Pr>|t|
Intercept	16.0589	0.2226	1124	72.15	<0.0001
Logv	-1.2146	0.0456	2208	-26.61	<0.0001
Visit	1.2622	0.1073	976	11.76	<0.0001
Visit*Logv	0.0454	0.0382	2208	1.19	0.2236

Looking at the results from the corresponding marginal model in Table 4.12 and those in Table 4.24 there seems to be just mild differences in parameter estimates and standard errors. Otherwise results are generally similar.

4.3 Multi-covariate models

So far the models fitted were sub-models explaining the rate of change of square root CD4+ count over time with a single covariate at a time. In this section the aim is to model the rate of change of square root CD4+ count over time but now including all potential covariates. We will start by fitting the marginal model and subsequently consider the corresponding random effects model.

4.3.1 Marginal model

When fitting the marginal model, a series of correlation structures of the repeated observation within an individual will be modelled. To aid in the model selection the Akaike's information criterion (AIC) will be used. Both REML and ML estimation will be applied but once an appropriate mean model is selected the final model will be fitted using the REML approach.

Covariance structures should be carefully selected to obtain valid inferences parameters for the fixed effects. If one ignores important correlations by using a model that is too simple, one risks increasing Type I error rate and underestimating standard errors (Littell et al., 2006). If the model is too complex, the power and efficiency is sacrificed (Littell et al., 2006). The covariance structure that fits the data best is used to estimate the fixed effects parameters. Table 4.25 gives examples of some of the common covariance structures.

Table 4.25: Different covariance structures

Structure	Example	Structure	Example
VC	$\begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{pmatrix}$	UN	$\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{pmatrix}$
CS	$\begin{pmatrix} \sigma_1^2+\sigma^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2+\sigma^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2+\sigma^2 \end{pmatrix}$	AR(1)	$\begin{pmatrix} \sigma^2 & \rho\sigma^2 & \rho^2\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ \rho^2\sigma^2 & \rho\sigma^2 & \sigma^2 \end{pmatrix}$

To fit a time-series-type or serial covariance structure in which correlation declines as a function of time, one can use any one of the more flexible spatial structures available in PROC MIXED (Littell et al., 2006). Spatial structures are also useful for unequally spaced longitudinal data. Note that unequally spaced longitudinal data can be viewed as a spatial process in one dimension (Littell et al., 2006). The spatial covariance structures do not make any assumptions about the

distance between measurements because they calculate the actual distance themselves. The sample variogram can be used as a diagnostic tool to select the covariance structure (Hallahan, 2003). Table 4.26 gives some of the spatial covariance structures.

Table 4.26: Spatial covariance structures

Structure	Example
Power	$\sigma^2 \begin{pmatrix} 1 & \rho^{d_{12}} & \rho^{d_{13}} \\ \rho^{d_{12}} & 1 & \rho^{d_{23}} \\ \rho^{d_{13}} & \rho^{d_{23}} & 1 \end{pmatrix}$
Linear	$\sigma^2 \begin{pmatrix} 1 & (1-\rho d_{12}) & (1-\rho d_{13}) \\ (1-\rho d_{12}) & 1 & (1-\rho d_{23}) \\ (1-\rho d_{13}) & (1-\rho d_{23}) & 1 \end{pmatrix}$
Exponential	$\sigma^2 \begin{pmatrix} 1 & \exp(-d_{12}/\rho) & \exp(-d_{13}/\rho) \\ \exp(-d_{12}/\rho) & 1 & \exp(-d_{23}/\rho) \\ \exp(-d_{13}/\rho) & \exp(-d_{23}/\rho) & 1 \end{pmatrix}$
Gaussian	$\sigma^2 \begin{pmatrix} 1 & \exp(-d_{12}^2/\rho^2) & \exp(-d_{13}^2/\rho^2) \\ \exp(-d_{12}^2/\rho^2) & 1 & \exp(-d_{23}^2/\rho^2) \\ \exp(-d_{13}^2/\rho^2) & \exp(-d_{23}^2/\rho^2) & 1 \end{pmatrix}$

Note that in the structures displayed above the parameters σ^2 and ρ constitute the set of parameters to be estimated. Candidate covariance structures including spatial correlation structures will be fitted to see which structure best agrees with the data. We will start by writing down the full model which is given by:

$$Y_{ij} = \beta_0 + \beta_1 S_i + \beta_2 G_i + \beta_3 A_i + \beta_4 L_{ij} + \beta_5 W_{ij} + \beta_6 t_{ij} + (\beta_7 S_i + \beta_8 G_i + \beta_9 A_i + \beta_{10} L_{ij} + \beta_{11} W_{ij}) t_{ij} + \varepsilon_{ij} \quad (4.13)$$

$\beta_6 t_{ij}$ is the time effect. S_i, G_i, A_i, W_{ij} and L_{ij} have been defined in models 4.2 to 4.6 earlier. The SAS program for this model is shown below.

proc mixed data=newdata method=reml covtest noclprint empirical;
class pid sex site visits;

*model sqrtcd4= site sex age logv weight visit visit*site visit*sex visit*age visit*logv visit*weight/solution;*
repeated visits/type=&cov local subject=pid ;
Title'Different covariance structures';
run;

The variable '&cov' should be replaced by each of the covariance structures and 'local' should only be used when fitting spatial or serial covariance structures so that residual error is accommodated in addition to the serial correlation. The graphical presentation in terms of AIC, AICC and BIC for each of the covariance structures is shown in Figure 4.3. The variogram in Section 2.6 (Figure 2.21) indicated or suggested the presence of serial correlation, measurement error and random effects. The spatial covariance structures also known as serial correlation structures are able to model the serial correlation and measurement error. When selecting a covariance structure for a model, only structures that make sense for the data should be considered (Hallahan, 2003).

Having looked at the current data, the best structure that makes sense is the AR(1), but such a structure is among the inferior structures according to Figure 4.3. One reason for this is possibly because of the unexpected incompleteness due to missing values rendering time intervals unequally spaced which is a requirement for the AR(1) structure. If one looks at Figure 2.18 and 2.19, two adjacent measurements are more correlated than those that are far apart and this tends to suggest some form of an AR correlation structure. Following the argument against AR(1) structure above the spatial covariance structure is a better substitute to the AR(1) in the presence of unequal intervals. Amongst the spatial correlation structure, the one with the smallest AIC is the spatial linear and thus we adopt such a covariance structure. The fit statistics for all the spatial covariance structures is shown in Table 4.27.

Table 4.27: Fit statistics for spatial covariance structures

	SP(GAU)	SP(EXP)	SP(LIN)	SP(POW)	SP(SPH)
-2 Log likelihood	19444.9	19452.2	19445.5	19452.2	19446.2
AIC	19455.9	19458.2	19451.5	19458.2	19452.2
AICC	19455.9	19458.2	19451.5	19458.2	19452.2
BIC	19471.0	19473.2	19466.5	19473.2	19467.3

In Figure 4.3 one should note that the following covariance structures namely: VC, AR(1), CS and the toeplitz seems to be inferior. The remaining covariance structures namely the spatial and the unstructured would be appropriate in this case. The disadvantage of the UN is that it has too many parameters to estimate and that lead to computational difficulties especially when subject specific effects are included in the model. In addition it does not account for any potential trends

in the correlations.

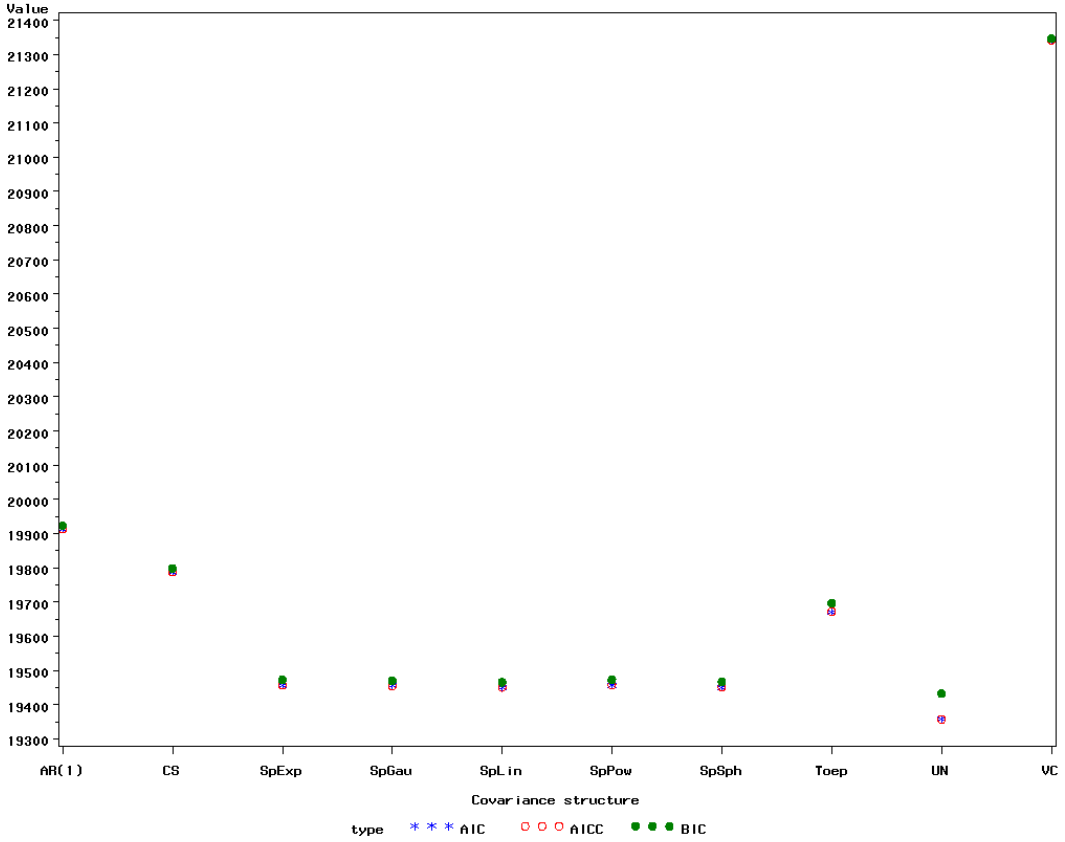

Figure 4.3 Model fit for each covariance structure

The SAS program to fit the model using spatial covariance structure is:

proc mixed data=newdata method=reml covtest noclprint empirical;

class pid sex site visits;

*model sqrtcd4= site sex age logv weight visit visit*site visit*sex visit*age visit*logv visit*weight/solution;*

repeated visits/type=sp(lin)(visit) local subject=pid ;

Title'Longitudinal model with spatial linear covariance structure';

run;

'Local' gives us the measurement error while 'sp(lin)(visit)' gives us serial correlation. The covariance parameter estimates for spatial linear structure are shown in Table 4.28. In the output the variance estimates are labelled as follows: Variance=σ_1^2, sp(lin)=ρ and residual=σ^2.

Table 4.28: Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error	Z value	Pr Z
Variance	11.9659	0.5022	23.83	<0.0001
SP(LIN)	0.1578	0.0101	15.63	<0.0001
Residual	2.1159	0.1807	11.71	<0.0001

The total variance ($\sigma_1^2 + \sigma^2$) for any given observation is given by 11.9659+2.1159=14.0818. It is satisfying to see that the SP(LIN) components are significantly justified with p-value <0.0001. Now that the covariance structure Σ_i has been selected to be the spatial linear, the next step is to fit the full model under ML estimation and remove insignificant fixed effects starting with the most insignificant one. The model that we are going to fit is called the marginal model. In this model there are no random effects and therefore no Z matrix as developed in equations 3.6 to 3.9 thus no covariance matrix G for b_i is modelled. In this model we assume that there is no individual to individual variability. The model with random effects will be shown in section 4.3. The results for the full model under ML estimation are shown in tables 4.29 and 4.30.

Table 4.29: Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error	Z value	Pr Z
Variance	11.8952	0.4988	23.85	<0.0001
SP(LIN)	0.1576	0.0101	15.60	<0.0001
Residual	2.1234	0.1803	11.78	<0.0001

The AIC for this model is 19414.6. The covariance parameter estimates under ML are similar to those in Table 4.28 where the REML estimation was used but note the slight underestimation of the variance components. This is because of the fact that REML takes into account the degrees of freedom whilst ML does not when estimating variance components.

Table 4.30: Solution for Fixed Effects

Effect	Sex	Estimate	Standard Error	DF	t value	Pr>|t|
Intercept		11.3582	0.7527	1108	15.09	<0.0001
Site	eThekwini	0.0571	0.2177	1108	0.26	0.7932
Sex	Female	0.6170	0.2346	1109	2.63	0.0087
Age		0.0039	0.0126	2771	0.31	0.7576
Logv		-1.1274	0.0489	2771	-23.05	<0.0001
Weight		0.0594	0.0084	2771	7.10	<0.0001
Visit		2.1837	0.2776	2771	7.87	<0.0001
Visit*Site	eThekwini	-0.1343	0.0778	2771	-1.73	0.0844
Visit*Sex	Female	0.2254	0.0877	2771	2.57	0.0102
Visit*Age		-0.0208	0.0051	2771	-4.08	<0.0001
Visit*Weight		-0.0033	0.0029	2771	-1.12	0.2627
Visit*Logv		-0.0182	0.0386	2771	-0.47	0.6371

Table 4.31: Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F value	Pr>F
Site	1	1109	0.07	0.7932
Sex	1	1109	6.91	0.0087
Age	1	2771	0.10	0.7576
Logv	1	2771	531.22	<0.0001
Weight	1	2771	50.46	<0.0001
Visit	1	2771	67.16	<0.0001
Visit*Site	1	2771	2.98	0.0844
Visit*Sex	1	2771	6.60	0.0102
Visit*Age	1	2771	16.66	<0.0001
Visit*Weight	1	2771	1.26	0.2627
Visit*Logv	1	2771	0.22	0.6371

Table 4.30 gives us the solution for fixed effects for the marginal model. Site, age and the interaction terms visit*site, visit*logv and visit*weight are not statistically significant and should be removed from the model starting with the most insignificant one of which is the interaction term visit*logv. Age will not be removed because it is involved in a significant interaction. Site will not be removed either because it is an important variable. Variables with subject matter importance should be kept in the model (Hallahan, 2003). It is noted that age and visit*logv were also insignificant in univariate models in section 4.2. But site and visit*site although insignificant now they were

significant in univariate model. Similarly the interaction term visit*weight was significant and now it is not significant. The model was then refitted after removing the interaction term visit*logv and the AIC dropped from 19414.6 to 19413.0 indicating a better fit. The next step is to remove the interaction term visit*weight with the p-value of 0.2953. The model was fitted again and the AIC dropped from 19413 to 19412.2. There were no other variables to be removed from the model. The final model is given by:

$$Y_{ij} = \beta_0 + \beta_1 S_i + \beta_2 G_i + \beta_3 A_i + \beta_4 L_{ij} + \beta_5 W_{ij} + \beta_6 t_{ij} + (\beta_7 S_i + \beta_8 G_i + \beta_9 A_i) t_{ij} + \varepsilon_{ij} \tag{4.14}$$

The final model was fitted under REML and the results are shown in Table 4.32, 4.33 and 4.34. The AIC for this model was 19438.

Table 4.32: Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error	Z value	Pr Z
Variance	11.9719	0.5012	23.89	<0.0001
SP(LIN)	0.1572	0.0100	15.67	<0.0001
Residual	2.1209	0.1792	11.83	<0.0001

The covariance parameter estimates for the reduced model in Table 4.32 are almost the same as parameter estimates in Table 4.28 where we fitted the full model. Note that although the residual variance is significant its parameters estimate has reduced drastically compared to in the case of univariate models fitted in section 4.2. Here much of the systematic variability in the outcome of interest has been accounted for via additional covariates and accommodating appropriate correlation structures within units or individuals.

The final model has 3 interaction terms, visit*site, visit*sex and visit*age, and 6 main effects. The likelihood ratio test comparing the full and reduced models gives a p-value of 0.4612 with 2 degrees of freedom. Therefore the reduced model is better than the full model. The results in Table 4.33 shows that there is no difference between eThekwini and Vulindlela in terms of mean CD4+ count post HAART. We noted that the intercept for the eThekwini site is greater than that for Vulindlela, but Vulindlela has the higher rate of CD4+ count gain compared to eThekwini.

There is a significant difference between males and females at the 5% level of significance. The significant gender or sex effect indicates that on average females started with the higher CD4+ count than males. In addition the interaction of sex and time is significant which implies that females have a significant higher rate of increase than males. It is interesting to note that the interaction of age and time (visit) is significant but the main effect of age is not significant. Thus

Table 4.33: Solution for Fixed Effects

Effect	Sex	Estimate	Standard Error	DF	t value	Pr>|t|
Intercept		11.7895	0.6886	1108	17.12	<0.0001
Site	eThekwini	0.0769	0.2180	1108	0.35	0.7245
Sex	Female	0.6275	0.2343	1109	2.68	0.0075
Age		0.0053	0.0125	2773	0.43	0.6688
Logv		-1.1492	0.0446	2773	-25.76	<0.0001
Weight		0.0530	0.0068	2773	7.82	<0.0001
Visit		1.9396	0.2020	2773	9.60	<0.0001
Visit*Site	eThekwini	-0.1487	0.0774	2773	-1.92	0.0546
Visit*Sex	Female	0.2152	0.0861	2773	2.50	0.0125
Visit*Age		-0.0213	0.0050	2773	-4.22	<0.0001

Table 4.34: Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F value	Pr>F
Site	1	1109	0.12	0.7245
Sex	1	1109	7.17	0.0075
Age	1	2773	0.18	0.6688
Logv	1	2773	663.83	<0.0001
Weight	1	2773	61.12	<0.0001
Visit	1	2773	115.57	<0.0001
Visit*Site	1	2773	3.70	0.0546
Visit*Sex	1	2773	6.24	0.0125
Visit*Age	1	2773	17.93	<0.0001

although at baseline age had no significant effect on CD4+ count, age seems to be significantly related to the rate of increase in CD4+ count. The parameter estimate of the interaction for age and time is -0.0213 which implies that the average rate of increase is inversely related to age. In other words, younger patients have the higher rate of change in CD4+ count than older patients. Whereas, the main effects of weight and log viral load are statistically different from zero, their interactions with time are not significant. This means that the rate of change in CD4+ count over time is not statistically significantly different whether one started HAART with lower or higher log viral load.

4.3.2 Random effects model

The random effects model or subject-specific model assumes that extra correlation arises among repeated response because the regression coefficients vary across individuals (Diggle et al., 1994, 2002). In univariate models we saw that the variances for the random intercept and slope were statistically significantly different from zero indicating that there is a between-subject variation. That variation does not only exist at baseline but it exists over time as well. So the random intercept and slope will be included in this model.

The sample variogram in Figure 2.21 showed that we have random effects, measurement error and serial correlation, but the question of which one is the main source of variation between random effects or serial correlation is an important topic in longitudinal data analysis. The within-subject variability which is directly related to the spacing of measurements is modelled by the covariance structure in Σ_i matrix via the REPEATED statement in SAS. We want to see if we can achieve similar results as in section 4.2 if we take into account both the within and between subject variation. That way each block $Z_i G Z_i'$ in equation (3.12) contributes the within subject correlation which varies with time (Hallahan, 2003) because the main covariate in Z_i is time. The full model is given by:

$$\begin{aligned} Y_{ij} = \beta_0 + \beta_1 S_i + \beta_2 G_i + \beta_3 A_i + \beta_4 L_{ij} + \beta_5 W_{ij} + \beta_6 t_{ij} + b_{0i} + \\ (\beta_7 S_i + \beta_8 G_i + \beta_9 A_i + \beta_{10} L_{ij} + \beta_{11} W_{ij} + b_{1i}) t_{ij} + \varepsilon_{ij} \end{aligned} \tag{4.15}$$

S_i, G_i, A_i, L_{ij} and W_{ij} are defined in section 4.2. b_{0i} and b_{1i} are the random intercept and slope respectively. This model will be fitted under REML estimation first so that we can be able to compare it's AIC with the one that we used to select the spatial linear structure in section 4.3.1. The SAS program to fit this model is:

```
proc mixed data=newdata method=reml covtest noclprint empirical;
class pid sex site visits;
model sqrtcd4= site sex age logv weight visit visit*site visit*sex
visit*age visit*logv visit*weight / solution ;
random intercept visit /type=un subject=pid;
repeated visits/type=sp(lin)(visit) local subject=pid ;
title' Longitudinal model with random effects and serial correlation';
run;
```

In model (4.15), Σ_i is modelled as spatial linear covariance structure and G is modelled as an unstructured covariance structure. The covariance parameter estimates showed that there is no

correlation between the random intercept and slope, so the G matrix was then modelled as variance component and the covariance parameter estimates under REML are shown in Table 4.35.

Table 4.35: Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error	Z value	Pr Z
Intercept	5.7685	0.4827	11.95	<0.0001
Visit	0.4502	0.0579	7.77	<0.0001
Variance	3.3889	0.4430	7.65	<0.0001
SP(LIN)	0.3690	0.0272	13.55	<0.0001
Residuals	2.5380	0.2243	11.32	<0.0001

The AIC for this model is 19415.3 which is smaller than 19451.5 that we used when selecting the spatial linear covariance structure in section 4.3.1 under the marginal model. Table 4.35 shows that despite having the within subject variation we also have the between subject variation. Parameter estimates for the spatial linear structure are less than those in Table 4.28 because the variation is now divided into two components, the within and between subject variation. In Table 4.28 we only allowed for within subject variation and the variation between subjects was ignored. This emphasizes that the two models in equations (4.13) and (4.15) are conceptually different. 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 shown in Tables 4.36, 4.37,4.38.

Table 4.36: Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error	Z value	Pr Z
Intercept	5.7460	0.4797	11.98	<0.0001
Visit	0.4458	0.0574	7.76	<0.0001
Variance	3.3671	0.4412	7.63	<0.0001
SP(LIN)	0.3692	0.0234	13.49	<0.0001
Residual	2.5401	0.2239	11.34	<0.0001

Table 4.37: Solution for Fixed Effects

Effect	Sex	Estimate	Standard Error	DF	t value	Pr>|t|
Intercept		11.5967	0.7365	1108	15.75	<0.0001
Site	eThekwini	0.1615	0.2167	1819	0.75	0.4563
Sex	Female	0.6000	0.2340	1819	2.56	0.0104
Age		0.0032	0.0127	1819	0.25	0.7994
Logv		-1.1376	0.0482	1819	-23.58	<0.0001
Weight		0.0567	0.0081	1819	7.03	<0.0001
Visit		2.0524	0.2725	953	7.53	<0.0001
Visit*Site	eThekwini	-0.1337	0.0776	1819	-1.72	0.0853
Visit*Sex	Female	0.2327	0.0883	1819	2.64	0.0085
Visit*Age		-0.0206	0.0051	1819	-3.99	<0.0001
Visit*Weight		-0.0021	0.0028	1819	-0.76	0.4501
Visit*Logv		-0.0071	0.0393	1819	-0.18	0.8558

Table 4.38: Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F value	Pr>F
Site	1	1819	0.56	0.4563
Sex	1	1819	6.57	0.0104
Age	1	1819	0.06	0.7994
Logv	1	1819	556.07	<0.0001
Weight	1	1819	49.43	<0.0001
Visit	1	953	62.50	<0.0001
Visit*Site	1	1819	2.96	0.0853
Visit*Sex	1	1819	6.95	0.0085
Age*Visit	1	1819	15.92	<0.0001
Weight*Visit	1	1819	0.57	0.4501
Logv*Visit	1	1819	0.03	0.8558

The variables site, age, and the interaction terms visit*site, visit*weight and visit*logv are statistically insignificant just like under the marginal model in section 4.3.1. The terms visit*weight and visit*logv will be removed from the model starting with visit*logv because it is the most insignificant effect. The model was fitted after removing the interaction term visit*log and the AIC dropped from 19387.5 to 19376.6. The model was fitted again and the term visit*weight was still insignificant with a p-value of 0.4650 thus finally removed from the model. After removing it, the AIC dropped from 19376.6 to 19375.1. The final model was fitted using the REML algorithm and

the results are shown in Tables 4.39, 4.40 and 4.41.

Table 4.39: Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error	Z value	Pr Z
Intercept	5.7904	0.4820	12.01	<0.0001
Visit	0.4501	0.0576	7.82	<0.0001
Variance	3.3686	0.4420	7.62	<0.0001
SP(LIN)	0.3697	0.0274	13.51	<0.0001
Residual	2.5408	0.2239	11.35	<0.0001

The AIC for this model is 19400.9 which is a considerable reduction compared to 19438, the AIC for the marginal model with the exact number of fixed effects in section 4.3.1. However this statement is made with caution depending on the intended scientific question to answer given that the two models are as stated before conceptually different. Significant variance estimates for the random effects associated with the intercept and linear time effect suggests that the intercepts and slopes vary across subjects.

Table 4.40: Solution for Fixed Effects

Effect	Sex	Estimate	Standard Error	DF	t value	Pr>|t|
Intercept		11.8134	0.6836	1108	17.28	<0.0001
Site	eThekwini	0.1731	0.2169	1821	0.80	0.4251
Sex	Female	0.6068	0.2336	1821	2.60	0.0095
Age		0.0039	0.0126	1821	0.31	0.7459
Logv		-1.1476	0.0444	1821	-25.88	<0.0001
Weight		0.0534	0.0067	1821	8.00	<0.0001
Visit		1.9063	0.2037	953	9.36	<0.0001
Visit*Site	eThekwini	-0.1434	0.0774	1821	-1.85	0.0641
Visit*Sex	Female	0.2248	0.0866	1821	2.60	0.0095
Visit*Age		-0.0208	0.0051	1821	-4.10	<0.0001

The inferences for the random effects model are similar to those of the marginal model in terms of magnitude and direction. However, according to the AIC the random effects model is better than the marginal model. The average intercepts for the eThekwini and Vulindlela populations are not statistically different with eThekwini having a slightly higher intercept of 0.1731 units greater than that of Vulindlela. Their site specific average slopes are also not significantly different.

Table 4.41: Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F value	Pr>F
Site	1	1821	0.64	0.4251
Sex	1	1821	6.75	0.0095
Age	1	1821	0.10	0.7559
Logv	1	1821	669.52	<0.0001
Weight	1	1821	64.06	<0.0001
Visit	1	953	110.47	<0.0001
Visit*Site	1	1821	3.43	0.0641
Visit*Sex	1	1821	6.74	0.0095
Visit*Age	1	1821	16.84	<0.0001

The intercept for females is 0.6068 greater than that for males and they are statistically significantly different. Age is not significant at baseline, this means that there was no statistically significant difference in CD4+ count at baseline for younger and older patients but the difference was observed in the follow up visits. Log viral load has a significant negative relationship with CD4+ count and that effect is visible in the exploratory analysis plot in Figure 2.15. The model here accounts for all the three sources of variability and correlation therefore the standard errors based on the current model are less conservative hence more reliable than when only serial and measurement error is accounted for in the evolution of CD4+ count post HAART.

4.4 Modelling baseline log viral load

In Section 4.3.1 and 4.3.2 one of our explanatory variables was log viral load. As it has been mentioned before, the viral load is collected on 6 monthly basis together with the CD4+ count. Patients started HAART with different viral load values, however most viral load measurements reached the undetectable level (threshold) as early as month 6 post-HAART initiation. The assay that was used in the laboratory was not able to detect any viral load less than 400 copies/ml. When the viral load is undetectable the laboratory result is stated as 400 copies/ml. So because of this, most viral load values are the same at month 6, 12 ,18 and 24. Figure 4.4 is a plot of the mean log viral over time for all the patients.

It is clear from Figure 4.4 that there is not much variation in measurements from month 6 to month 24. In this supplementary analysis, only the baseline log viral load will be included as a covariate in the model. We thought that it might be a good idea to fit a model where we control for initial log viral load for each patient. We want to see if the baseline log viral load can improve the prediction of CD4+ count post HAART.

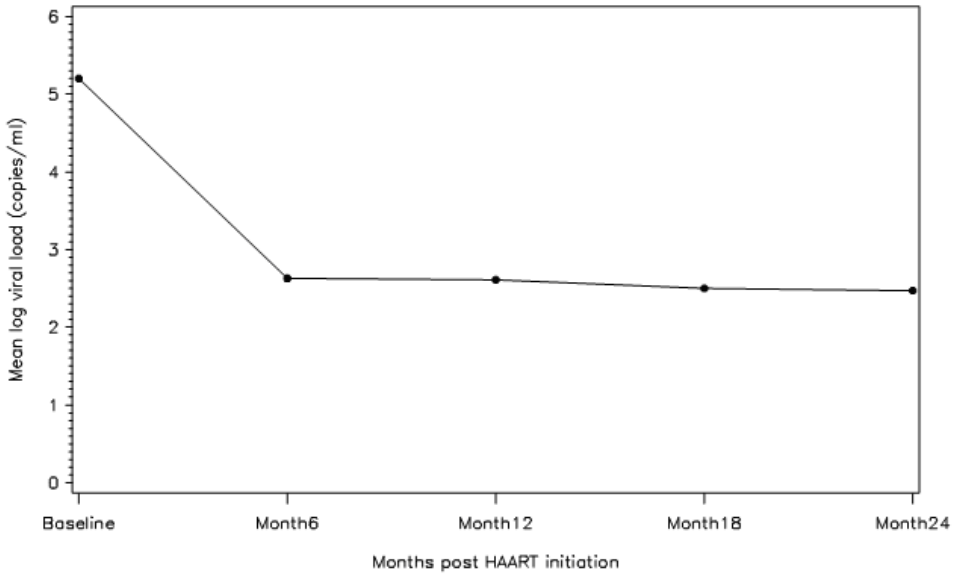

Figure 4.4 Mean log viral load over time

4.4.1 Marginal model

In order to fit the marginal model we will use the same procedure as we did in section 4.3.1. We will evaluate the best covariance structure for the Σ matrix, then fit the model using the selected covariance to determine the mean structure. We will also fit the random effects model. The model is given by:

$$Y_{ij} = \beta_0 + \beta_1 S_i + \beta_2 G_i + \beta_3 A_i + \beta_4 L_i + \beta_5 W_{ij} + \beta_6 t_{ij} + (\beta_7 S_i + \beta_8 G_i + \beta_9 A_i + \beta_{10} L_i + \beta_{11} W_{ij}) t_{ij} + \varepsilon_{ij} \quad (4.16)$$

where L_i is the baseline log viral load for the i^{th} patient. The definition for other variables is the same as in the previous sections. The fit statistics for each covariance structure is shown in Figure 4.5.

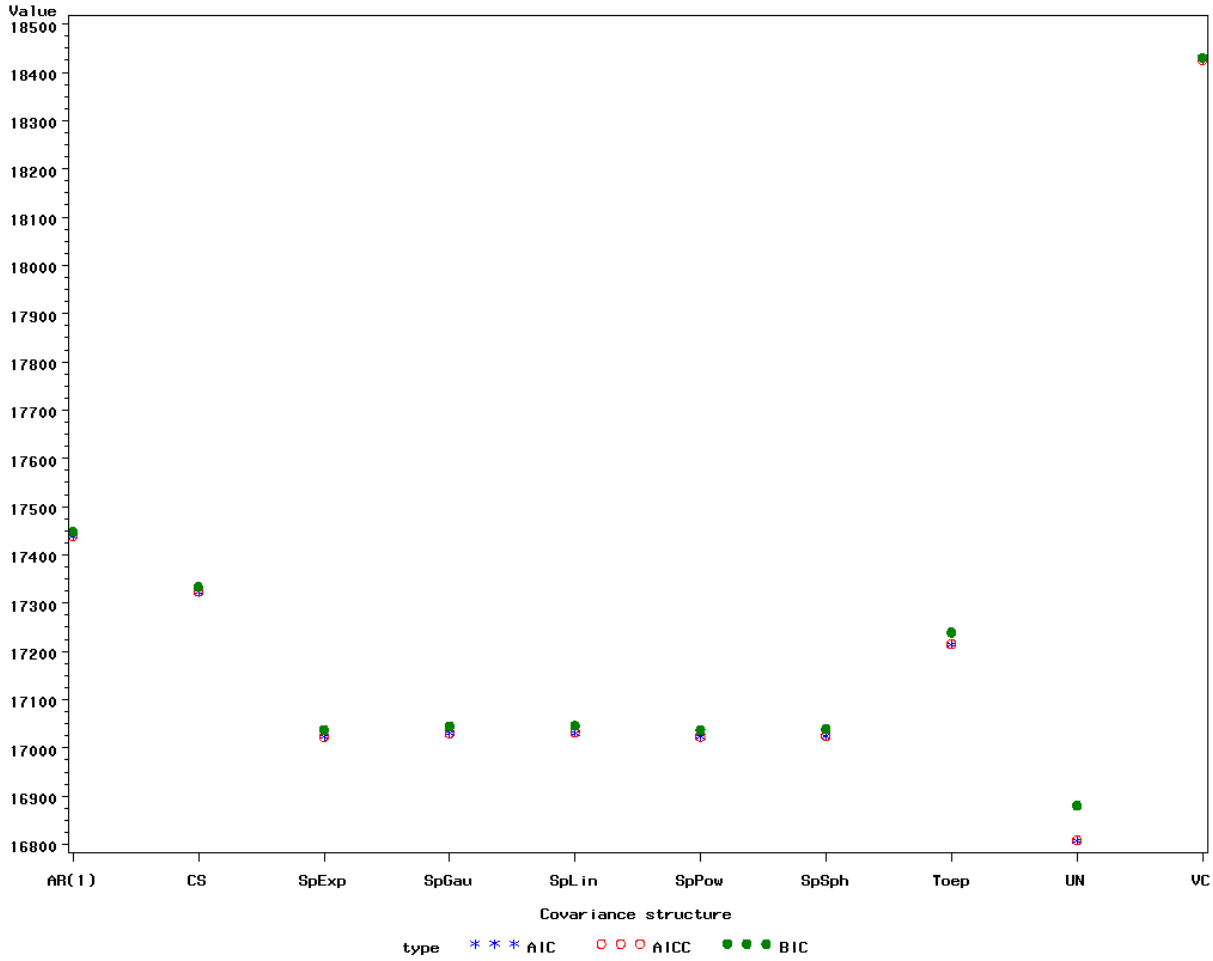

Figure 4.5 Model fit for each covariance structure

From Figure 4.5 we see that the UN and all the spatial covariance structures are the superior structures compared to CS, AR(1), Toeplitz and VC. The UN structure is too parametric (i.e. many parameters) and also lacks structure such as the ability to accommodate serial correlation. Among the spatial family the structure with the smallest AIC was the spatial power and also the exponential spatial both with an AIC of 17023.4. The results are shown in Table 4.42, 4.43 and 4.44. The spatial power structure was adopted.

If one compares these results to what we have in section 4.3.1 and 4.3.2, we note that the intercept for the Vulindlela site is now bigger than that for the eThekwini site. However, the two intercepts are still not statistically different from each other. We also note that interaction term visit*weight is now statistically significant. The interaction term visit*baselogv is also statistically significant in contrast to visit*logv in section 4.3.1 and 4.3.2 which was not statistically significant. It is

Table 4.42: Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error	Z value	Pr Z
Variance	14.8191	0.7111	20.84	<0.0001
SP(POW)	0.7219	0.0233	31.03	<0.0001
Residuals	0.7834	0.4557	1.72	0.0428

Table 4.43: Solution for Fixed Effects

Effect	Sex	Estimate	Standard Error	DF	t value	Pr>|t|
Intercept		4.5508	1.1889	871	3.83	0.0001
Site	eThekwini	-0.3684	0.2437	871	-1.51	0.1311
Sex	Female	0.6107	0.2628	871	2.32	0.0204
Age		0.0006	0.0145	2421	-0.04	0.9682
Baselogv		-0.4545	0.1497	871	-3.04	0.0025
Weight		0.1277	0.0100	2421	12.75	<0.0001
Visit		3.0581	0.4500	2421	6.80	<0.0001
Visit*Site	eThekwini	-0.1184	0.0917	2421	-1.29	0.1968
Visit*Sex	Female	0.2262	0.1045	2421	2.16	0.0305
Visit*Age		-0.0150	0.0059	2421	-2.53	0.0114
Visit*Weight		-0.0213	0.0035	2421	-6.14	<0.0001
Visit*Baselogv		0.1362	0.0629	2421	2.16	0.0306

obvious that the interaction term visit*weight is affected by the presence of log viral load as a time dependent covariate. In order to check if this was true we fitted model (4.13) in section 4.3.1 without the covariate logv and the interaction term visit*logv. Under this modification the interaction term visit*weight was found to be significant. We also swapped roles by removing weight and visit*weight terms to check if visit*logv is going to be significant and interestingly it was not. So it is only weight that is affected by log viral load and not vice versa. There are no fixed effects to be removed from this model.

Table 4.44: Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F value	Pr>F
Site	1	871	2.28	0.1311
Sex	1	871	5.40	0.0204
Age	1	2421	0.00	0.9682
Baselogv	1	871	9.22	0.0025
Weight	1	2421	162.60	<0.0001
Visit	1	2421	49.81	<0.0001
Visit*Site	1	2421	1.67	0.1968
Visit*Sex	1	2421	4.68	0.0305
Visit*Age	1	2421	6.42	0.0114
Visit*Weight	1	2421	37.69	<0.0001
Visit*Baselogv	1	2421	4.68	0.0306

4.4.2 Random effects model

We will now fit the random effects model where we take into account the random effects namely the random intercept and random slope. The model is given by:

$$Y_{ij} = \beta_0 + \beta_1 S_i + \beta_2 G_i + \beta_3 A_i + \beta_4 L_i + \beta_5 W_{ij} + \beta_6 t_{ij} + b_{0i}$$
$$(\beta_7 S_i + \beta_8 G_i + \beta_9 A_i + \beta_{10} L_i + \beta_{11} W_{ij} + b_{1i}) t_{ij} + \varepsilon_{ij} \tag{4.17}$$

We will use the spatial power correlation structure from Section 4.4.1 to model the within subject serial correlation as well as the UN for the between subject variability. The results for the covariance parameter estimates are shown in Tables 4.45.

Table 4.45: Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error	Z value	Pr Z
UN(1,1)	0	.	.	.
UN(2,1)	0.8758	0.1771	4.95	<0.0001
UN(2,2)	0.0249	0.0911	0.27	0.3924
Variance	12.3468	0.7700	16.04	<0.0001
SP(POW)	0.6083	0.0387	15.74	<0.0001
Residuals	0.0890	0.6146	0.14	0.4425

The AIC for this model 16984.5. The model was fitted and the covariance parameter estimate for the random intercept UN(1,1) was found to be zero and so there is no variability in random intercepts. One can also see that the variance for the random slopes is also insignificant. Since we are interested in serial correlation we are going to try other spatial covariance structures. Verbeke and Molenberghs (2000) argued that if one is interested in the serial correlation function, it is

usually sufficient to fit and compare a series of serial correlation models. The other option was to reduce the covariance parameters. The random intercept was removed to see whether the variance for the slope becomes significant. The model was then fitted without the random intercept and the results are shown in Table 4.46.

Table 4.46: Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error	Z value	Pr Z
Visit	0.3335	0.0752	4.44	<0.0001
Variance	12.4426	0.7974	15.60	<0.0001
SP(POW)	0.7116	0.0301	23.61	<0.0001
Residuals	1.1760	0.5123	2.30	0.0109

The AIC for this model is 17001.6 which is greater than 16984.5 from the model where we have both the random intercept and slope. The next step was to test whether dropping the random intercept makes the model better than having both the random intercept and slope. The likelihood ratio test was performed which subsequently gave us a p-value <0.0001 with 1 degree of freedom. Therefore we reject the hypothesis and conclude that the model with the slope only is not better than the model with both the random intercept and slope. So we have decided to try each and everyone of the spatial covariance structure and see if all of them give the same results because of the strong belief that there is a between subject variability and it's existence is from baseline and throughout all the other visits.

The summary for each of the spatial covariance structures is provided as follows: The Gaussian, exponential and spherical structures show that there is no variability in random intercepts and slopes. The linear structure on the other hand shows that there is variability in random intercepts and slopes. So the random effects model will be fitted using the spatial linear covariance structure for the within subject variation and the UN for the between subject variation. The covariance parameter estimates are shown in Table 4.47.

Table 4.47: Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error	Z value	Pr Z
UN(1,1)	2.8993	0.8477	3.42	0.0003
UN(2,1)	0.6489	0.2232	2.91	0.0037
UN(2,2)	0.1842	0.1057	1.74	0.0406
Variance	7.2375	0.7510	9.64	<0.0001
SP(LIN)	0.3745	0.0156	23.94	<0.0001
Residuals	2.1852	0.2995	7.30	<0.0001

The AIC for this model is 16992.0 including variability in random intercepts and slopes. The rest of the results are shown in Table 4.48 and 4.49.

Table 4.48: Solution for Fixed effects

Effect	Sex	Estimate	Standard Error	DF	t value	Pr>|t|
Intercept		5.5278	1.1402	872	4.85	<0.0001
Site	eThekwini	-0.2695	0.2412	1662	-1.12	0.2639
Sex	Female	0.5975	0.2614	1662	2.29	0.0224
Age		0.0049	0.0143	1662	0.35	0.7296
Baselogv		-0.4908	0.1456	1662	-3.37	0.0008
Weight		0.1122	0.0093	1662	12.10	<0.0001
Visit		2.6575	0.4428	758	6.00	<0.0001
Visit*Site	eThekwini	-0.1345	0.0916	1662	-1.47	0.1419
Visit*Sex	Female	0.2382	0.1051	1662	2.27	0.0236
Visit*Age		-0.0161	0.0059	1662	-2.71	0.0067
Visit*Weight		-0.0162	0.0033	1662	-4.86	<0.0001
Visit*Baselogv		0.1652	0.0626	1662	2.64	0.0083

Table 4.49: Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F value	Pr>F
Site	1	1662	1.25	0.2639
Sex	1	1662	5.22	0.0224
Age	1	1662	0.12	0.7296
Baselogv	1	1662	11.37	0.0008
Weight	1	1662	146.31	<0.0001
Visit	1	758	39.13	<0.0001
Visit*Site	1	1662	2.16	0.1419
Visit*Sex	1	1662	5.13	0.0236
Visit*Age	1	1662	7.37	0.0067
Visit*Weight	1	1662	23.59	<0.0001
Visit*Baselogv	1	1662	6.97	0.0083

The results from the random effects model are also similar to the results from the marginal model especially in terms of direction, even though they slightly differ in magnitude. But if one were to compare the marginal and the random effects model in terms of the AIC, the random effects model has the smallest AIC and thus is the best model. As stated before, since we are dealing with a Gaussian response, relating the marginal and random effects model is a straight forward exercise but caution is needed when dealing with a non-Gaussian response.

In this model we have 3 significant interaction terms between two continuous variables and it is a good idea to interpret them graphically as it has been mentioned before in section 4.2.1. However, it should be noted that a positive rate of change (or increase) in CD4+ count is associated with sex specifically females, younger baseline age, lower weight and higher baseline log viral load. A greater increase in CD4+ count with therapy has been associated with younger age and higher baseline viral load (Kaufmann et al., 2003). We will only plot baseline log viral and weight because age was plotted in section 4.2.1 and the estimates are still following the same direction. The graph showing the interaction between time and the two continuous variables namely baseline log viral load and weight are shown in Figure 4.6 and 4.7 respectively.

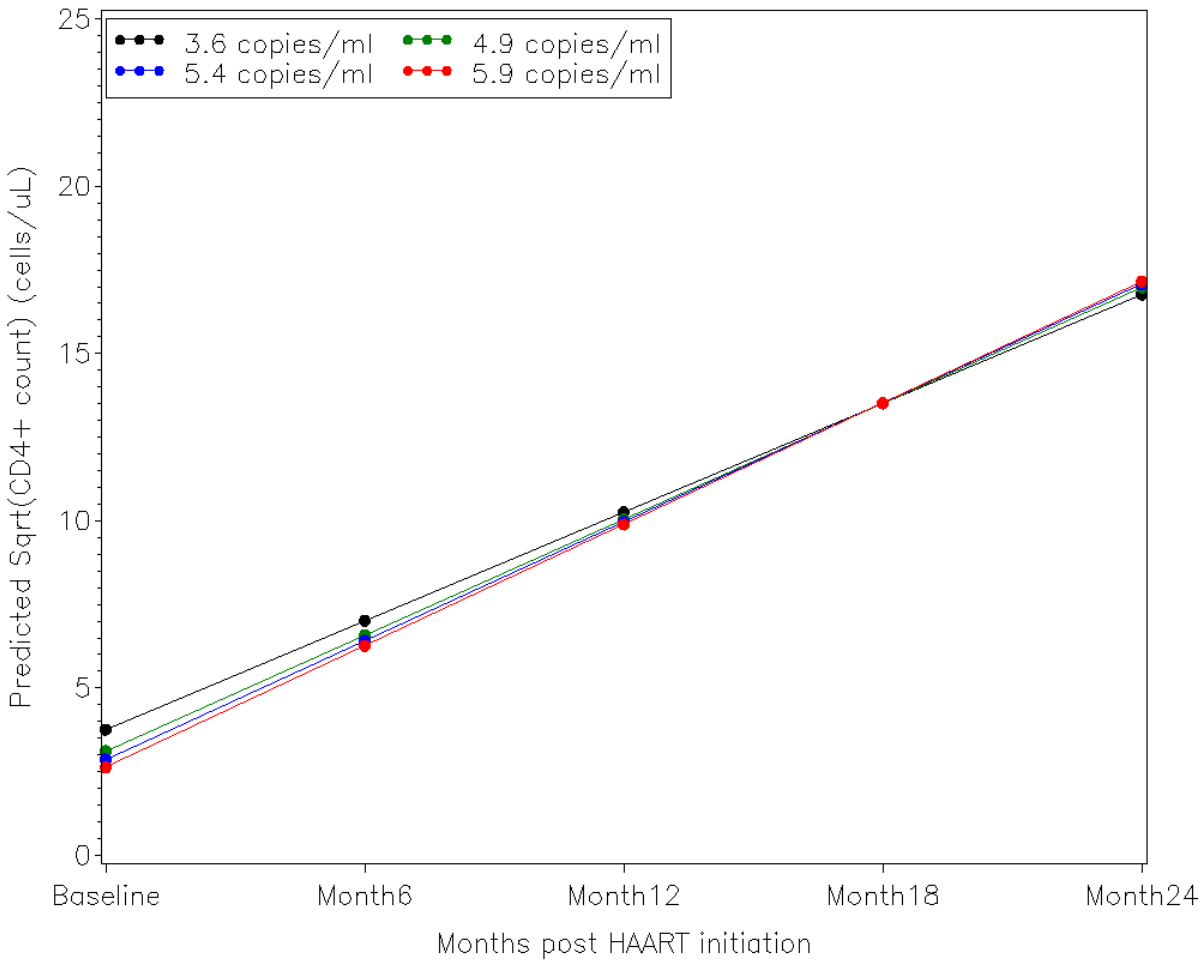

Figure 4.6 Interaction between time and baseline log viral load (copies/ml)

The estimate for the interaction term visit*baselogv in Table 4.48 is 0.1652 and it tells us that the rate of change in square root CD4+ count for patients who started HAART with high baseline log viral load is greater than for those who started with low baseline log viral load. Figure 4.6 also gives us the same results.

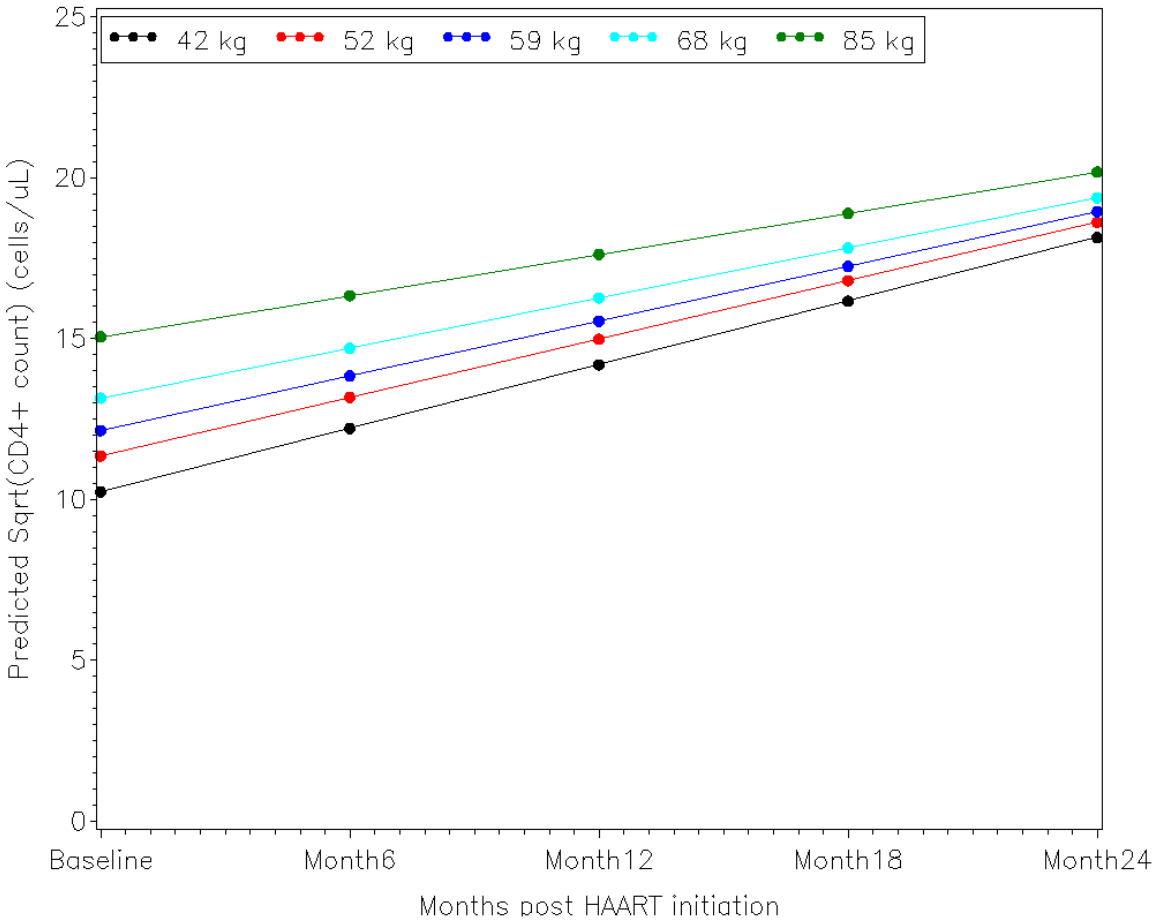

Figure 4.7 Interaction between time and weight (kg)

Figure 4.7 shows that the rate of change in square root CD4+ count for patients who started HAART with low weight is greater than for those who started with higher weight. This is in conjunction with the estimate for visit*weight which is -0.0162.

4.5 Model diagnostics

Residuals are frequently used to evaluate the validity of the assumptions of statistical models and may also be employed as tools for model selection (Nobre and da Motta Singer, 2007). There are 3 types of residuals that accommodate the extra source of variability present in linear mixed models (Nobre and da Motta Singer, 2007). They are:

i. Marginal residuals given by, $\hat{\varepsilon} = y - X\hat{\beta}$

ii. Conditional residuals given by, $\hat{\varepsilon} = y - X\hat{\beta} - Z\hat{b}$

iii. The best linear unbiased predictor (BLUP), $Z\hat{b}$

Each type of the above residual is useful to evaluate the assumptions of the linear mixed model (3.9). The residuals in (i) are used to assess the fixed effects specification of the model. The subject specific residuals in (ii) are used to assess unusual or outlying subjects and whether the random effects were selected properly. The subject specific residuals should be small with a correct choice of random effects (Hallahan, 2003). Now that we have fitted the multi-covariate models in sections 4.3 and 4.4, it is important to assess if the normality assumptions for residuals and random effects were not violated. In section 4.3 the random effects model was chosen over the marginal model on the basis of the AIC. Here we will assess if the chosen models do not violate the normality assumptions.

One should recall that residuals and random effects are assumed to be normally distributed. It has been discussed in section 3.5.1 that fixed effects parameter estimates and standard errors are robust with respect to mis-specification of the random effects distribution, which follows from the theory of generalized estimating equations (Liang and Zeger, 1986). However, the violation of the normality assumption does affect the parameter estimates and standard errors of the random effects. First we plot the residuals as well as the residuals against predicted values for the marginal and random effects models in section 4.3. We will also plot the random intercept and slope for the random effects model. These plots are shown in Figure 4.8, 4.9 and 4.10.

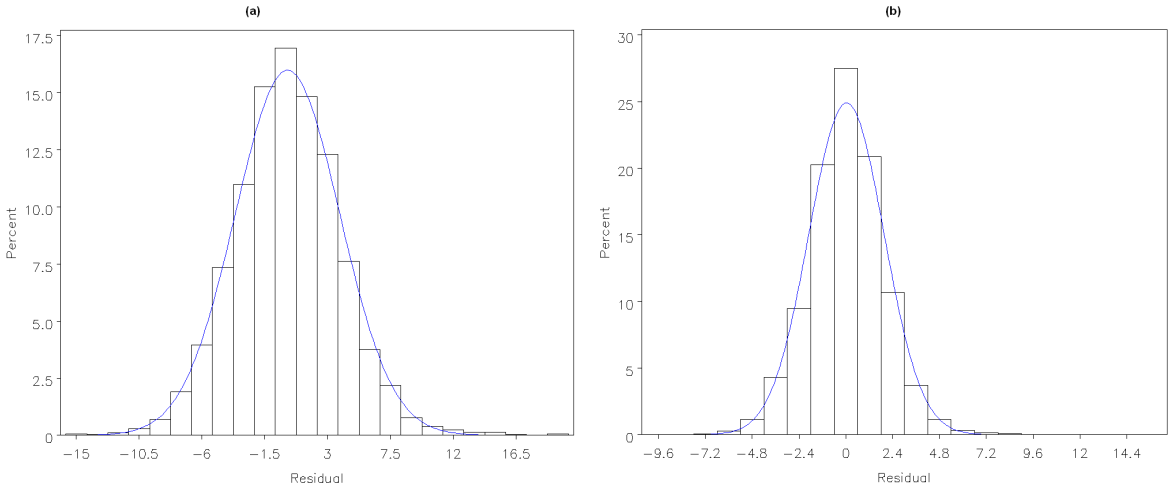

Figure 4.8 Residuals analysis

Figure 4.8 (a) and (b) are histogram plots of the residuals for marginal and random effects models respectively. Both graphs show that the residuals are normally distributed and thus one can infer that the normality assumption is not violated under either model assumption and specification.

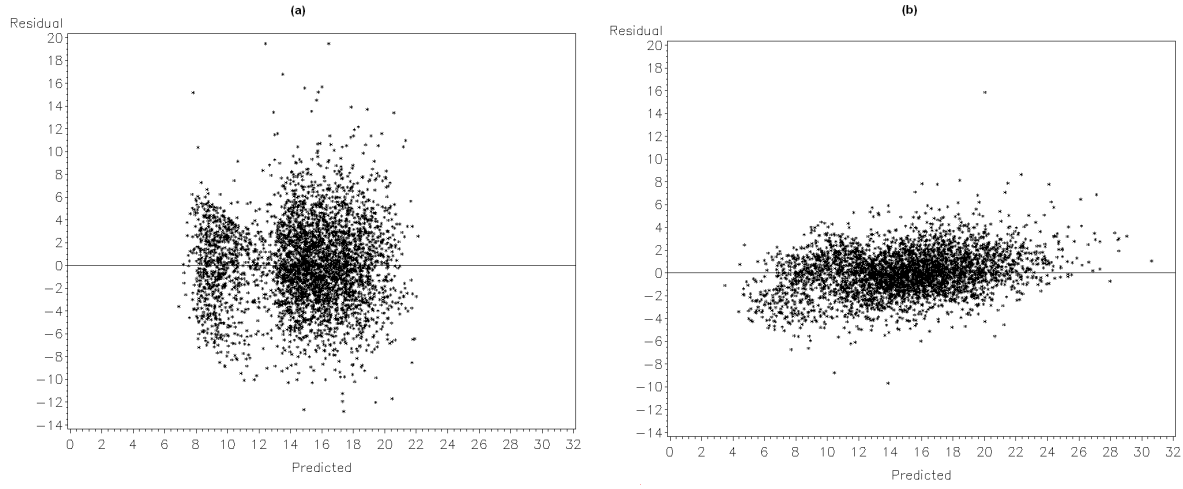

Figure 4.9 Residuals vs. predicted values

Figure 4.9 (a) and (b) respectively shows the graph of residuals plotted against the predicted values for the marginal and random effects models. Few large positive and negative residuals in Figure 4.9 (a) indicate possible outliers. The plot in Figure 4.9 (b) includes random effects and values seem to be scattered around 0 in comparison with Figure 4.9 (a).

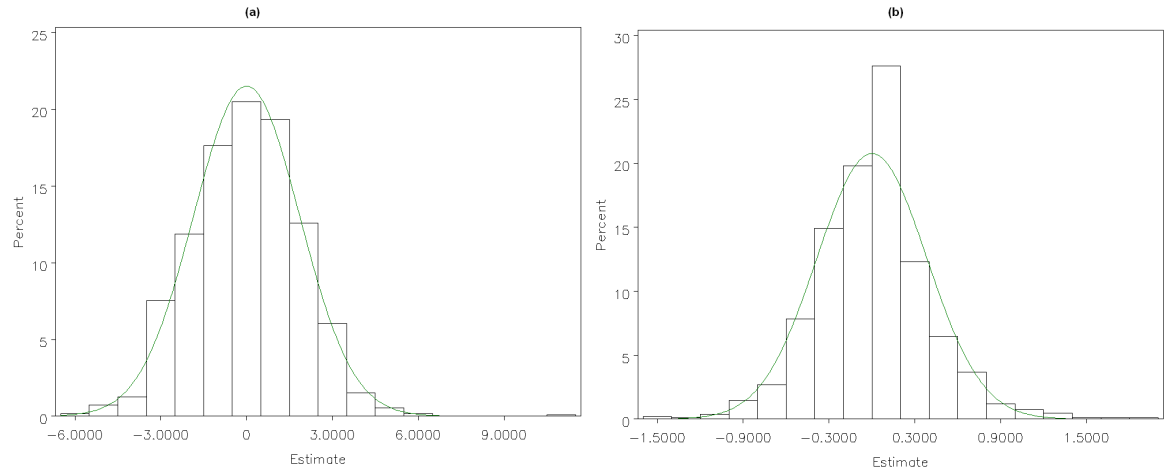

Figure 4.10 Distribution of random intercept and slope

Figure 10 (a) and (b) shows the distribution of random intercept and slope respectively. This graph shows that the normality assumption for both the random intercept and slope is not violated because histograms are indicative of a bell shaped distribution. The residual histogram plots for the final marginal and random effects models in this thesis discussed in section 4.4 are shown in Figure 4.11 to 4.13. One should note that the difference between the models in section 4.3 and 4.4 is that, in section 4.4 we used baseline log viral load as a covariate in the model. Whereas, in section 4.3 we used the log viral load at all visits.

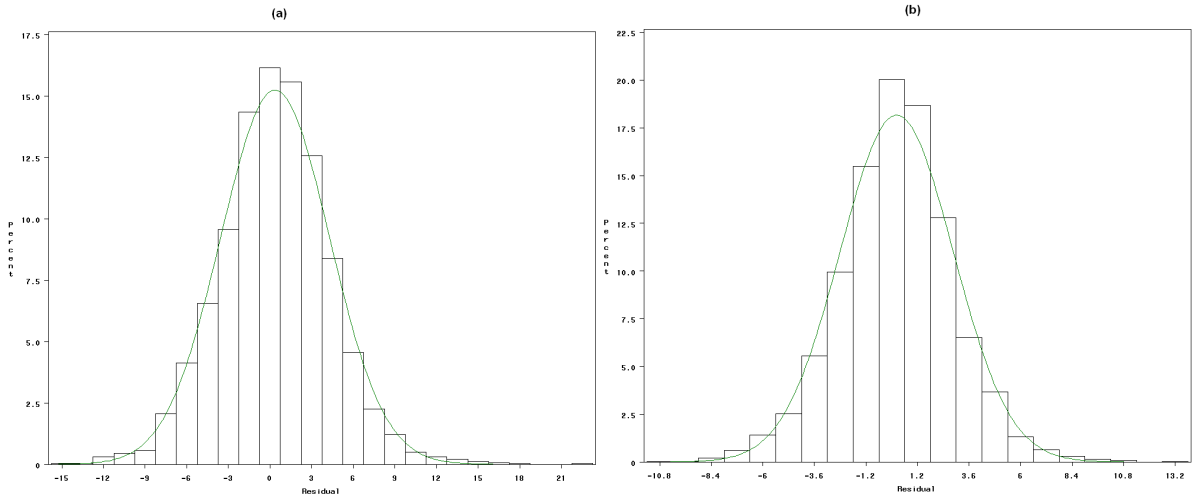

Figure 4.11 Residuals analysis

It is once again noted that, the normality assumption for residuals is not violated in both the marginal (Figure 4.11 (a)) and random effects model (Figure 4.11 (b)). There are no systematic

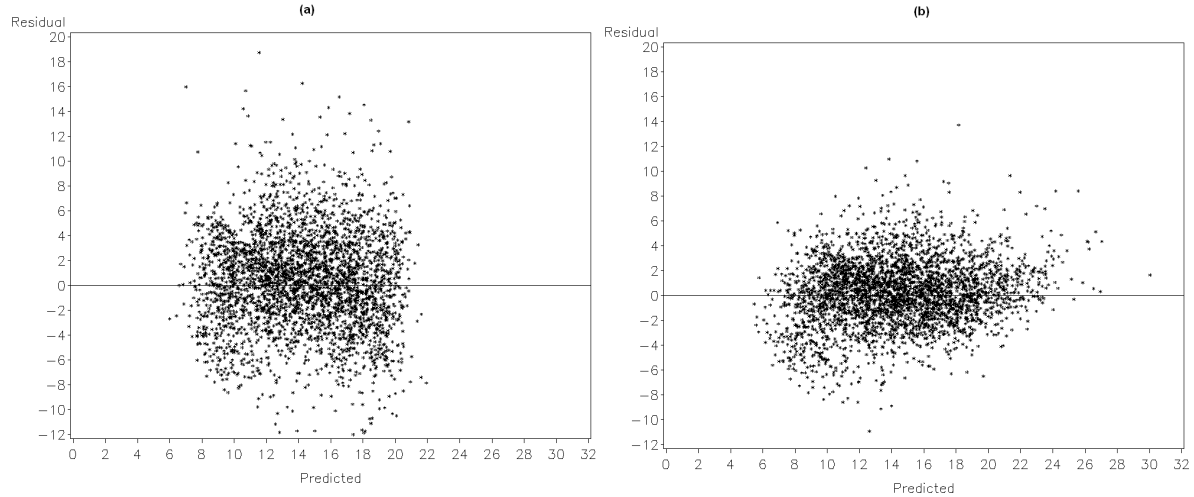

Figure 4.12 Residuals vs. predicted values

trends in the plot of residuals against predicted values for the marginal model in Figure 4.12 (a). Nonetheless the graph shows that there are outliers. The degree of outliers is substantially reduced when we consider the plots of residuals against predicted values for conditional model as seen in Figure 4.12 (b).

The normality assumption of random intercept and slope still holds and thus the inference for random effects is not affected. In section 4.3 and 4.4, the random effects models were chosen over the marginal models and the model diagnostics show that the normality assumption was

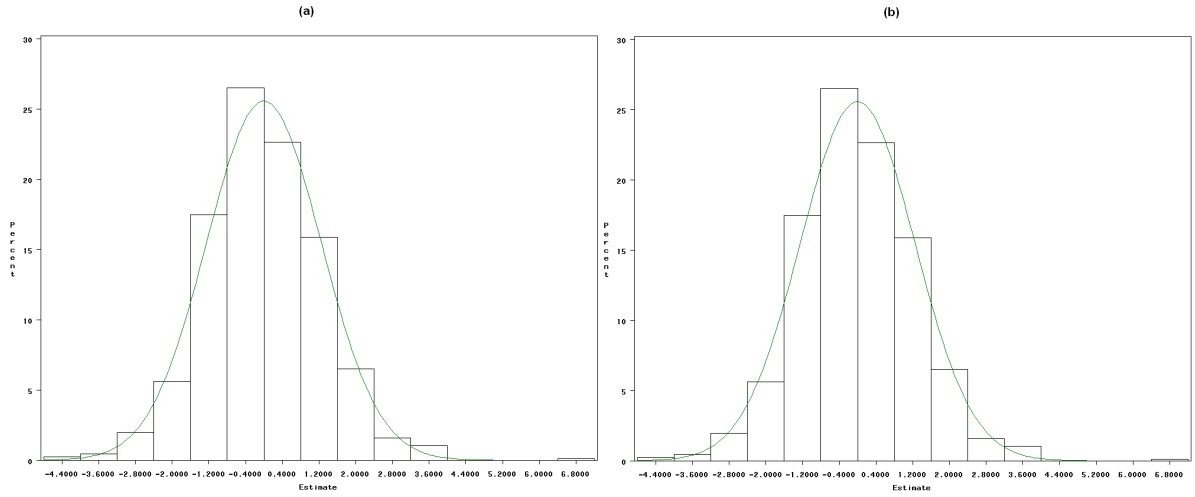

Figure 4.13 Distribution of random intercept and slope

not substantially violated in the chosen models. It should be pointed out that the final random effects model included the serial correlation in addition to measurement error. It has been noted by previous authors (Verbeke et al., 1998; Serroyen et al., 2009) that failure to include serial correlation when it exists can severely compromise the normality assumption of the model. Thus a key component in longitudinal data such as the one in the current study, it was necessary to include serial correlation to improve the fit of the final model.

Chapter 5

Missing data

5.1 Introduction

Most prospective studies including clinical trials use repeated measurements of laboratory markers to track disease progression and to evaluate new therapies (Touloumi et al., 2002). However, one major problem with such studies is that some patients end up having missing data due to reasons such as death, loss to follow up, and withdrawal. Missing data are data that the investigator intended to collect, but for one reason or another did not (Carpenter and Kenward, 2006) and it should be avoided if possible. Software for analyzing unbalanced longitudinal data, such as SAS Proc MIXED is now available to practitioners. These analysis tools are valuable in that they incorporate all the available information in the data and can even eliminate bias resulting from an analysis confined to the complete cases (Little, 1995). For data being studied in the current project, the possible reasons for missing data were due to death, loss to follow up, relocation, withdrawal and some patients are being transferred out to ARV clinics closer to their homes.

This means that for some patients we have the follow up data up to a certain point then they dropout. There are patients who also missed at least one or two follow up visits and attended their subsequent follow up visits and we refer to that kind of missing data as *intermittent missing data* (Carpenter and Kenward, 2006). It is essential to assess the pattern of missing data and apply the appropriate statistical analysis (Wisniewski et al., 2006) in order to adequately understand the problem at hand. To incorporate incompleteness into the modelling process, we need to understand the nature of the missing value mechanism and its implications for statistical inference (Molenberghs and Kenward, 2007).

5.2 Missing data processes

There are three possible missing data processes due originally to Rubin (1976) and later Little and Rubin (2002). A process is called missing completely at random (MCAR) when the probability of an observation being missing does not depend on observed or unobserved measurements. For example, in relation to the current data set, some observations could be missing due to technical and logistical issues in the laboratory. Or because some patients were unable to attend for some reason not related to his or her illness due to ARV's side effects (for example, family crisis).

However, many mechanisms that initially seem to be MCAR may turn out not to be. For example, a patient in a clinical trial may be lost to follow up after falling under a bus; however if it is a psychiatric trial, this may be an indication of poor response to treatment or due to worsening psychiatric condition. Likewise, if a response to a postal questionnaire is missing because the questionnaire was lost or stolen in the post, this may not be random but rather reflect the area in which the sorting office is located. If the assumption of data is MCAR, analysing only those with fully observed data gives sensible results but in practice trial data are rarely MCAR (Carpenter and Kenward, 2006). Assuming the data is MCAR means that there are no variables in the dataset which predict why the observation is missing.

After considering MCAR, we now look at data missing at random (MAR). Usually there is an association between the chance of patient withdrawal and observations, baseline and (in longitudinal follow-up) measurements prior to withdrawal. In this case, it is not sensible to include in the analysis only those with complete data (Carpenter and Kenward, 2006). For example, suppose that the patient's worse health at baseline is associated both with increased risk of withdrawal and poor response to HAART. In the current data the study did enroll very sick patients with very low CD4+ count and they also had opportunistic infections and they had difficulties coming to the clinic and subsequently died within 6 to 12 months of HAART initiation. Analysing data from the patients who remain up to the end of the trial will thus give an over optimistic view of HAART effect.

It is possible that patients responded well to treatment and were able to suppress the virus also withdrew because they felt better and did not see the point of going back to the clinic. Thus, if we can identify or account for those variables which are associated with an increased risk of withdrawal, we can carry out a sensible analysis and MCAR in that case is not sensible (Carpenter and Kenward, 2006). If we have a fully observed variables whose values affect the chance of seeing missing data, those missing data are not MCAR but MAR.

If data are neither MCAR nor MAR, then the data is missing not at random (MNAR) and this missing mechanism is termed non-ignorable. This means that even accounting for all the available observed information, the reason for observations being missing still depends on the unseen observations themselves. Carpenter and Kenward (2006) give the following example, assuming 12 week forced expiratory volume (FEV_1) was MCAR conditional on baseline FEV_1. Now suppose that even conditional on baseline the chance of seeing FEV_1 at 12 weeks depends on the value at 12 weeks, that is 12 week FEV_1 is MNAR. Unfortunately it is not easy to tell from the data at hand whether the missing observations are MCAR, MNAR or MAR although one can distinguish between MCAR and MAR. In our case the best assumption that we make about the CAT data that is used for this thesis is that it is *missing at random* (MAR). Dealing with the case of MNAR will require some untestable assumptions and this type of missing data process methods such as sensitivity analysis which is beyond the scope of the current work.

5.3 Missing data frameworks

Before discussing the missing data frameworks, we will look at the missing data terminology based on the standard framework of Rubin (1976) and Little and Rubin (2002). This terminology allows us to place formal conditions on the missing value mechanism which determine how the mechanism may influence subsequent inferences (Molenberghs and Kenward, 2007). Assume that for each independent unit i=1, ..., N we have measurements Y_{ij} where $j = 1, \ldots, n_i$. The outcomes are grouped into a vector Y_i=$(Y_{i1}, \ldots, Y_{in_i})'$. For each occasion j we define

$$R_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \text{ is observed.} \\ 0 & \text{otherwise.} \end{cases}$$

The missing data indicators R_{ij} are organized into a vector R_i of parallel structure to Y_i. Thus a prior Y_i can be partitioned into two subvectors such that Y_i^0 is the vector containing those Y_{ij} for which $R_{ij} = 1$, and Y_i^m contains the remaining components. These subvectors are referred to as the *observed* and *missing* components respectively (Molenberghs and Kenward, 2007). The *complete* data refers to the vector Y_i and this is the outcome vector that would have been recorded if no data had been missing. Thus the full data (Y_i, R_i) consist of the complete data, together with the missing data indicators.

There are three missing data frameworks that we are going to discuss. They are called selection, pattern-mixture and shared-parameter modelling frameworks. When the data are incomplete due to the operation of a random (missing value) mechanism the appropriate starting point for a model is the full data density

$$f(y_i, r_i | X_i, W_i, \theta, \psi) \tag{5.1}$$

where X_i and W_i denote design matrices for the measurements and missingness mechanism respectively. The corresponding parameter vectors are θ and ψ respectively. The selection model factorization is given by

$$f(y_i, r_i|X_i, W_i, \theta, \psi) = f(y_i|X_i, \theta)f(r_i|y_i, W_i, \psi), \tag{5.2}$$

where the first factor is the marginal density of the measurement process and the second one is the density of the missingness process, conditional on the outcomes. A pattern-mixture approach starts from the reverse factorisation. This factorization is given by

$$f(y_i, r_i|X_i, W_i, \theta, \psi) = f(y_i|r_i, X_i, \theta)f(r_i|W_i, \psi). \tag{5.3}$$

The pattern-mixture model allows for a different response model for each pattern of missing values, the observed data being a mixture of these weighted by the probability of each missing value or dropout pattern (Molenberghs and Kenward, 2007). In a shared-parameter model, a set of random effects b_i is assumed to drive both the Y_i and R_i processes. Thus the shared-parameter factorization is given by

$$f(y_i, r_i|X_i, W_i, \theta, \psi, b_i) = f(y_i|r_i, X_i, \theta, b_i)f(r_i|W_i, \psi, b_i). \tag{5.4}$$

A sensible assumption is that Y_i and R_i are conditionally independent, given the random effects b_i. Molenberghs and Kenward (2007) reported that the natural parameters of selection models, pattern-mixture models and shared-parameter models have different interpretations, and transforming one statistical model from one of the frameworks to another is generally not straightforward.

5.4 Methods for handling missing data

There are several commonly used approaches that one can use to handle missing data. These include complete cases (CC), last observation carried forward (LOCF) and marginal and conditional mean imputation. A complete case analysis excludes patients with missing observations. If the missingness mechanism is MCAR, a complete case analysis is sensible, although it may well not use all the available information in the data. However, if the missingness mechanism is not MCAR, complete case analysis is not sensible (Carpenter and Kenward ,2006).

The CC method suffers from several drawbacks such as the substantial loss of information leading to inefficient estimators (Molenberghs and Kenward, 2007). In LOCF, if a patient withdraws and their subsequent responses are missing then we set their missing responses equal to their last observed response. However Carpenter and Kenward (2006) argue that LOCF is not sensible when data is MCAR. Molenberghs and Kenward (2007) argue that in the LOCF very strong and often unrealistic assumptions have to be made to ensure the strong validity of this method.

One has to believe that a subject's measurements stay at the same level from the moment of drop out onwards. The marginal and conditional imputation, also known as marginal mean imputation replaces the missing observation by the average of the observed values for that variable and this is clearly problematic for categorical variables. However the marginal mean imputation ignores all the other variables in the data set, using it reduces the associations in the data set. Also, imputing all the missing observations to the same value is clearly wrong, and will underestimate the variability in the unseen data (Carpenter and Kenward, 2006).

5.5 Application to CAT study: Predictors of withdrawal

A sensible MAR analysis must condition or adjust for variables predictive of withdrawal (Carpenter and Kenward, 2006). Carpenter and Kenward (2006) also stated that some useful exploratory techniques are using t-tests or cross tabulations to investigate the association between baseline variables and withdrawal. It can also be useful to look, at each time point, whether there is a difference in response between patients who do, and do not, return for further visits. More formally, logistic regression and/or survival (withdrawal) analysis can be useful to establish key independent predictors of withdrawal. We then checked how many patients had CD4+ count measured at each visit and the results are shown in Table 5.1.

Table 5.1: No of patients remaining at each visit

Visit	eThekwini	Vulindlela	**Total**
Baseline	409	767	1176
Month 6	363	631	994
Month 12	340	573	913
Month 18	317	537	854
Month 24	299	501	800

Table 5.1 does not provide us with the number of patients we are left with at month 24, but the number of patients that had CD4+ count measured. There are patients who don't have CD4+ count measured at month 18 or month 24 but have other variables like viral load and weight measured and they obviously did not drop out. One can see that by month 24 there were only 800 (68.0%) patients with CD4+ count measured out of the 1176 we started with. The eThekwini site contributes 25.4% while the Vulindlela site contributes 42.6%. However, if one looks at how many patients each site has, the eThekwini site has 299 (73.1%) patients out of the 409 they started with. On the other hand, Vulindela has 501 (65.3%) patients out of the 767 they started with.

The next step was to check how many patients dropped out. Those are patients who did not attend all their visits consecutively. We found that out of the 1176 patients we started with, 341 (29.0%) dropped out. We will now establish the baseline independent predictors of withdrawal. We will start by performing Chi-square test on categorical variables and t-test on continuous variables. We will then structure a survival analysis model similar to the Cox proportional hazards model to identify significant predictors of withdrawal.

Table 5.2: Sex and withdrawal

	Withdrew	Not withdrew	p-value
Female	221 (27.3)	590 (72.8)	
Male	120 (32.9)	245 (64.1)	0.049

The p-value in Table 5.2 implies that there is an association or relationship between withdrawal and sex. The proportion of males who withdrew is greater than that for females.

Table 5.3: Site and withdrawal

	Withdrew	Not withdrew	p-value
eThekwini	95 (23.2)	314 (76.8)	
Vulindlela	246 (32.1)	521 (67.9)	0.002

There is also a relationship between site and withdrawal with Vulindlela having a higher withdrawal rate than eThekwini. For CD4+ count and weight we used the last observation prior to withdrawal for patients who withdrew and the month 24 CD4+ count for those who did not withdrew. However, for log viral load we used the baseline observation as there is no variation in log viral load after month 6 post HAART initiation. The mean CD4+ count for patients who withdrew and those who did not was 142 and 364.1 cells/μL respectively. These are the means at the measurement of occasion just prior to withdrawal. The p-value from the independent samples t-test shows that the two means are statistically significantly different (p=<0.0001).

The mean weight was also statistically significant (p=<0.0001) with patients who withdrew having the lowest mean of 59.6 kg as compared to those who did not withdraw with a mean weight of 68.8 kg. However, there was no significant difference between the mean age for the two groups (p=0.275). The mean log viral load for the two groups was also significantly different (p=0.037). So we will check if site, sex, CD4+ count, log viral load and weight are predictors of withdrawal using the survival analysis. The outcome of interest is the time to withdrawal and the Cox proportional hazard model gave the following results.

Table 5.4: Survival analysis

Parameter	Hazard ratio	95% C.I.	p value
Sex (ref=Female)	1.248	0.942 - 1.654	0.1230
Site(ref=eThekwini)	1.305	0.984 - 1.730	0.0642
CD4+ count	0.619	0.577 - 0.663	<0.0001
Log viral load	1.066	0.866 - 1.312	0.5476
Weight	0.983	0.972 - 0.994	0.003
Age	0.943	0.868 - 1.025	0.169

Weight and CD4+ count are significantly associated with withdrawal. Sex, site, age and log viral load are not significant. One should remember that the CD4+ count and viral load are highly negatively correlated. In order to avoid multicollinearity, we will remove one variable from the model and refit the model to see how this affects the model results. We started by removing log viral load from the model and the results are shown in Table 5.5.

Table 5.5: Survival analysis

Parameter	Hazard ratio	95% C.I.	p value
Sex (ref=Female)	1.283	1.007 - 1.635	0.044
Site(ref=eThekwini)	1.367	1.066 - 1.753	0.014
CD4+ count	0.615	0.580 - 0.652	<0.0001
Weight	0.986	0.978 - 0.995	0.003
Age	0.952	0.887 - 1.022	0.174

After removing the log viral load from the model, sex and site became significant. If we remove CD4+ count from the model the p-value for the log viral load becomes 0.043 which is different than 0.547 when both were in the model. However, sex and site became insignificant. Therefore the most plausible approach is to include either CD4+ count or log viral in the model but not both. Results 5.5 are similar to what we saw when testing the relationship between withdrawal and all other covariates univariately using cross-tabulation analysis. Since the CD4+ count is our outcome of interest we will use the model with CD4+ count only. The results in Table 5.5 tells us that for every 50 cells/μL increase in the last CD4+ count the probability of withdrawal decreases by 38.5%. Age is still not a predictor of withdrawal, but for every 5 year increase in age the probability of withdrawal decreases by 4.8%.

Males are more likely to withdraw than females whereas patients from Vulindlela are more likely to withdraw than those from the eThekwini site, that is holding other covariates fixed. So the predictors of withdrawal are site, sex, weight and CD4+ count. In the analysis in Chapter 4 we adjusted for all the variables predictive of withdrawal, so the assumption that we made about our data being MAR seems to be correct or sensible. In fact the use of proc MIXED with the repeated statement correctly, which is likelihood based, makes sure all the available data is subjected to analysis both complete and incomplete cases thus accounting for missing data under the MAR assumption (Mwambi et al., 2009).

Chapter 6

Discussion and conclusion

The exploratory data analysis shows that there is no difference in CD4+ count over time for patients in rural and urban site. However, we noted a significant difference in CD4+ count over time between males and females as well as between different age groups. Results from the linear mixed models, both marginal and random effects models showed no statistical difference between the eThekwini and Vulindlela site in terms of the CD4+ count improvement over time. There was no statistical difference in their intercepts as well as their slopes. On the other hand, there was a statistical significant difference between males and females with females having the higher rate of increase in CD4+ count over time.

It is difficult to point out whether the difference between males and females is attributable to adherence or biomedical factors. Patients of different age groups started with on average the same CD4+ count, but their rate of change in CD4+ count was statistically significantly different with the younger patients doing better than their older counterparts. The study conducted by Kaufmann et al. (2002) showed younger patients showed greater increase in CD4 cell count than older subjects. However, the rate of change in CD4+ count over time was the same whether patients started HAART with lower or higher weight. Similar results were observed for the log viral load as well. So the CD4+ count increase is associated with age and sex. Younger patients had the greatest CD4+ gain than older patients and females also had the greatest CD4+ increase than males. The interesting findings upon fitting the baseline log viral instead of the log viral at all visits was that, the interaction terms between time and the two variables weight and baseline log viral were now statistically significant. Using baseline log viral load in the model shows that the rate of change in CD4+ count over time was different for patients with lower weight compared to those with higher weight.

Similarly, there was a significant different rate of change in CD4+ count over time for patients who started with lower baseline log viral load compared to those with higher baseline log viral load. The study conducted by Kaufmann et al. (2003) showed that higher baseline viral load was significantly associated with larger increase in CD4+ count. So, fitting the baseline log viral in the model shows that the rate of change in CD4+ count is associated with sex, age, weight, baseline log viral and not site. A higher rate of change in CD4+ count is specifically associated with females, younger age, lower weight and higher baseline log viral load.

Having assumed that the data that we used for analysis is MAR, we found that sex, site, CD4+ count and weight are predictors of a patient withdrawing from the project. For future work, one might want to look at modelling CD4+ count over time without taking the square root. That way the CD4+ count may be allowed to follow a Poisson distribution and therefore generalized linear mixed models will be more appropriate instead of linear mixed models. Nonetheless square root CD4+ count and log viral load have become universally acceptable transformation both in the medical and statistical community such that the utility of count distributions such as the Poisson may not add more value. This is because the CD4+ count and viral load counts per unit of measurements are in large numbers requiring exact distributional analysis as would be in the case of sparse counts.

Future work possible includes full integration of missing data analysis in the evolution of success of HAART on HIV infected patients. One other exciting areas of advancement is to link the current analysis to a dynamic HIV/AIDS model and inform the process using the current data in estimating some key disease transition parameters. Incomplete data methodologies such as interval censored outcome can be used to address the question of non-adherence to scheduled and intermittent visits. It will also be a novel idea to compare different HAART studies in the context of meta analytic modelling approach in order to derive precise measures of post HAART treatment success to HIV infected patients.

References

1. Abdool Karim, S. S. and Abdool Karim, Q. (2005). *HIV/AIDS in South Africa*: Cambridge University Press.

2. Battegay, M., Nuesch, R., Hirschel, B. and Kaufmann, G. R. (2006). Immunological recovery and antiretroviral therapy in HIV-1 infection. *Lancet Infectious Diseases* **6**, 280-287.

3. Bucy, R. P., Hockett, R. D., Derdeyn, C. A., Saag, M. S., Squires, K., Sillers, M., Mitsuyasu, R. T. and Kilby, J. M. (1999). Initial increase in blood CD4(+) lymphocytes after HIV antiretroviral therapy reflects redistribution from lymphoid tissues. *Journal of Clinical Investigation* **103**, 1391-1398.

4. Burzykowski, T., Molenberghs, G. and Buyse, M. (2005). *The Evaluation of Surrogate Endpoints*. New York: Springer-Verlag.

5. Carpenter J.R., and Kenward M.G. (2006). *Missing Data in Clinical Trials-a Practical Guide*: UK National Health Service, National Co-ordinating Centre for Research on Methodology.

6. Department of Health (2006). Report National HIV and Syphilis Ante-Natal Sero-prevalence Survey in South Africa (2005).

7. Der, G. and Everitt, B. S. (2006). *Statistical Analysis of Medical Data using SAS*: Taylor & Francis Group, LLC.

8. Diggle, P. J., Zeger, S. L. and Liang, K. Y. (1994). *Analysis of Longitudinal Data*: University Press Inc., New York.

9. Diggle, P. J., Heagerty, P. J., Liang, K. Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data (2nd ed.)*: Oxford Science Publications. Oxford: Clarendon Press.

10. Douek, D. C., McFarland, R. D., Keiser, P. H., Gage, E. A., Massey, J. M., Haynes, B. F., Polis, M. A., Haase, A. T., Feinberg, M. B., Sullivan, J. L., Jamieson, B. D., Zack, J. A., Picker, L. J. and Koup, R. A. (1998). Changes in thymic function with age and during the treatment of HIV infection. *Nature* **396**, 690-695.

11. Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. (2004). *Applied Longitudinal Analysis*: John Wiley & Sons, Inc., Hoboken, New Jersey.

12. Gandhi, N. R., Moll, A., Sturm, A. W., Pawinski, R., Govender, T., Lalloo, U., Zeller, K., Andrews, J. and 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. *Lancet* **368**, 1575-1580.

13. Ghosh, D. (2008). Semiparametric inference for surrogate endpoints with bivariate censored data. Biometrics 64, 149-156.

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

15. Hedeker, D. (2004). *An introduction to growth modeling. In D. Kaplan (Ed.), Quantitative Methodology for the Social Sciences*: Thousand Oaks CA: Sage Publications.

16. Joint United Nations Programme on HIV/AIDS (UNAIDS)/WHO (2006). AIDS Epidemic Update.

17. Kaufmann, G. R., Bloch, M., Finlayson, R., Zaunders, J., Smith, D. and Cooper, D. A. (2002). The extent of HIV-1-related immunodeficiency and age predict the long-term CD4 T lymphocyte response to potent antiretroviral therapy. *AIDS* **16**, 359-367.

18. Kaufmann, G. R., Perrin, L., Pantaleo, G., Opravil, M., Furrer, H., Telenti, A., Hirschel, B., Ledergerber, B., Vernazza, P., Bernasconi, E., Rickenbach, M., Egger, M., Battegay, M. and S. H. C. S. Group (2003). CD4 T-lymphocyte recovery in individuals with advanced HIV-1 infection receiving potent antiretroviral therapy for 4 years - The Swiss HIV cohort study. *Archives of Internal Medicine* **163**, 2187-2195.

19. Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963-974.

20. Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

21. Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D. and Schabenberger, O. (2006). *SAS for Mixed Models: Second Edition*. Cary: SAS Institute Inc., Cary, NC, USA.

22. Little, R. J. A. (1995). Modelling the Drop-Out Mechanism in Repeated-Measures Studies. *Journal of the American Statistical Association* **90(431)**, 1112-1121.

23. Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data. (2nd edition)*. New York: John Wiley & Sons, Inc.

24. Malone, J. L., Simms, T. E., Gray, G. C., Wagner, K. F., Burge, J. R. and Burke, D. S. (1990). Sources of Variability in Repeated T-Helper Lymphocyte Counts from Human Immunodeficiency Virus Type 1-Infected Patients - Total Lymphocyte Count Fluctuations and Diurnal Cycle Are Important. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology* **3**, 144-151.

25. McCulloch, C. E., Searle, S. R. and Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models*: John Wiley & Sons, Inc.

26. Mocroft, A., Phillips, A. N., Gatell, J., Ledergerber, M. B., Fisher, M. M., Clumeck, N., Losso, M., Lazzarin, A., Fatkenheuer, G. and Lundgren, J. D. (2007). Normalisation of CD4 counts in patients with HIV-1 infection and maximum virological suppression who are taking combination antiretroviral therapy: an observational cohort study. *Lancet* **370**, 407-413.

27. Molenberghs, G. and Kenward, M. G. (2007). *Missing Data in Clinical Studies*: John Wiley & Sons, Ltd.

28. Molenberghs, G. and Verbeke, G. (2001). A review on linear mixed models for longitudinal data, possibly subject to drop out. *Statistical Modelling* **1(4)**, 235-269.

29. Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer Science+Business Media, Inc.

30. Mwambi, H., Ramroop, S., Shkedy, Z. and Molenberghs, G. (2009). A frequentist approach to estimating the force of infection and the recovery rate for a respiratory disease among infants in coastal Kenya. *Statistical Methods in Medical Research*, February 2009, doi: 10.1177/0962280208098666

31. Nattrass, N. (2008). Gender and Access to Antiretroviral Treatment in South Africa. *Feminist Economics* **14**, 19-36.

32. Nobre, J. S. and da Motta Singer, J. (2007). Residual analysis for linear mixed models. *Biometrical Journal* **49**, 863-875.

33. O'Brien, W. A., Hartigan, P. M., Martin, D., Esinhart, J., Hill, A., Benoit, S., Rubin, M., Lahart, C., Wray, N., Finegold, S. M., George, W. L., Dickinson, G. M., Klimas, N., Diamond, G., ZollaPazner, S. B., Jensen, P. C., Hawkes, C., Oster, C., Gordin, F., Labriola, A. M., Spivey, P., Matthews, T., Weinhold, K., Drusano, G. and Egorin, M. J. (1996). Changes in plasma HIV-1 RNA and CD4+ lymphocyte counts and the risk of progression to AIDS. *New England Journal of Medicine* **334**, 426-431.

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

35. Rubin, D. B. (1976). Inference and Missing Data. *Biometrika* **63**, 581-590.

36. Serroyen, J., Molenberghs, G., Aerts, M., Vloeberghs, E., De Deyn, P. P. and Verbeke, G. (2009). Flexible estimation of serial correlation in linear mixed models. *Jornal of Applied Statistics* **0**, 0-0.

37. Sherr, L., Lopman, B., Kakowa, M., Dube, S., Chariwa, G., Nyamukapa, N., Oberzaucher, N., Cremin, I. and Gregson, S. (2007). Voluntary counselling and testing: uptake, impact on sexual behaviour, and HIV incidence in a rural Zimbabwean cohort. *AIDS* **21**, 851-860.

38. Taylor, J. M. G. and Yu, M. G. (2002). Bias and efficiency loss due to categorizing an explanatory variable. *Journal of Multivariate Analysis* **83**, 248-263.

39. Tollerud, D. J., Clark, J. W., Brown, L. M., Neuland, C.Y., Pankiwtrost, L. K., Blattner, W. A. and Hoover, R. N. (1989). The Influence of Age, Race, and Gender on Peripheral-Blood Mononuclear-Cell Subsets in Healthy Nonsmokers. *Journal of Clinical Immunology* **9**, 214-222.

40. Touloumi, G., Pocock, S. J., Babiker, A. G. and Darbyshire, J. H. (2002). Impact of missing data due to selective dropouts in cohort studies and clinical trials. *Epidemiology* **13**, 347-355.

41. van Walraven, C., and Hart , R. G. (2008). Leave 'em Alone - Why Continuous Variables Should Be Analyzed as Such. *Neuroepidemiology* **30**:138-139.

42. Verbeke, G., Lesaffre, E. and Brant, L. J. (1998). The detection of residual serial correlation in linear mixed models. *Statistics in Medicine* **17**, 1391-1402.

43. Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data Analysis*. Springer-Verlag New York, Inc.

44. Wisniewski, S. R., Leon, A. C., Otto, M. W. and Trivedi, M. H. (2006). Prevention of missing data in clinical research studies. *Biol Psychiatry* **59**, 997-1000.