UNIVERSITY OF KWAZULU-NATAL

Masters Thesis

THE ROLE OF IMMUNE-GENETIC FACTORS IN MODELLING LONGITUDINALLY MEASURED HIV BIO-MARKERS INCLUDING THE HANDLING OF MISSING DATA

Author:
Nancy ODHIAMBO

Supervisor:
Prof. Henry MWAMBI

Co-Supervisor:
Dr. Thomas ACHIA

A thesis submitted in fulfilment of the requirements for the degree of Master of Science in statistics

Declaration of Authorship

I, Odhiambo Nancy, hereby declare that this thesis I am submitting is entirely my own original work except where otherwise indicated.

Signed: _____

Date: _____

Supervisor: _____

Signed: _____

Date: _____

(UNIVERSITY OF KWAZULU-NATAL)

Abstract

School of

Mathematics, Statistics and Computer Science

Master of Science

THE ROLE OF IMMUNE-GENETIC FACTORS IN MODELLING LONGITUDINALLY MEASURED HIV BIO-MARKERS INCLUDING THE HANDLING OF MISSING DATA

by Nancy ODHIAMBO

Since the discovery of AIDS among the gay men in 1981 in the United States of America, it has become a major world pandemic with over 40 million individuals infected world wide. According to the Joint United Nations Programme against HIV|AIDS epidermic updates in 2012, 28.3 million individuals are living with HIV world wide, 23.5 million among them coming from sub-saharan Africa and 4.8 million individuals residing in Asia. The report showed that approximately 1.7 million individuals have died from AIDS related deaths, 34 million ±50%know their HIV status, a total of 2.5 million individuals are newly infected, 14.8 million individuals are eligible for HIV treatment and only 8 million are on HIV treatment (Joint United Nations Programme on HIV |AIDS and health sector progress towards universal access: progress report, 2011).

Numerous studies have been carried out to understand the pathogenesis and the dynamics of this deadly disease (AIDS) but, still its pathogenesis is poorly understood. More understanding of the disease is still needed so as to reduce the rate of its acquisition. Researchers have come up with statistical and mathematical models which help in understanding and predicting the progression of the disease better so as to find ways in which its acquisition can be prevented and controlled.

Previous studies on HIV|AIDS have shown that, inter-individual variability plays an important role in susceptibility to HIV-1 infection, its transmission, progression and even response to antiviral therapy. Certain immuno-genetic factors (human leukocyte antigen (HLA), Interleukin-10 (IL-10) and single nucleotide polymorphisms (SNPs)) have been associated with the variability among individuals.

In this dissertation we are going to reaffirm previous studies through statistical modelling and analysis that have shown that, immuno-genetic factors could play a role in susceptibility, transmission, progression and even response to antiviral therapy. This will be done using the Sinikithemba study data from the HIV Pathogenesis Programme (HPP) at Nelson Mandela Medical school, University of Kwazulu-Natal consisting of 451 HIV positive and treatment naive individuals to model how the HIV Bio-markers (viral load and CD4 count) are associated with the immuno-genetic factors using linear mixed models.

We finalize the dissertation by dealing with drop-out which is a pervasive problem in longitudinal studies, regardless of how well they are designed and executed. We demonstrate the application and performance of multiple imputation (MI) in handling drop-out using a longitudinal count data from the Sinikithemba study with log viral load as the response. Our aim is to investigate the influence of drop-out on the evolution of HIV Bio-markers in a model including selected genetic factors as covariates, assuming the missing mechanism is missing at random (MAR). We later compare the results obtained from the MI method to those obtained from the incomplete dataset. From the results, we can clearly see that there is much difference in the findings obtained from the two

analysis. Therefore, there is need to account for drop-out since it can lead to biased results if not accounted for.

Acknowledgements

I would like to thank Prof.Henry Mwambi and Dr.Thomas Achia for the support and guidance they provided me with throughout this research period.

I also wish to extend my utmost gratitude to Prof.Thumbi Ndung'u for proving me with the data collected from the SK study under the HPP programme, without the data I would have done nothing constructive and recommendable.

I would also like to express my gratitude to Oscar, Ali and Artz for always being ready to answer my problems whenever would go to them.

Lastly, I thank all those who assisted, encouraged and supported me in one way or another during this research period and the Almighty God for giving me courage, strength and determination in conducting this research study despite the difficulties I passed through.

Contents

List of Figures

List of Tables

In loving memory of my dad, Henry Odhiambo Opiyo. Daddy thanks a lot for the support and encouragements you gave me through out this research period even though God took you away from us before you could witness the end of it.

Chapter 1

Introduction

This chapter gives a summary of HIV|AIDS, its discovery, how it's acquired, when and how it was discovered. It then illustrates some previous studies that have shown that AIDS is associated with the immuno-genetic factors. Then, discussing briefly about the immuno-genetic factors and the HIV Bio-markers, stating the problem and finally outlining the objectives for the study.

1.1 Background

Acquired Immune deficiency Syndrome (AIDS) was first reported in the United States of America in 1981 among the gay men. Since then it has become a major world pandemic with over 40 million individuals infected (Fauci et al., 2003). According to the Joint United Nations Programme against HIV|AIDS epidermic updates in 2012, 28.3 million individuals are living with HIV world wide, 23.5 million among them coming from sub-saharan Africa and 4.8 million residing in Asia. Approximately 1.7 million individuals have died from AIDS related deaths, 34 million $\pm 50\%$ know their HIV status, a total of 2.5 million individuals are newly infected, 14.8 million individuals are eligible for HIV treatment and only 8 million are on HIV treatment. The good news was that the number of new HIV infections globally in 2011 reduced by $700,000$ more compared to 2001 showing that Africa had cut AIDS related deaths by one third in the past six years (Joint United Nations Programme on HIV |AIDS and health sector progress towards universal access: progress report, 2011).

Previous studies on HIV|AIDS have shown that inter-individual variability plays an important role in susceptibility to HIV-1 infection, its transmission, progression and response to antiviral therapy. This is as a result of a number of factors such as: host

susceptibility, genetics and immune function among others. Even though those individuals who get exposed to the virus develops full blown AIDS at a certain instance, different individuals develop AIDS at different intervals and this has in part been associated with difference in genetic make-up among individuals (Chatterjee, 2010; Kaur and Mehra, 2009).

According to the Center for Disease Control and Prevention (CDC), genes do influence susceptibility to HIV infection and progression to AIDS. Previous studies have also shown that, the proportion of individuals with two copies (homozygous) of Chemokine receptor ($CCR5 - \triangle 32$) are rare and seem to have a strong protection against HIV infection while those with a copy (heterozygous) with the $CCR5 - \triangle 32$ allele shows delay in progression (Kaur and Mehra, 2009; Silva and Stumpf, 2004). Certain human leukocyte antigens (HLA) genotypes have also been associated with HIV progression i.e. $HLA - B35, HLA - A * 2301, HLA - A_2 \mid 6802$ are associated with fast disease progression while $HLA - B44$, $HLA - DRB * 01$ are associated with resistance to HIV-1 infection (Trachtenberg and Erlich, 2001).

Generally the human genome sequence in most cases is identical between any two individuals and variations in it contribute to the phenotypic differences including susceptibility to or protection against diseases (i.e. AIDS). The general mammalian mutation rate is low about (2×10^{-9}) on average per base pair per year and the majority of inter-individual genetic variability is inherited (Kumar and Subramanian, 2002; Wolfe et al., 1989). There are several types of variations in the human genome, ranging from a single base pair to thousands of base-pairs in size: single nucleotide polymorphisms (SNPs), repeat polymorphisms (minisatelites and microsatellites) and small insertions or deletions are some of the forms of variations (polymorphisms) in the human genome. Polymorphisms occurs when two or more clearly different phenotypes exists in the same population of species. Phenotype is the observable characteristics of an organism. Previous studies conducted by Weber and May (1989); Litt and Luty (1989) showed that these variations can be used as genetic markers in the search for genetic factors underlying human diseases. A genetic marker is a gene or deoxyribonucleic acid (DNA) sequence with a known location on a chromosome that can be used to identify individuals or species.

Based on the difference in genetic make-up among individuals, a proportion of individuals do develop resistance to infection hence referred to as long term progressors (elite controllers) while others progress to full blown AIDS very fast thus, referred to us fast progressors. The fast progressors are those individuals who rapidly progress to AIDS within four years after primary HIV-1 infection if they fail to take medication. Their progression not only depends on their genetic make-up but also is influenced by the HIV-subtype they are infected with. Those individuals exposed to HIV-1 subtypes C, D

and G are 8 times more likely to develop AIDS than those exposed to subtype A (Kanki et al., 1999).

Despite the numerous studies and technologies developed to study the disease, its pathogenesis is still poorly understood. More research to understand the factors that influence its progression is still required in order to better inform prevention, control and treatment strategies against the virus. Viral load and CD4 counts are still the commonly used HIV Bio-markers to monitor its progression and status in human beings. A better understanding of the genetic make-up of an individual is also important as it plays a crucial role in susceptibility to HIV|AIDS. The alarming increase in statistics for those who are living with HIV have accelerated much research into the pathology of HIV|AIDS so as to reduce its spread and eventually eradicate it.

1.2 Human Immuno-deficiency Virus (HIV)

Human immuno-deficiency virus (HIV) is the virus that causes this deadly disease (AIDS). HIV is one of the virus known as retroviruses. It attacks the most vulnerable part of the human body, the immune system (Al-Jabri, 2007). After getting into the body, the virus kills or damages cells of the body's immune system destroying the body's ability to fight infections thus making the body of an infected individual to be easily attacked by other opportunistic infections. The virus is passed on from one person to the next via bodily fluids i.e. semen, vaginal fluid and blood. This can occur during sexual contact (anal, vaginal or oral), exposure to infected body fluids or tissues, from mother to child during delivery or breast feeding; known as vertical transmission (Al-Jabri, 2007). An individual who contacts the virus (HIV) may live for many years before developing AIDS depending on her|his genetic make-up.

AIDS is not a specific illness, but a collection of different symptoms or conditions that manifest in the human body due to the weakened immune system after the virus has destroyed so much of the body's defences that immune cell counts fall to critical levels of an infected person. What makes this virus (HIV) difficult to understand is, its replication that is highly error prone resulting into a mutation rate of approximately (3×10^{-5}) per base per replication cycle. This results into a tremendous degree of HIV genetic variability within a single human host. Implying that, each HIV infected individual carries an entire population of viruses with each viral particle potentially comprised of different genetic material (Foulkes, 2009).

1.3 HIV Bio-markers

A Bio-marker is a substance, physiological characteristic or a gene that indicates, or may indicate the presence of a disease, a physiological abnormality or a psychological condition. The commonly used Bio-markers in predicting the progression of HIV to a state of the HIV disease (AIDS) are cluster of differentiation 4 (CD4) count and viral load.

1.3.1 Cluster of Differentiation 4 (CD4)

Discovered in 1970, originally known as leu 3 and T4 before referred to as CD4 in 1984. It's a glycoprotein found on the surface of immune cells such us T helper cells, monocytes, macrophages and dendrite cells. At times referred to us T helper cells since they send signals to other immune cells. They send the signal to CD8 cells which in turn destroy and kills the infection or virus. It is a co-receptor that assist T cells receptor (TCR) to activate its T cell following an interaction with an antigen presenting cell. Using its part that reside inside the T cell, CD4 amplifies the signal generated by the TCR by recruiting an enzyme known as the tyosine kinase (lck) which is essential for activating many molecules involved in the signalling cascade of an activated T cell. Its measured in cells per mm^3 and when depleted in the body, the body is left vulnerable to a wide range of infections. For a healthy HIV negative individual, the CD4 count should lie between $(600 - 1200)$ and if the CD4 count lies between $(350 - 600)$ for a healthy HIV positive individual then it's considered very good. The immune system is termed weakened whenever the CD4 count lies between $200 - 350$ and an individual is referred to as having AIDS if the CD4 count is below 200 (Walker, 2004; Reddy et al., 2011).

1.3.2 Viral Load

Viral load refers to the actual number of virus in the blood which is counted by carrying out a blood test i.e. polymerase chain reaction (PCR) expressed in copies per ml and in certain circumstances can be used to observe an individual's response to antiretroviral treatment. The viral load and the CD4 count of an individual are indirectly related. The higher the viral load the lower the CD4 count (this is because the virus destroys the CD4 counts) and the lower the viral load the higher the CD4 count (this is because when the number of virus particles in the body is low then, the body is given an opportunity to generate and maintain more CD4 counts). As indicators of the immune system strength for an HIV infected person, high CD4 count and low viral load count implies a strong immune system and vice-versa. However it should be noted that the correlation

between the two markers is mediated by other factors in body immune system and also depends at what stage of the disease we are referring to. These stages include the acute, asymptomatic, symptomatic and AIDS stages.

1.4 Immuno-Genetic Factors

Immuno-genetic factors are components of the genetic make-up of an individual that cause variability among individuals and can be used as disease markers in some instances. In this section, we describe some of the examples of the immuno-genetic factors: human leukocyte antigen (HLA), interleukin-10 (IL-10) and single nucleotide polymorphisms (SNPs).

1.4.1 Human Leukocyte Antigen (HLA)

Major histocompatibility complex (MHC) is a cell surface molecule encoded by a large gene family in all vertebrates. In humans it is also called the human leukocyte antigen (HLA). HLA's are just but a class of proteins on the surface membrane of cells which can be sub-divided into three categories namely: class I (A, B, C), class II (DP, DM, DOA, DOB, DQ and DR) and class III. The super locus contains a large number of genes found on chromosome 6. They encode cell surface antigen presenting proteins and has many other functions.

The roles played by different HLA's includes: (1) class I present peptides from inside the cell (including viral peptides if present). These peptides are produced from digested proteins that are broken down in the proteasomes, (2) class II present antigens from outside of the cell to T-lymphocytes and (3) class III HLA's encode components of the complement system. Other roles includes: disease defence, they are also the major cause of organ transplant rejections and also may protect against or fail to protect against cancer (Levsky and Singer, 2003).

Mutations in HLA's at times are associated with certain diseases such as Type I diabetes and coelic disease. The possibility of two unrelated individuals to poses identical HLA's molecules on all loci is very low and this is the main reason behind diversity of HLA's in the human population which also do promote the aspect of disease defence in HLA's (Brennan and Kendrick, 2006). Loci (singular locus) is the specific location of a gene or DNA sequence on a chromosome.

1.4.2 Interleukin-10 (IL-10)

Interleukin-10 were discovered in 1970, they are a subset of a larger group of cellular messenger cells called cytokines which modulate cellular behaviour. They are not stored within cells like other cytokines but secreted rapidly and briefly in response to a stimulus. Interleukin-10 (IL-10) is an important immuno-regulatory cytokine in humans (Asadullah et al., 2001). It does play a role in regulation of inflammatory responses and in the pathology of human auto immune disease. It suppress the T-cell immune responses which are the defence mechanisms that fight against foreign substances. It also inhibits major histocompatibility complex class I expression and plays an important role in the development of infectious diseases (Asadullah et al., 2001).

1.4.3 Single Nucleotide Polymorphisms (SNPs)

This is a DNA sequence variation occurring when a single nucleotide (A, T, C or G) in the genome differs between members of a biological species. In the human genome they occur in average once per 300 base pair, but their density do vary upto ten fold between different regions of the genome (Carlson et al., 2004). Though they are greater in number, different populations in the world have specific SNPs which clearly indicates the evolutionary history of human populations (Meier and Robinson, 2005). Due to their bi-allelic nature they are less informative compared to microsatellites in that their heterozygosity is limited to 50% (Ziegler et al., 2010). Therefore, to achieve high level of heterozygosity multiple SNPs are assembled together as haplotypes.

SNPs make-up about 90% of all human genetic variations in the human genome making them the most common markers for gene mapping because they are: (1) Abundant, (2) Have low mutation rate of $\approx 10^{-8}$ per generation, thus very stable. (3) Some are located in the genes, so they can be viewed as candidate variants for diseases. (4) Their automated detection. They arise as a result of misincorporation of nucleotides during replication or chemical and physical mutagenesis which can give rise to base substitutions in a DNA sequence. In principle, SNPs could be bi, tri- or tetra-allelic variations but tri- and tetra-allelic SNPs are very rare (Brookes, 1999). Most are di-allelic in nature and occur whenever one base is substituted for another. They are short in length and as a result of that they can be amplified using PCR (polymerise chain reaction) or ARMS (amplification refractory mutation system) and separated using gel electrophoresis or using time of flight (TOF) (Ziegler et al., 2010). In humans all combinations of substitution polymorphisms are observed with A|G substitution SNPs (including reverse complete T|C) being the most prevalent (Wang and Moult, 2001).

1.5 Literature Review on Immuno-Genetic Factors in Relation to HIV Bio-markers

As per the World Health Organisation progress report 2012, HIV|AIDS is still a major challenge facing the world. Previous studies have shown that there is an increasing evidence that the human leukocyte antigen (HLA) molecules do influence the rate of HIV-1 progression after infection (Martin et al., 2007). There are numerous polymorphisms of HLA molecules and as a result of this, the response to specific immuno-dominant HIV-1 epitodes among individuals differ depending on a person's HLA background (genetic make-up) (Borghans et al., 2007). The frequency of HLA molecules in the human population have also been associated with the rate of HIV-1 progression (Borghans et al., 2007). Individuals with rare HLA molecules have a high probability of being heterozygous at the HLA loci, thereby expected to induce an immune response against a larger diversity of peptides than those who are homozygous at the HLA loci. Heterozygous individuals are associated with slow progression of HIV-1 whereas homozygous individuals are associated with fast progression of HIV-1 (Gao et al., 2001; Carrington and O'Brien, 2003).

According to studies done by Gillespie et al. (2002) and Klein et al. (1998), HLA-B*57 molecules have been found to have immune control over HIV-1 and slow progression of HIV-1 to AIDS. HLA-B*58 and HLA-B*63 also posses binding motifs that are similar to HLA-B*57 therefore are also associated with good immune control of HIV-1 (Carrington and O'Brien, 2003; Frahm et al., 2005). Therefore, HLA-B*58, HLA-B*63, HLA-B*57 and HLA-B*27 are associated with slow progression of HIV-1 (Borghans et al., 2007).

According to Barker et al. (1998) they found out that some individuals ($< 5\%$) remain healthy after contracting HIV even without starting treatment as the anti-retroviral (ARVs) as a result of there genetic make-up. There findings showed that, those individuals who remain healthy their status are usually not influenced by either socio-economic or behavioural characteristics but is associated with the major histocompatibillity complex (MHC) alleles. They showed that CD4 counts among the slow progressors was low ($p < 0.0001$) compared to the fast progressors.

Studies from the Chinese population reported in Huang et al. (2009) also showed that, individuals who have HLA-B*5801, HLA-A*3303|haplotypes and are heterozygous for BW4-BW6 were more likely to be resistant to HIV progression to AIDS. HLA-B*14 (p=0.001) was associated with non-progression while HLA-B*27, HLA-B*57 and HLA-B*35 were associated with slow progression accordance to studies carried out by Roger (1998) which reaffirmed that HLA-A*23, HLA-A*24, HLA-A*26, HLA-B*21 and HLA-B*38 were associated with fast disease progression while HLA-B*17, HLA-B*27, HLA-B*51 and HLA-B*57 were associated with slow disease progression.

Looking at the previous studies on Interleukin-10, IL-10(-592) and IL-10(-1082) have shown some relationship with the susceptibility and pathogenesis of HIV-1 which varies depending on the infection phase (Naicker et al., 2009). The findings from Naicker et al. (2009) were reaffirmed by studies carried out using the European Americans genetic data which showed that, IL-10(-592) is associated with increased susceptibility to HIV-1 infection and progression to AIDS at a very high rate in the late stages of the disease. Studies done by Naicker et al. (2009) showed that individuals who were homozygous IL-10(-592) for the AA genotype have a higher chance of being infected with HIV-1. According to studies done by Shin et al. (2000), they showed that those individuals with IL-10(-592) A genotype were at a higher risk of contracting HIV compared to those having the wild type allele (most common allele in the population of study). This was reaffirmed by studies carried out by Naicker et al. (2012) which also showed that IL-10(-592) AA genotype was associated with reduction of CD4 T cells (p=0.0496) while those individuals with IL-10(-1082) GG genotype have higher levels of IL-10 compared to those with -1082 AA|AG (p=0.0006). They concluded their study by stating that, IL-10 promoter genetic variants might influence the rate of HIV-1 disease progression by regulating IL-10 levels and the breath of CD8 T cell immune response. According to Naicker (2011) they showed that individuals who were homozygous for the mutation at the IL-10(-592) AA genotype were 2.78 times more likely to contract HIV compared to those individuals who were homozygous wild type (CC) genotype at the same position (p=0.023). Polymorphisms associated with decreased IL-10 production have been associated with increased likelihood of human immuno-deficiency virus type one (HIV-1) acquisition mostly in the late stages of the disease suggesting that, high IL-10 production may reduce susceptibility to HIV-1 infection and protect against disease progression (Naicker et al., 2009).

Previous studies on single nucleotide polymorphisms showed that APOBEC3G transcription rapidly down regulated upon HIV-1 infection thereby reducing the level of CD4 counts and increasing the level of viral particles in the blood. In that, the level of APOBEC3G among the HIV negative individuals was high compared to the HIV positive individuals even though there is no any relationship between APOBEC3G and the HIV Bio-markers (Reddy et al., 2010). This was reaffirmed by studies done by Jin et al. (2007), they showed that there is no any association between APOBEC3G mRNA levels with the HIV Bio-markers. Studies carried on Peptidyl Prolyl Isomerase A (PPIA) among African American examined the effect of single nucleotide polymorphisms (SNPs) and the haplotypes within the PPIA gene on HIV-1 longitudinal history cohort and they found that among the eight SNPs they tested, two promoter SNPs (SNP3 and SNP4) were not only in perfect linkage disequilibrium with one another but were also related with rapid CD4 T cell decrease (p=0.003) (An et al., 2007). This was also observed among the European Americans where the same alleles were found to be

associated with rapid progression of HIV-1 to AIDS. Studies on Transportin 3 Protein (TNPO3 or TRN-SR2) with HIV-1 showed that, TNPO3 do assist in HIV-1 replication in the nucleus and that it may assist preintergration complex (PIC) maturation in the nucleus (Diaz-Griffero, 2012) .

In conclusion, the pathogenesis and progression of HIV|AIDS do not entirely depend on the geographical location of an individual and his|her socio-economic factors but also influenced by the genetic make-up of an individual in accordance with the major histocompatibility complex. Previous studies have shown that certain individuals are slow progressors while others are fast progressors and those with two copies of chemokin receptor ($CCR5\Delta32$) remained healthy for a period of time without starting on ARV's (An et al., 2011). Therefore, inter-individual variability does influence susceptibility of HIV|AIDS.

1.6 The Mixed Effects Model

These are models that contain both the fixed effects and the random effects part. The general form of the model can be decomposed into two additive components namely; the fixed effects part and the random effects. Fixed effects are the effects attributable to a finite set of levels of a factor that occur in the data while random effects are attributable to an infinite level of a factor, of which only a random sample are allowed to occur in the data (McCulloch, 2006; Verbeke and Molenberghs, 2005). They are relevant to our study because: (1) It is useful in settings where measurements are made on clusters of related statistical units i.e. students within classrooms, or to repeated measurements on each subject over time. (2) The model allows a wide variety of correlation patterns to be explicitly modelled. (3) They take into consideration variation that is not generalised to the independent variables (Verbeke and Molenberghs, 2005; Diggle et al., 2002).

1.7 Missing Data

Data is considered missing whenever there is no information about a variable of interest may be as a result of the design used, chance or unforeseen circumstances. It is a very common problem with any real study and if not dealt with properly can lead to biased inference or at times inefficient analysis. The missing data mechanism can either be missing completely at random (MCAR) which implies that, the missing process does not depend on Y_i at all, missing at random (MAR), if the missingness depends on the unobserved response or missing not at random (MNAR) which allows the missing process to depend on the unobserved responses (Little and Rubin, 1987).

1.8 Problem Statement

In this study, our aim is to use data from the Sinikithemba cohort study collected by the HIV Pathogenesis Programme (HPP) at Nelson Mandela Medical School, University of KwaZulu-Natal in Durban. The data consists of 451 HIV positive and treatment naive individuals. We use the data to model the bivariate response variable (CD4 counts and viral load) with immuno-genetic factors to reaffirm the results of previous studies and to also deal with missing data which is a common challenge to longitudinal studies.

1.9 Objectives of the Study

(i) To investigate statistical methods to show that HLA class I alleles do influence HIV-1 susceptibility and progression to AIDS causing an individual to be a long term progressor or a slow progressor depending on his|her genetic make-up.

(ii) To investigate the importance of IL-10 promoter polymorphisms ($IL{-}10(-1082), IL{-}10(-592), IL-10(-3575)$) using the Sinikithemba data consisting of 451 HIV-1 C naive individuals to see whether there is any association between IL-10(-3575), IL-10(-1082) and IL-10(-592) genotypes with the HIV-1 Bio-markers (CD4 counts and viral load).

(iii) To see the association between the SNPs (TNPO3, APOBEC3G, $PSIP_{rs12339417}$ and $PPIA_{1650}$) with the HIV Bio-markers.

(iv) To understand the use of multiple imputation method to account for missing data in both the outcome and covariates for longitudinal data subject to drop out.

The next chapter illustrates exploratory data analysis using the Sinikithemba data consisting of 451 HIV-1 C naive individuals.

Chapter 2

Exploratory Data Analysis

2.1 Introduction

In 2005 HIV Pathogenesis Program (HPP) initiated a Sinikithemba study (SK) at Nelson Mandela Medical School, University of Kwazulu-Natal in Durban. The cohort enrolled 451 HIV naive individuals of which 359 (79.06 %) were females and 92 (20.399 %) were males including some incorporated from a previous study. Among the 451 individuals, 184 remained active during the whole study period, 126 defaulted, 38 died and 103 dis-enrolled from the study at some point. Baseline characteristics such as: (1) High resolution HLA typing was done with genomic (DNA) with single stranded conformation polymorphism (PCR), (2) Cellular immunology, (3) CD4 counts, (4) Viral load and (5) Demographic characteristics such as gender were collected.

This was a follow up study where for CD4 count measurements visits were scheduled after every three months and for viral load measurements visits were scheduled after every six months. This was to help in accruing data to help in understanding the progression and the pathogenesis of the disease (AIDS). The viral loads were measured with the Roche Amplicor assay version 1.5 (Roche NJ) while the CD4 count was measured by flow cytometry (Becton Dickinson, San Jose, CA). Other important information regarding entry into anti-retroviral treatment (ART) was updated throughout the study and recorded at the time of administration. As time elapsed the participants (individuals) were monitored carefully, given clinical care throughout the study and provided with feedback regarding CD4 counts and viral loads levels. Counselling and guidance from expertise such as experienced doctors and nurses, among others were also provided to the participants. They were referred to the government sector medical care centre for ARV initiation as per the national guidance given by South African government whenever their CD4 counts fell below 350mm|μL cells for more conservative visits. They were

advised to start highly active antiretroviral therapy (HAART) once their CD4 count fell below 200mm|μL cells (South Africa National Department of Health, 2004).

2.2 Study Design

This study (the Sinikithemba study) can be categorised into two ways: (1) Observational cohort study; because a group of individuals are followed prospectively over time or (2) Longitudinal study; because Bio-marker measurements of the same individual (participant) are taken repeatedly at an intervals of 3 and 6 months for CD4 counts and viral load respectively, over time at regular but not necessarily equal intervals as time elapses because of mistimed visits.

Longitudinal studies posses some advantages over others. The advantages include:

(i) Allows gathering of information on intra-individual changes within individuals.

(ii) Helps in observing change or evolution of a process overtime i.e. increase of viral load or decrease of CD4 counts with time which are in this case related to the disease process.

(iii) It does account for individual to individual heterogeneity which is not possible in other studies.

The main challenge with longitudinal studies is that; it may lead to incomplete data due to drop-out or intermittent missingness at times which needs to be accounted for during the modelling process.

2.3 HIV Bio-markers

2.3.1 CD4 counts

Figure 2.1 below shows CD4 counts over time in months (after every three months) for all the 451 individuals in the Sinikithemba study. From Figure 2.1, we can clearly see that CD4 counts decreases with time except for certain individuals (outliers) whose CD4 counts increases with time. Figure 2.1 clearly shows existence of variability between and within the individuals which will be accounted for using the mixed model approach at the modelling stage.

Figure 2.2 and Figure 2.3 shows CD4 counts for females and males respectively. From the two figures we are able to see that, the behaviour pattern of CD4 counts with time in

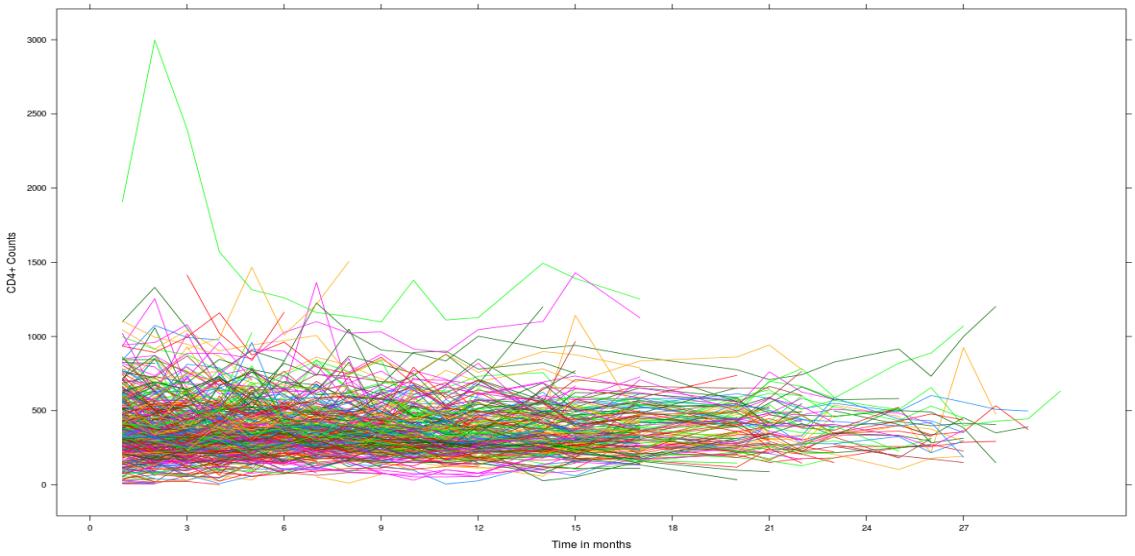

FIGURE 2.1: CD4 counts at different time points

both cases is similar with time except for some few outliers whose CD4 counts behaviour is unique. Even though the number of males are few (92) compared to the females (359) in the study.

Figure 2.4 shows the mean CD4 counts for females and males on the same graph to show if there is any association between the two categories of individuals (males and females). From Figure 2.4 we are able to deduce that there is a slight difference between the two categories of individuals. The plot in Figure 2.4 indicates that, the mean CD4 counts in females is slightly high compared to that in males but the trend in both cases remains the same. As time elapses we are able to observe the difference in CD4 counts between the two groups. Increase in time causes no change to our initial observations for the mean CD4 counts between the individuals. The gap between the two groups starts widening as from time-point 15. This could be attributed to low turnout for males due to death or drop out by some individuals from the study.

As shown in Table 2.1 we are able to see that, the number of individuals in the study decreased with time. It was highest at the initial stages of the study but as time elapses the individuals kept on decreasing in number maybe as a result of death or withdrawal by some individuals from the study.

We performed a normality test on CD4 count data to see whether the data satisfies the normal distribution assumptions. Our results showed that the CD4 count data violates the normality rules as depicted by Figure 2.5. Thus, the need to carry out remedial measures such as data transformation.

When we performed a log transform on the CD4 count data and carried out the normality tests on the transformed data, there were some slight improvements on adherence to the

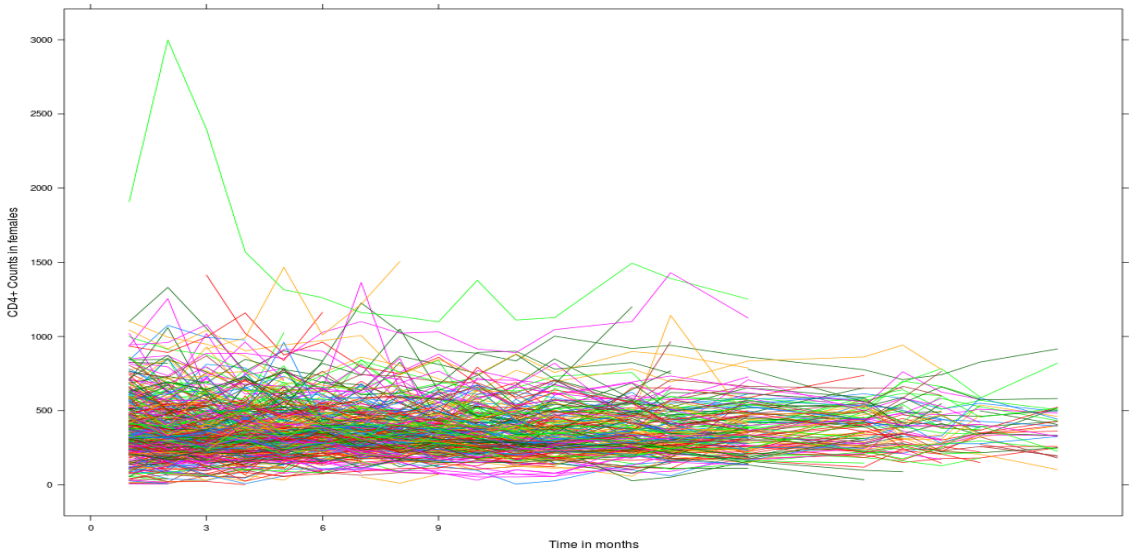

FIGURE 2.2: CD4 counts over time after every three months in females

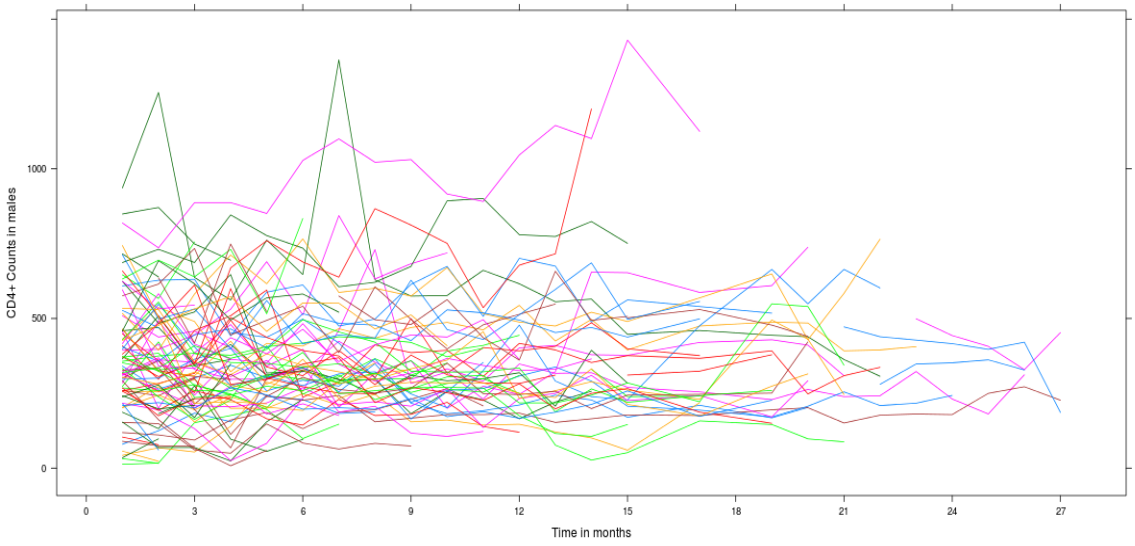

FIGURE 2.3: CD4 counts over time after every three months in males

normality assumptions compared to when the normality tests were carried out on the original CD4 count data. Even though there were some slight improvements, the log transform data still violated the normality rules. Therefore, the need to perform another transformation on the CD4 count data. Figures 2.5 and 2.6 show the histogram plots for the original and log-transformed CD4 count data.

On the other hand the square root transformed CD4 count data showed good observable improvements on normality adherence. However both the Kolmogorov-Smirnov and the Cramer von-mises tests showed some slight significance probably due to the test detecting the longer than normal tails when normality test were performed on the square root CD4 count data. The kurtosis and the skewness of the square root CD4 count data illustrated

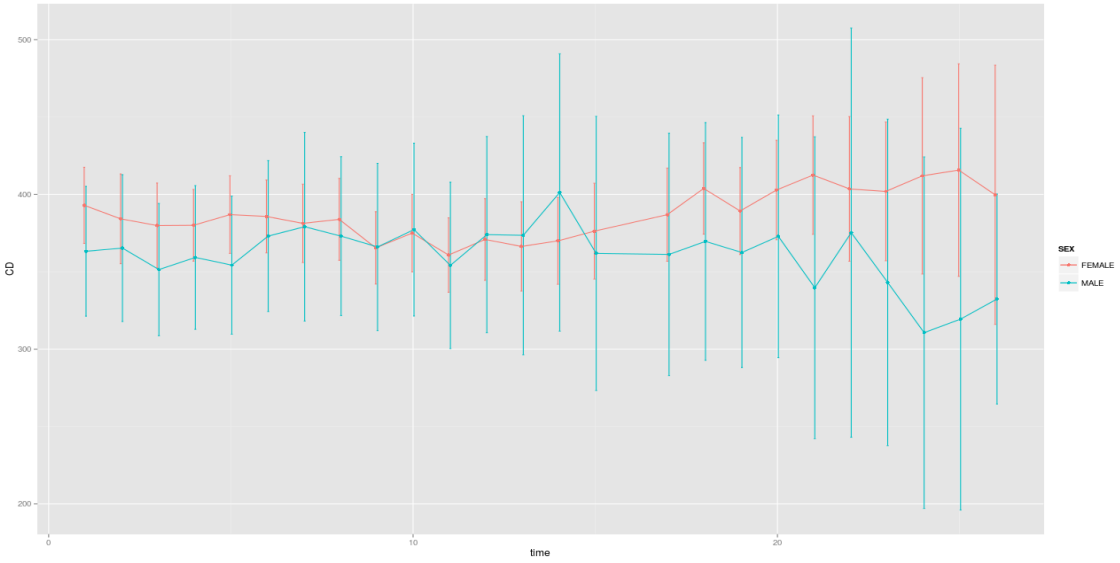

FIGURE 2.4: Mean CD4 counts for all the individuals

TABLE 2.1: The number and percentage of participants present at a given follow up time for CD4 count

No of participants observed per time point	Time points	Percentage of participants at a given time point
1	451	100
2	382	84.70
3	371	82.26
4	353	78.27
5	321	71.18
6	296	64.75
7	274	60.75
8	262	58.09
9	254	56.32
10	241	53.44
11	232	51.44
12	219	48.56
13	207	45.9
14	195	43.24
15	185	41.02
16	178	39.47
17	164	36.36
18	156	34.57
19	132	29.27
20	109	24.17
21	80	17.74
22	62	13.75
23	48	10.64
24	42	9.31
25	35	7.76
26	31	6.87
27	23	5.1
28	11	2.44

on Table 2.2 showed a slight difference from that of a normal distribution. Due to its better performance compared to the log transformation, the square root CD4 count data will be used for our further analysis instead of the log transformed CD4 count data. The histogram and the QQ plots on Figure 2.7 support the choice of square root CD4 count as the most preferred form of transformation on the original CD4 counts data.

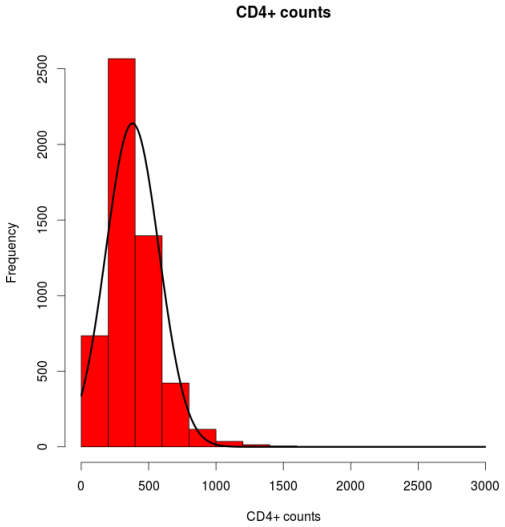

FIGURE 2.5: Histogram for CD4 counts

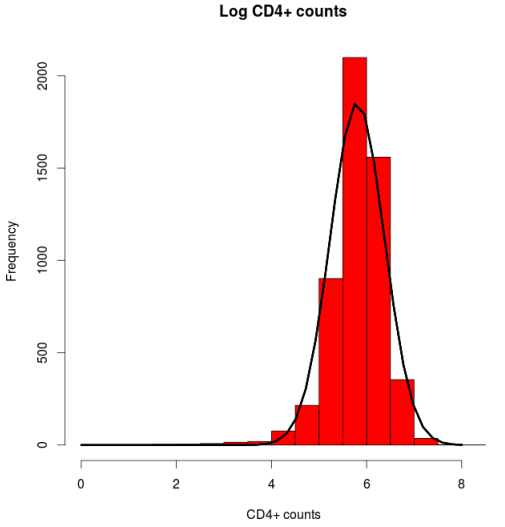

FIGURE 2.6: Histogram for log CD4 counts

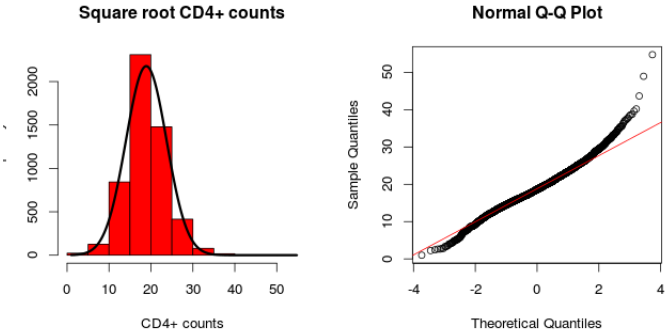

FIGURE 2.7: Histogram for square root CD4 counts and its QQ plots

TABLE 2.2: Descriptive statistics for square root CD4 counts

Mean	18.88
Median	18.63
Mode	16.37
Standard deviation	4.85
variance	23.52
Range	53.75
Interquartile range	5.98
Skewness	0.3946
Kurtosis	1.7454

TABLE 2.3: Test statistic

Test	Statistic	P -Value
Lilliefors (Kolmogorov Smirnov)	0.0367	0.0029
Cramer von-mises	2.21	< 0.0001

2.3.2 Viral Load

The viral load for the individuals was collected from the 12^{th} of August 2003 to 15^{th} of April 2011. Figure 2.5 shows a graph of all the individuals at different time points.

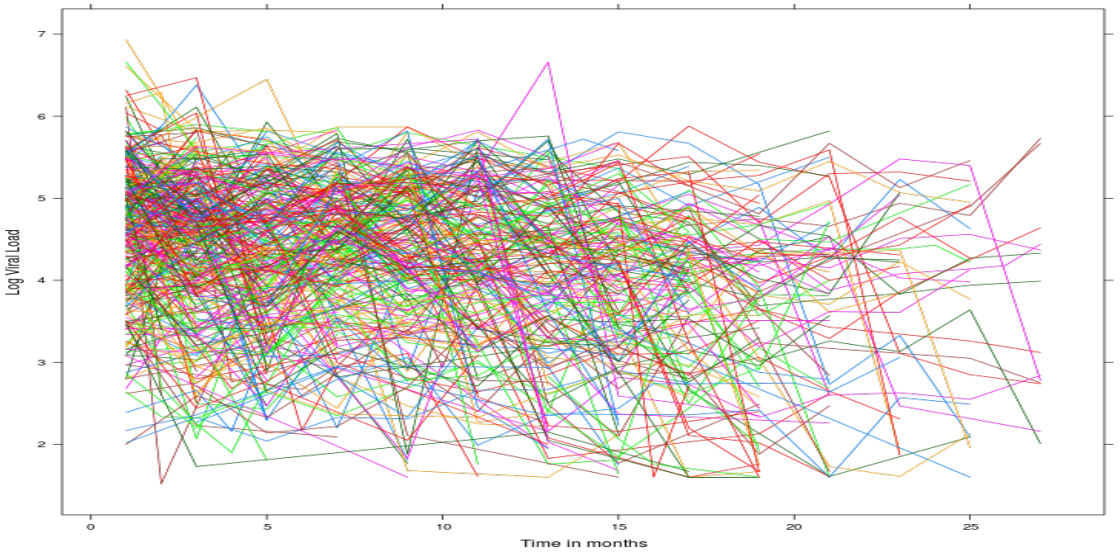

FIGURE 2.8: Data on log viral load for all the individuals

We also plotted the log viral loads on different graphs to observe the trend of viral load in the two groups of individuals (males and females) to see whether there were any observable difference basing on gender. See Figures 2.9 and 2.10.

We then present the number and percentage of participants present at a given follow up time; see Table 2.4. From Table 2.4 we notice that, the log viral load data also follows the same trend as for the CD4 count data. At the initial stages of the study the number

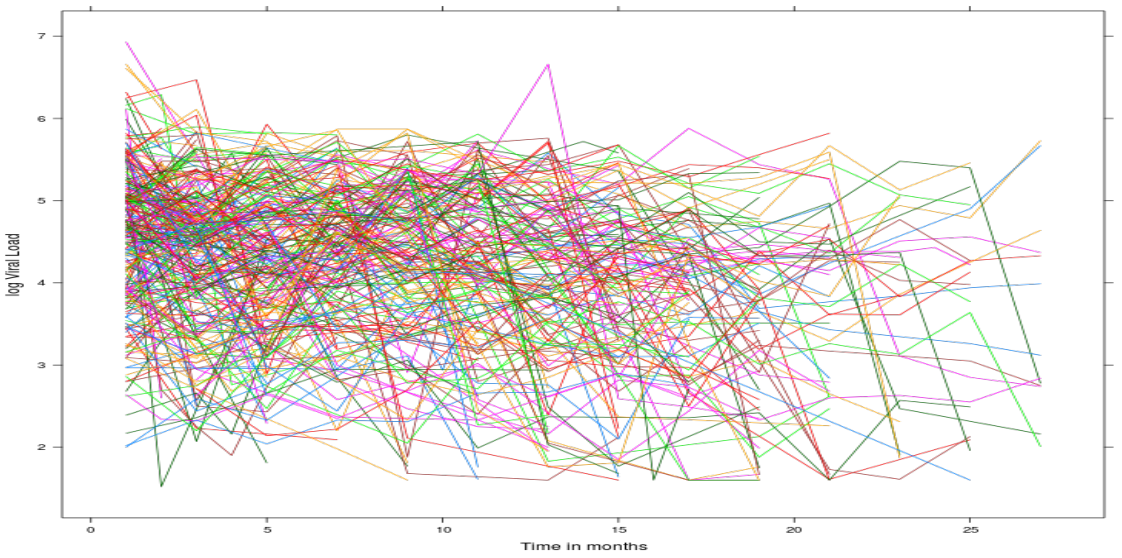

FIGURE 2.9: Data on viral load for females at different time points

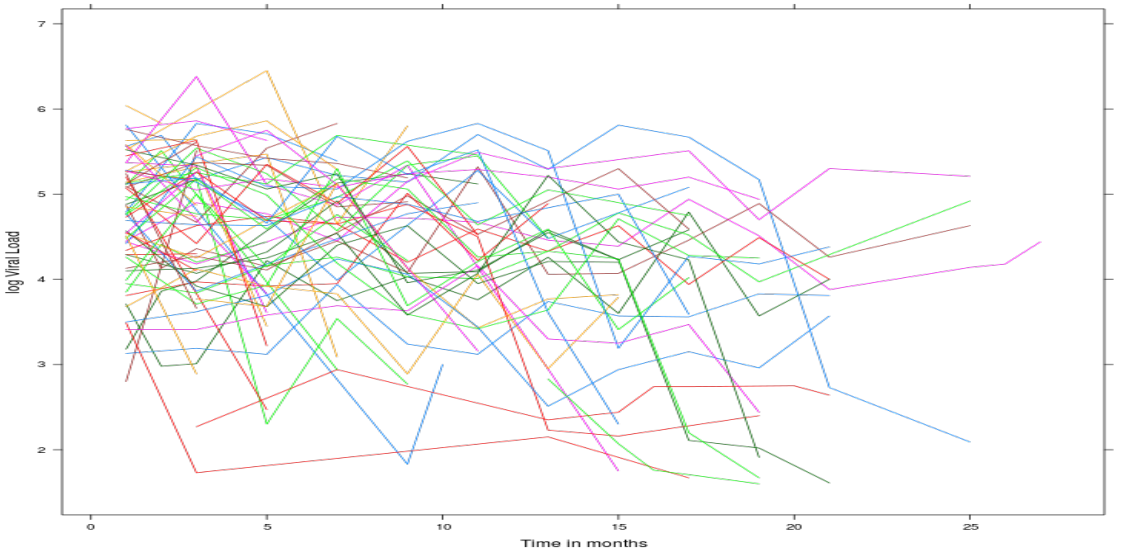

FIGURE 2.10: Data on log viral load in males at different time points

of individuals who turned up for the study were high and as time went by the number of individuals kept on decreasing.

Just as observed with the CD4 count data, the viral load data also violates the normality assumptions as depicted on Figure 2.11. Therefore, some remedial measures such as variable transformation may be necessary. A log transform was performed on the viral load data and when normality tests were performed, normality conditions were still violated but not as severe as noted with the original viral load data. This was supported by the Kolmogorov-Smirnov ($p < 0.0001$) and Crammer von-misses ($p < 0.0001$) tests which showed lack of normality adherence when tested on the log viral load data. For

TABLE 2.4: The number and percentage of participants present at a given follow up time for viral load

No of participants observed per time point	Time points	Percentage of participants at a given time point
1	427	94.68
2	361	80.04
3	319	70.73
4	276	61.2
5	252	55.88
6	227	50.33
7	207	45.9
8	177	39.25
9	152	33.70
10	112	24.83
11	70	15.52
12	40	8.87
13	29	6.43
14	14	3.10

further analysis a log transform will be more preferable than the original viral load data. See Figure 2.11.

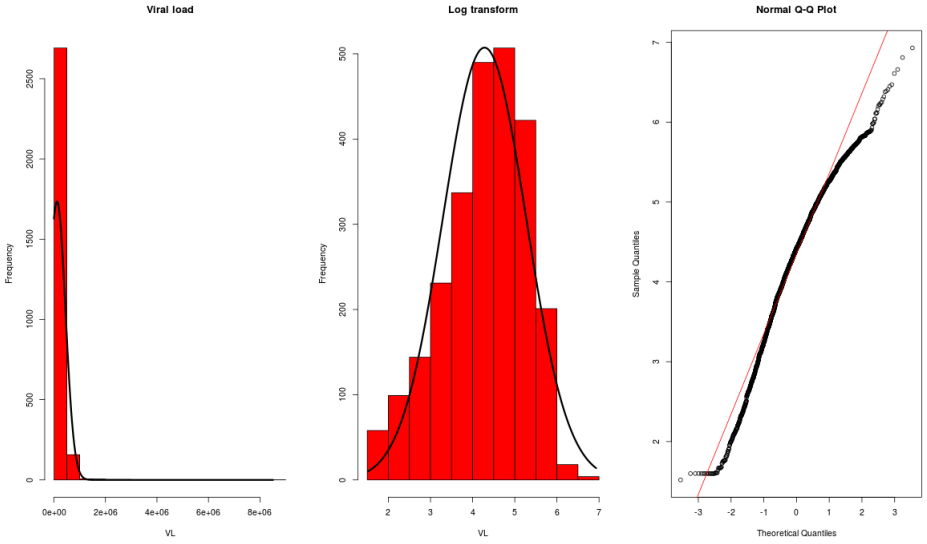

FIGURE 2.11: Histogram for viral load and the QQ plots.

The descriptive statistic illustrated in Table 2.5 for the log viral load transform shows that, the kurtosis and the skewness for the log viral load are not any different from that of a normal distribution. Therefore, the log transform remains the preferred choice for further analysis.

We also carried out the Spearman correlation test to see whether there were any relationship between CD4 count and viral load at different time points for all the individuals in the study. Looking at Figure 2.12 and Figure 2.13 we are able to deduce that, there is a positive correlation between the different time points for CD4 counts and viral load ($p < 0.0001$). Therefore, there exits an association between the different time points for the HIV Bio-markers.

TABLE 2.5: Descriptive statistics for log transform on viral load

Mean	4.282
Median	4.420
Mode	4.33
Standard deviation	0.9872
variance	0.97463
Range	5.410
Interquartile range	1.3550
Skewness	-0.56331
Kurtosis	-0.1395

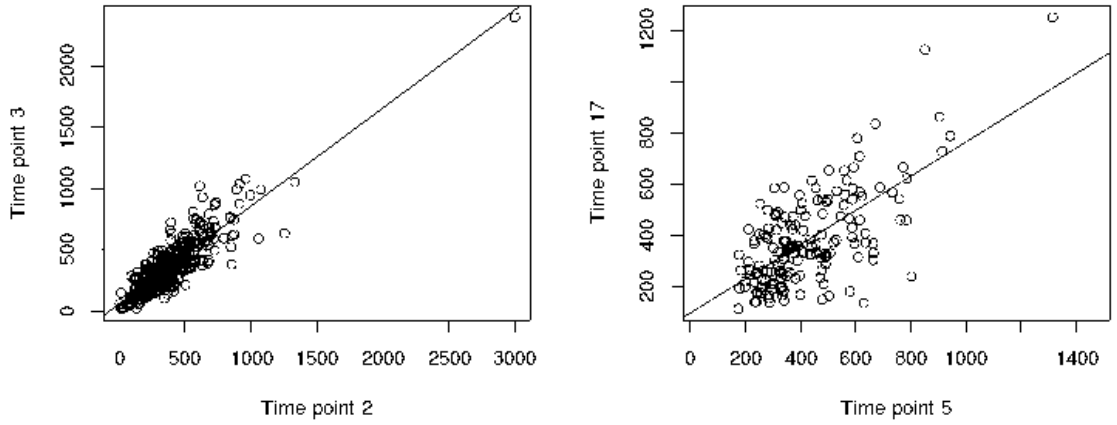

FIGURE 2.12: Diagrams illustrating the correlation of CD4 counts at different time points

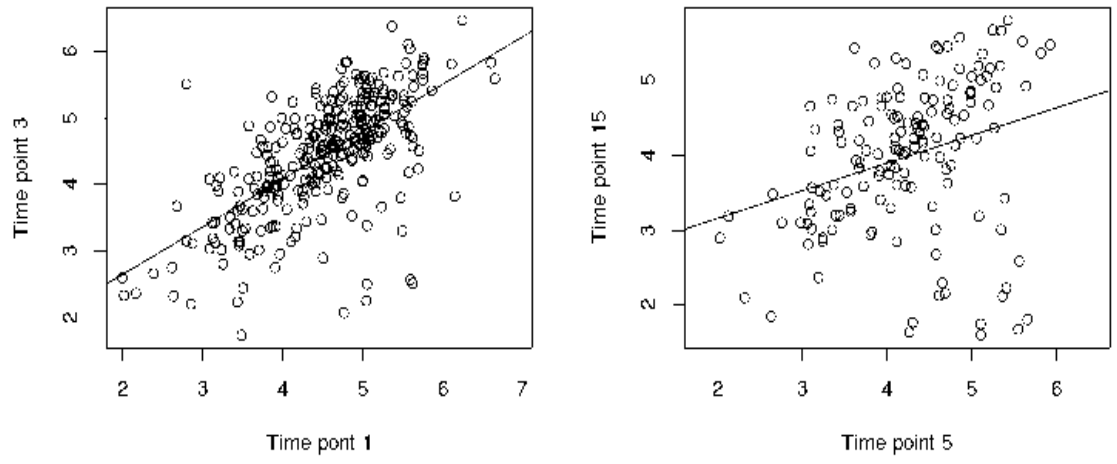

FIGURE 2.13: Diagrams illustrating the correlation of viral load at different time points

2.4 Human Leukocyte Antigen (HLA) Subtypes

HLA class I alleles have been associated with relatively successful control of viral replication and slow disease progression. Its locus (HLA) which codes for HLA class I alleles (which present pathogenesis derived peptides on the cell surface of infected cells for recognition by CD8 T lymphocytes) shows a significant variation among individuals from different geographical locations which has been attached to differences in HIV|AIDS disease progression and natural viral control (Rajapaksa et al., 2012).

Table 2.6 gives some of the most frequent HLA subtypes and their frequencies in the Sinikithemba study.

TABLE 2.6: Summary for the most frequent HLA subtypes in the SK study

HLA subtype	Frequency	HLA subtype	Frequency
HLA-A*02	55	HLA-B*0801	45
HLA-A*3001	81	HLA-B*4201	92
HLA-A*29	71	HLA-C*03	55
HLA-A*2301	81	HLA-C*04	78
HLA-A*6802	71	HLA-C*10701	96
HLA-B*5802	79	HLA-C*07	158
HLA-B*1503	91	HLA-C*0202	96
HLA-B*4201	97	HLA-C*0602	101

We performed Kruskal wallis rank sum test to see whether there were any association or variability between the different HLA class I subtype alleles. We compared the various categories of HLA subtypes: (HLA-A and HLA-B), (HLA-B and HLA-C) and lastly between (HLA-A and HLA-C); (p=0.02427), (p <0.0001) and (p=0.003028) respectively. The $p-$values showed some significance therefore there exists some association and variation between the different HLA's class I subtypes. We then plotted box and whisker plots for CD4 count and log viral load against the HLA class I alleles. Figure 2.14 and Figure 2.15 suggests that HLA-B alleles are highly associated with high and low viral loads and CD4 counts compared to HLA-A and HLA-C alleles.

2.5 Interleukin-10 Promoter Polymorphisms (IL-10)

Out of the 451 individuals that were present in the study, 50 had AA, 210 had CA and 191 had the CC IL-10(-592) genotype, 221 had AA, 181 had AG and 48 had GG IL-10(-1082) genotype and lastly 19 had AA, 168 had TA and 264 had TT IL-10(-3575) genotype. Genotype is the genetic make-up of an individual. Table 2.7 gives a summary of the sample distribution of IL-10 polymorphisms. IL-10 is bi-allelic implying that it has two alleles at a specific locus. An allele is generally a group of genes, while a gene is the molecular unit of heredity of a living organism (Feero et al., 2010).

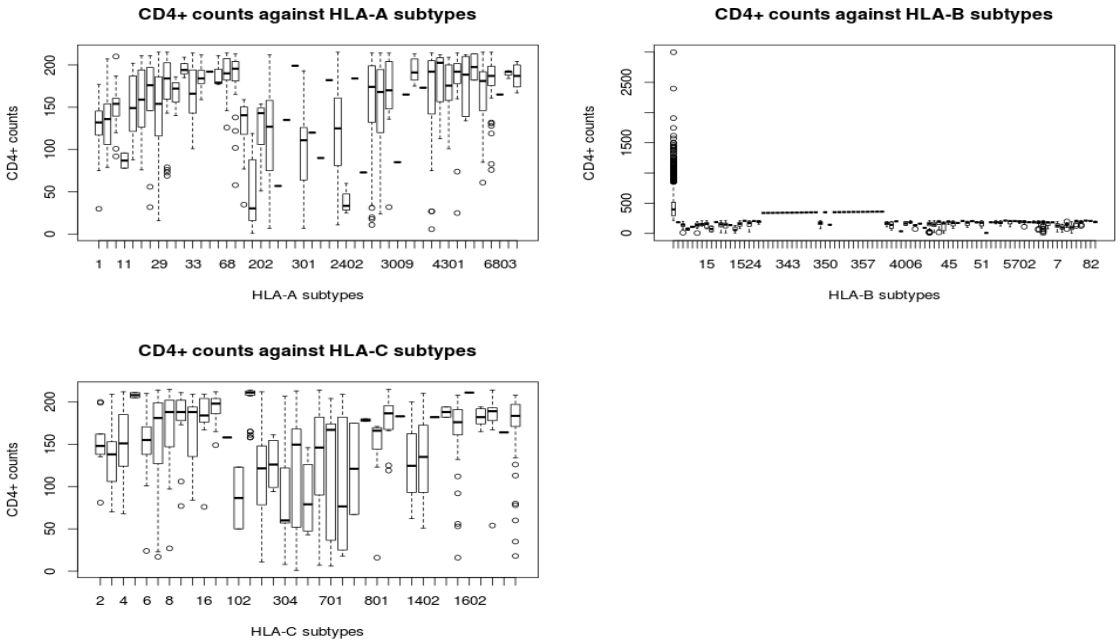

FIGURE 2.14: Trends of CD4 counts in different HLA subtypes

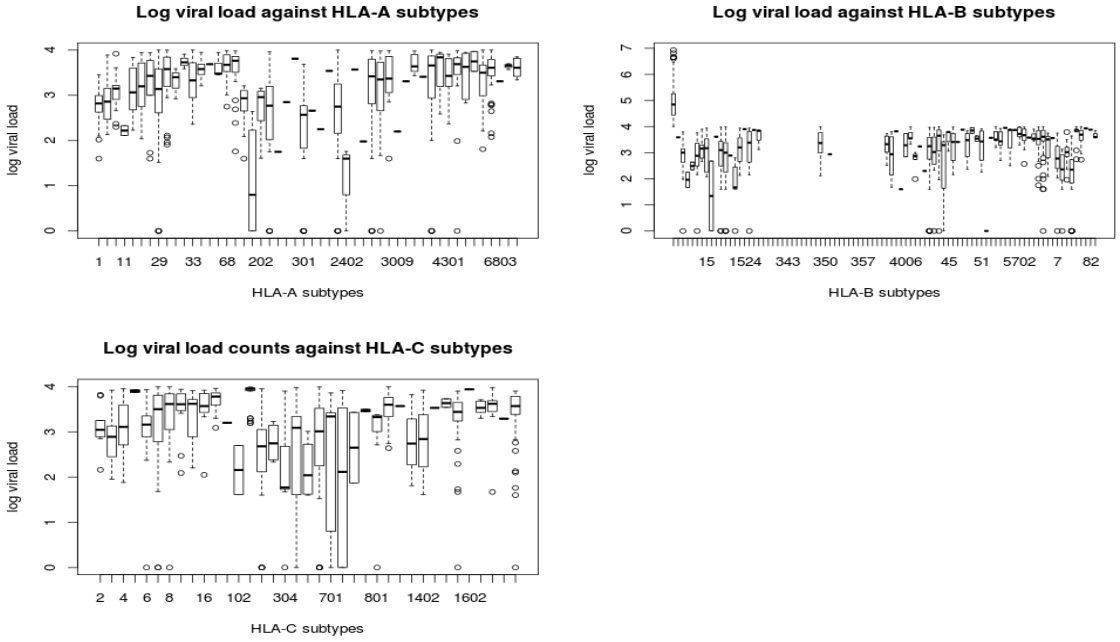

FIGURE 2.15: Trends of VL in different HLA subtypes

TABLE 2.7: Summary of Interleukin-10 promoter polymorphisms

IL-10(-592)	IL-10(-1082)	IL-10(-3575)
AA 50	AA 221	AA 19
CA 210	AG 181	TA 168
CC 191	GG 48	TT 264
Total 451	451	451

We calculated the frequencies of all the available alleles in the study and found out that, for IL-10(-3575) its minor allele frequency was 22.84% (A) while the major allele frequency was 33.48 (T), for IL-10(-592) the minor allele frequency was 32.26% (C) while the major allele frequency was 34.37% (T). Finally, for IL-10(-1082), the allele frequencies were 54.32% (G) and 69.18% (A) for the minor and major allele frequencies respectively. We then performed the Hardy and Weinberg test. Hardy and Weinberg (1908) introduced an important principle based on the relationship between genotype and allele frequencies in a population (formally known as the Hardy-Weinberg equilibrium (HWE) test. They based their principle under the following assumptions: (1) Random mating implying that mating does not depend on the genotype, every individual has the same opportunity of mating with any individual of the opposite sex, a state known as panmixia. (2) No selection or migration which implies that, if the probability to reproduce depends on the genotype at the relevant locus then a selection has taken place which should not be the case. (3) No mutation, since mutation results in the shift of the equilibrium, as a result of this, the allele and genotype frequency do not remain fixed and stable from generation to generation. (4) No population stratification meaning the population of interest should be homogeneous. (5) Infinite population size so as to avoid the sampling errors, the population has to be large. Based on these assumptions the equilibrium is not distorted (Ziegler et al., 2010). To see whether these alleles obeyed Hardy and Weinberg principle, we performed HWE test and found the associated $p-$values for the test as; IL-10(-3575) (p=0.2845), IL-10(-1082) (p=0.2694) and IL-10(-592) (p=0.5322) showing that they were all in HWE.

The Kruskal-Wallis test between the median CD4 counts and IL-10(-592) (p=0.3641), IL-10(-1082) (p=0.1453) and IL-10 (-3575) (p=0.0979) genotypes showed no significance and when a pairwise comparison test using wilcoxon test was performed between the various genotypes still there was no any observable significance. We also performed Kruskal-Wallis test for median log viral load and the various Interleukin-10 promoter polymorphisms and noted no observable significance. The $p-$values were 0.7204, 0.2531 and 09795 for IL-10(-592), IL-10(-1082) and IL-10(-3575) respectively. The pairwise comparison using wilcoxon test showed no significance for IL-10(-592) and IL-10(-1082) except for IL-10(-3575) ("AA", "TA") genotypes which showed some significant difference between median log viral load and IL-10(-3575) ("AA", "TA") genotypes.

2.6 Single Nucleotide Polymorphisms (SNPs)

To measure the expression of APOBEC3G, TNPO3, $PPIA_{1650}$ and $PSIP_{rs12339417}$ real time PCR was used and in identifying the variants in the SNPs, DNA sequencing and Tag-Man genotyping were performed. From the individuals in the study, for $PPIA_{1650}$

we found out that, 144 individuals had the AA genotype, 120 individuals had the AG genotype and 54 individuals had the GG genotype while 154 had missing information. The corresponding allele frequencies were 22.89% A while G had a frequency of 22.82%. Thus the minor allele was G with a frequency of 22.82%. The minor allele frequency refers to the frequency at which the least common allele occurs in a given population. For APOBEC3G we found out that, 230 individuals had the CC genotype, 137 individuals had the CG genotype and 30 individuals had the GG genotype while 54 had missing information. The corresponding allele frequencies were 1.82% C while G had a frequency of 1.48%. Thus the minor allele is G with a frequency of 1.48%. For TNPO3 we found out that, 298 individuals had the AA genotype, 82 individuals had the AG genotype and 9 individuals had the GG genotype while 62 had missing information. The corresponding allele frequencies were 87.15% A while G had a frequency of 77.76%. Thus the minor allele is G with a frequency of 77.76%. Lastly for $PSIP_{rs12339417}$ we found out that, 43 individuals had the CC genotype, 175 individuals had the CT genotype and 172 individuals had the TT genotype while 61 had missing information. The allele frequencies were 33.46% C while T had a frequency of 33.07%. Thus the minor allele is T with a frequency of 33.07%. The highest number of genotypes in the population are termed homozygous wild-type and the rare ones are termed homozygous recessive. We also performed HWE test to see whether the alleles obeyed HWE principle. The $p-$values were not significant therefore showing that the alleles were not in Hardy Weinberg equilibrium.

The next chapter illustrates the theory behind the linear mixed models.

Chapter 3

Linear Mixed Models

3.1 Introduction

In this chapter we present the theory behind linear mixed models. It is our model of choice since the Sinikithemba study (our study) involves repeated measures over time. Our main objective for this study is to find out the relationship between the HIV Biomarkers (CD4 count and viral load) and selected covariates. This chapter describes the statistical models used in the subsequent chapters. It also describes how the parameters were estimated.

3.2 Model Description

3.2.1 Linear Mixed Model

Linear mixed model procedures extends the general linear model so that the data is allowed to display correlation and non constant variability. It deals with continuous longitudinal data assumed to be normally distributed and in its simplest form the model can be written as:

$$Y_i = X_i\beta + Z_iU_i + \epsilon_i \tag{3.1}$$

for

$$i = 1, 2, \ldots, N,$$

where Y_i is an $n \times 1$ response vector for the i^{th} individual, such that Y_{ij} denotes the $j^{th}(j = 1, 2, 3, \ldots, n)$ observation made at time t_{ij} for the individual, X_i is the model matrix for the fixed effects for the observations in individual i, Z_i is the model matrix for the random effects for observations in individual i, U_i is the random effect coefficient vector for individual i, β is the vector of fixed effect coefficient, N is the number of subjects and ϵ_i is the errors for observations in individual i (Laird and Ware, 1982). The assumptions are; $U_i \sim N(0, G)$, where G denotes the covariance matrix of random effects U_i; $\epsilon_i = (\epsilon_{i1}, \cdots, \epsilon_{in_i})' \sim N(0, R_i)$; $U_1, \ldots, U_n, \epsilon_1, \ldots, \epsilon_n$ are independent; and R_i is the covariance matrix of error vector ϵ_i for observations in individual i (Verbeke and Molenberghs, 2009). The elements in G and R_i are known as variance components.

We note that,

$$\begin{bmatrix} U_i \\ \epsilon_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} G & 0 \\ 0 & R_i \end{bmatrix} \right).$$

If we assume that $\epsilon_1, \ldots, \epsilon_n$ are independent then, it follows that $R_i = \sigma^2 I$ where I is the identity matrix .

We know that missing data is one of the challenges encountered with longitudinal studies. Previous studies have shown that Gaussian theory estimation procedures for mixed models which consists of the maximum likelihood (ML) and the restricted maximum likelihood (REML) are some of the methods that can be used to deal with such challenges.

3.3 Estimation Methods

3.3.1 Maximum Likelihood Estimation (MLE)

Maximum likelihood estimation was introduced by R.A Fisher in 1912. The aim of this method is to construct an estimator for an unknown parameter θ.

The mixed effects model shown in Equation 3.1 can also be expressed as a marginal model of the form

$$Y_i \sim N(X_i\beta, Z_i G Z_i' + R_i).$$

We should note that inferences based on the marginal model do not explicitly assume the presence of random effects representing the natural heterogeneity between subjects

Ramroop (2008). Let β denote the vector of fixed effects coefficients and α denote the vector of all variance components in G and R_i. It then follows that the variance covariance matrix V_i of Y_i is α dependent. Thus we can let $\theta = (\beta', \alpha')$ denote the vector of all parameters in the marginal model (Ramroop, 2008). The marginal likelihood function is expressed as:

$$L_{ML}(\theta) = \prod_{i=1}^{N} \left\{ (2\pi)^{-\frac{n_i}{2}} |V_i(\alpha)|^{-\frac{1}{2}} \exp\big(-\frac{1}{2}(Y_i - X_i\beta)' V_i^{-1}(\alpha)(Y_i - X_i\beta) \right\} \tag{3.2}$$

given the above assumptions holds. Where $V_i(\alpha)$ is the matrix of variance components and if α is known, the MLE of β is given by:

$$\hat{\beta}(\alpha) = \left(\sum_{i=1}^{N} (X_i' W_i X_i) \right)^{-1} \left(\sum_{i=1}^{N} (X_i' W_i y_i) \right) \tag{3.3}$$

where $W_i = V_i^{-1}(\alpha)$. The expression of $\hat{\beta}(\alpha)$ in Equation 3.3 assumes that α is known otherwise an estimate of α may be used. The estimates for α_{ML} and β_{ML} can be obtained from maximizing $L_{ML}(\theta)$ with respect to θ that is with respect to α and β respectively (Verbeke and Molenberghs, 2000).

3.3.2 Restricted Maximum Likelihood Estimation (REML)

REML was first postulated by Thompson in 1962 and introduced by Patterson and Thompson in 1971. It was developed because maximum likelihood estimation of the variance components does not account for the loss of degrees of freedom used in estimating the fixed parameters. It finds a linear transformation on the response Y such that the resulting vector does not contain the fixed effects (Boldman and Van, 1991).

Consider a sample of N observations $Y_1, Y_2, \ldots, Y_N$. Suppose μ is known then the MLE of σ^2 is well known to be given by:

$$\hat{\sigma}^2 = \sum_{i=1}^{N} \frac{(Y_i - \mu)^2}{N} \tag{3.4}$$

and $\hat{\sigma}^2$ is unbiased for σ^2. However when μ is unknown the sample mean is replaced by Y and MLE is written as:

$$\hat{\sigma}^2 = \sum_{i=1}^{N} \frac{(Y_i - \bar{Y})^2}{N} \tag{3.5}$$

and $\hat{\sigma}^2$ is biased for σ^2 since $E(\hat{\sigma^2}) = \frac{N-1}{N}\sigma^2$ and furthermore the estimator underestimates the variance. However when μ is unknown an unbiased estimate for σ^2 can still be found which is given by:

$$s^2 = \sum_{i=1}^{N} \frac{(Y_i - \bar{Y})^2}{N-1}. \tag{3.6}$$

To generalize the above REML expression consider the combined formulation of the data model as:

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix} \sim N \left\{ \begin{pmatrix} \mu \\ \vdots \\ \mu \end{pmatrix}, \sigma^2 I_N \right\}$$

which can also be expressed as:

$$Y \sim N(\mu 1_N, \sigma^2 I_N),$$

where I_N is equal to the N dimensional identity matrix and 1_N is the N dimensional vector containing only ones. Now lets transform the data vector Y such that μ vanishes from the likelihood. Let

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

where A' is a $(N-1) \times N$ matrix with elements $A_{i,i} = 1$, $A_{i,i+1} = -1$ and zero elsewhere of the form:

$$A' = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix} = A'Y \sim N(0, \sigma^2 A'A).$$

Using the above transformation the REML estimate of σ^2 is given by:

$$s^2 = \sum_{i=1}^{N} \frac{(Y_i - \bar{Y})^2}{N-1}, \tag{3.7}$$

which is the unbiased estimator for σ^2 (Wener, 2009).

REML estimation can be extended to the linear mixed model as shown below.

For easy manipulation combine all models of the form $Y_i \sim N(X_i\beta, V_i)$ into one model given by $Y \sim N(X\beta, V)$, where $V_i = Z_i G Z_i' + R_i$ and

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix}, X = \begin{pmatrix} X_1 \\ \vdots \\ X_N \end{pmatrix}, V(\alpha) = \begin{pmatrix} V_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & V_N \end{pmatrix}.$$

Using a similar approach as in the simple case the data can be transformed to be orthogonal to X so that

$$\hat{Y} = A'Y \sim N(0, A'V(\alpha)A).$$

The MLE of α based on $\hat{Y}$ is known as the REML estimate and is denoted by $\alpha_{R\hat{E}ML}$. The resulting estimate $\hat{\beta}(\alpha_{REML})$ for β is denoted by $\hat{\beta_{R\hat{E}ML}}$.

We need to note that the estimates of $\alpha_{R\hat{E}ML}$ and $\hat{\beta_{R\hat{E}ML}}$ can be obtained by maximizing

$$L_{REML}(\theta) = \left| \sum_{i=1}^{N} X_i' V_i(\alpha)^{-1} X_i \right|^{-\frac{1}{2}} L_{ML}(\hat{\theta}) \tag{3.8}$$

with respect to $\theta = (\alpha', \beta')$ (Verbeke and Molenberghs, 2000; Ramroop, 2008).

Generally, REML estimation involves applying the maximum likelihood method to linear functions of Y i.e. $A'Y$ where A' is designed such that $A'Y$ does not have fixed effects. The main objective of REML is to adjust the variance without being biased which cames as a result of using ML estimation.

3.4 Parameter Estimation

3.4.1 Estimation of Fixed Effects

Let $Y_i = (Y_{i1}, Y_{i2}, Y_{i3}, \ldots, Y_{in_i})$ where Y_{ij} is the j^{th} observation of the i^{th} individual in a longitudinal study consisting of N individuals where the i^{th} individual has n_i observations. Assume the observation $Y_i \sim MVN(X_i'\beta, V_i)$. The extended likelihood formulation of this data is defined as:

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

and it has two types of fixed parameters: (1) Fixed effect regression parameter β and (2) Variance covariance parameter α not neglecting the variance covariance parameters of U_i.

To estimate the fixed parameters we maximize L by taking its partial derivative of the log-likelihood with respect to β and equating to 0. First we consider the log-likelihood contribution from a single individual. Assuming α is known we have:

$$\begin{aligned}
\frac{\partial l}{\partial \beta} =& \frac{\partial}{\partial \beta}\left(-\frac{1}{2}(Y_i - X_i\beta)' V_i^{-1}(\alpha)(Y_i - X_i\beta)\right) \\
=& \frac{\partial}{\partial \beta}\left(-\frac{1}{2}(V_i^{-1}(\alpha)Y - V_i^{-1}(\alpha)X_i\beta)'(Y_i - X_i\beta)\right) \\
=& \frac{\partial}{\partial \beta}\left(-\frac{1}{2}(Y_i' V_i^{-1}(\alpha) Y_i' V_i^{-1}(\alpha) X_i\beta - \beta' X_i' V_i^{-1}(\alpha) Y_i + \beta' X_i' V_i^{-1}(\alpha) X_i\beta)\right) \\
=& -\big(X_i' V_i^{-1}(\alpha) X_i\beta - X_i' V_i^{-1}(\alpha) Y_i\big),
\end{aligned} \tag{3.9}$$

equating $\frac{\partial l}{\partial \beta}$ to zero we have:

$$\begin{aligned}
X_i' V_i^{-1}(\alpha) X_i\beta - X_i' V_i^{-1}(\alpha) Y_i =& 0 \\
X_i' V_i^{-1}(\alpha) X_i\beta =& X_i' V_i^{-1}(\alpha) Y_i,
\end{aligned} \tag{3.10}$$

Thus:

$$\hat{\beta}(\alpha) = (X_i' W_i X_i)^{-1}(X_i' W_i Y_i). \tag{3.11}$$

Extending it to N individuals we have:

$$\hat{\beta}(\alpha) = \left(\sum_{i=1}^{N} \left(X_i' W_i X_i\right)\right)^{-1} \left(\sum_{i=1}^{N} \left(X_i' W_i Y_i\right)\right)$$

where $W_i = V_i^{-1}(\alpha)$ (Wener, 2009; Bryan, 2011)

3.4.2 Estimation of Random Effects

Consider the model shown below for each individual i:

$$Y_i = X_i\beta + Z_i U_i + \epsilon_i. \tag{3.12}$$

If the $\text{Cov}(U_i, Y_i) = GZ'$ where Y_i is the response vector and U_i is the vector of the individual specific parameters, then we have

$$\begin{bmatrix} Y \\ U \end{bmatrix} \sim N\left(\begin{bmatrix} X\beta \\ 0 \end{bmatrix}, \begin{bmatrix} V & ZG \\ GZ' & G \end{bmatrix} \right).$$

The prediction of the random effects given Y is given by the conditional mean

$$\begin{aligned} \hat{U}_i =& E(U_i|Y_i) \\ =& E(U_i) + Cov(U_i|Y_i)(Var(Y_i))^{-1}(Y_i - E(Y_i)) \\ =& GZ_i' V_i^{-1}(\alpha)(Y_i - X_i\beta). \end{aligned}$$

It can easily be shown that $E(\hat{U}_i) = 0$ and

$$Var(\hat{U}_i) = GZ_i' \left\{ W_i - W_i X_i \left(\sum_{i=1}^{N} X_i' W_i X_i \right)^{-1} X_i' W_i \right\} Z_i G'. \tag{3.13}$$

We need to note that suppose we access the variation error using Equation 3.13 the variation in $(\hat{U} - U)$ will be underestimated since this expression ignores the variability in the random effects, U_i (Laird and Ware, 1982; Verbeke and Molenberghs, 2000). Hence,

$$Var(\hat{U}_i - U_i) = G - GZ_i' W_i Z_i G + GZ_i' W_i X_i \left(\sum_{i=1}^{N} X_i' W_i X_i \right)^{-1} X_i' W_i Z_i G \tag{3.14}$$

is the most appropriate expression for variability (Bryan, 2011; Laird and Ware, 1982).

3.4.3 Estimation of Unknown Variance Components

If the covariance matrices are unknown but we know the estimate of fixed θ. Then we can say that, the estimates of G and R_i are also known. Let Var $(Y_i) = Z_i G Z_i' + R_i = V_i$ be estimated by:

$$\hat{V}_i = \hat{R}_i + Z_i \hat{G} Z_i' = \hat{W}_i^{-1}. \tag{3.15}$$

From Equation 3.15 we are then able to estimate α and U_i using the weighted least squares shown in Equation 3.11. This is done by replacing W_i with an estimate of $\hat{W}_i$. We denote these estimates by $\hat{\alpha}(\hat{\theta})$ and $\hat{U}_i(\hat{\theta})$ (Laird and Ware, 1982; Verbeke and Molenberghs, 2000).

3.5 Random Intercept Model

Consider the random intercept model:

$$Y_{ij} = X_{ij}^T \beta + a_i + \epsilon_{ij}, \tag{3.16}$$

where $i = 1, 2, \ldots, N$, $a_i \sim N(0, \sigma_a^2)$ is the random subject effect, $\epsilon_{ij} \sim N(0, \sigma^2)$ are within subject measurement errors. Further a_i and ϵ_{ij} are assumed to be independent of each other. From Equation 3.16 above we can get the mean response over time for the i^{th} individual. The conditional mean of Y_{ij} given the subject specific effect a_i is given by $E(Y_{ij}|a_i) = X_{ij}^T\beta + a_i$ and the marginal mean given by $E(Y_{ij}) = X_{ij}^T\beta$. Where β represents the mean changes over time in the population of interest, a_i represents the i^{th} individual deviation from the population mean intercept after the effects of covariates have been accounted for. The marginal response variance for each response is given by:

$$\begin{aligned} Var(Y_{ij}) =& Var(X_{ij}^T\beta + a_i + \epsilon_{ij}) \\ =& Var(a_i + \epsilon_{ij}) \\ =& \sigma_a^2 + \sigma^2 \end{aligned} \tag{3.17}$$

and the marginal covariance between any two pair of responses (Y_{ij} and Y_{ik}) is given by:

$$\begin{aligned} Cov(Y_{ij}, Y_{ik}) =& Cov(X_{ij}^T\beta + a_i + \epsilon_{ij},\ X_{ik}^T\beta + a_i + \epsilon_{ik}) \\ =& Cov(a_i + \epsilon_{ij}, a_i + \epsilon_{ik}) \\ =& Cov(a_i, a_i) \\ =& \sigma_a^2. \end{aligned} \tag{3.18}$$

Therefore:

$$Cov(Y_{ij}, Y_{ik}) = \frac{Cov(Y_{ij}, Y_{ik})}{\sqrt{Var(Y_{ij})Var(Y_{ik})}} = \frac{\sigma_a^2}{\sigma_a^2 + \sigma^2} = \rho, \tag{3.19}$$

indicating that the introduction of a random intercept a_i induces correlation among the repeated measures in longitudinal data by allowing for subject to subject variability through σ_a^2.

The random effect models for longitudinally measured or observed data were first described by Laird and Ware (1982). Implying that a model with a random intercept only leads to exchangeable or compound symmetry correlation structure given by:

$$\begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

3.6 Random Intercept and Slope Model

Consider the model:

$$Y_{ij} = \beta_1 + \beta_2 t_{ij} + a_{1i} + a_{2i} t_{ij} + \epsilon_{ij} \tag{3.20}$$

or

$$Y_{ij} = X'_{ij}\beta + Z'_{ij} a_i + \epsilon_{ij}, \tag{3.21}$$

where $X'_{ij} = (1 \quad t_{ij}), Z'_{ij} = (1 \quad t_{ij})$, $a_{1j} \sim N(0, b_{11})$ and $a_{2i} \sim N(0, b_{22})$ are the random intercept and the random slope respectively. $Cov(a_{1i}, a_{2i}) = b_{12}$ and $\epsilon_{ij} \sim N(0, \sigma^2)$ are within subject measurement errors. Note that $a_i = (a_{1i}, a_{2i})$ and $\beta = (\beta_1, \beta_2)$.

The conditional mean of Y_{ij} given the subject specific effect a_i is given by:

$$E(Y_{ij}|a_i) = \beta_1 + \beta_2 t_{ij} + a_{1i} + a_{2i} t_{ij} \tag{3.22}$$

and the marginal mean and the variance of Y_{ij} is therefore given by:

$$E(Y_{ij}) = \beta_1 + \beta_2 t_{ij}$$

and

$$Var(Y_{ij}) = h_0^2 + h_1^2 t_{ij} + 2 t_{ij} h_{01} + \sigma^2$$

respectively.

The

$$Cov(Y_{ij}, Y_{ik}) = Cov(\beta_1 + \beta_2 t_{ij} + a_{1i} + a_{2i} t_{ij} + \epsilon_{ij}, \beta_1 + \beta_2 t_{ik} + a_{1i} + a_{2i} t_{ik} + \epsilon_{ik}), \tag{3.23}$$

which can be expanded to:

$$\begin{aligned} Cov(Y_{ij}, Y_{ik}) =& E(Y_{ij}Y_{ik}) - E(Y_{ij})E(Y_{ik}) \\ =& h_0^2 + t_{ij}t_{ik}h_1^2 + h_{01}t_{ik} + h_{10}t_{ij} \\ =& h_0^2 + t_{ij}t_{ik}g_1^2 + h_{01}(t_{ik} + t_{ij}). \end{aligned} \tag{3.24}$$

Thus:

$$Corr(Y_{ij}, Y_{ik}) = \frac{h_0^2 + t_{ij}t_{ik}h_1^2 + h_{01}(t_{ij} + t_{ik})}{\sqrt{Var(Y_{ij})Var(Y_{ik})}} \tag{3.25}$$

3.7 Mean and Covariance Structure

3.7.1 Mean structure

For balanced data, an average can be calculated for each occasion separately and standard errors for the means calculated. Plots of such summary quantities can help in telling whether there is a linear or non-linear average trend. Increasing standard errors in this case can be due to drop-outs which is a common challenge in longitudinal or repeated measures data. For unbalanced data, the time scale can be discretized and simple averaging within intervals calculated (Ramroop, 2008).

3.7.2 Selection of Mean Structure

Suppose the null hypothesis of interest is given by: $H_0 : \gamma \in \theta_{\gamma,0}$, for some subspace $\theta_{\gamma,0}$ of the parameter space θ_γ of the variance components γ. Let L_{ML} denote the ML likelihood function in Equation 3.8 and $-2\ln\lambda_N$ be the likelihood ratio test which can also be defined as:

$$-2\ln\lambda_N = -2\ln\left(\frac{L_{ML}(\hat{\theta_{\gamma,0}})}{L_{ML}(\hat{\theta_\gamma})}\right), \tag{3.26}$$

where $\hat{\theta_{ML,0}}$ and $\hat{\theta_{ML}}$ are the maximum likelihood estimates obtained from maximizing L_{ML} over $\theta_{\gamma,0}$ and θ_γ respectively. It then follows from classical likelihood theory that under some regular conditions $-2\ln\lambda_N$ asymptotically follows the chi-square distribution under H_0 with degrees of freedom equal to the difference between the dimension of θ_γ and the dimension $\theta_{\gamma,0}$ (Verbeke and Molenberghs, 2009).

The LR shown in Equation 3.26 is commonly used to test the fit of the mean structures of two models where one model is a special case of the other, or nested within the other model (Verbeke and Molenberghs, 2009).

3.7.3 Covariance Structures

The main difference between a univariate regression for independent observations and a multivariate model for repeated measures is that, the results for each individual are bound to be correlated over time. There are a number of covariance structures that can be assumed to account for such correlation. A summary of some of covariance structures that can be used are listed in Table 3.1. For more details about covariance structures see (Jones, 1993).

TABLE 3.1: Summary of covariance structures

Structure	Description	No of parameters	$(i,j)^{th}$ element
CS	Compound symmetry	2	$\sigma_{ij} = \sigma_1 + \sigma^2 1(i=j), where 1(i=j) = 1$
AR(1)	Autoregressive (1)	2	$\sigma_{ij} = \sigma^2\rho^{\lvert i-j\rvert}$
UN	Unstructured	$\frac{a(a+1)}{2}$	if i=j and 0 if i $\neq$ j, $\sigma_{ij} = \sigma_{ij}$
TOEP	Toeplitz	a	$\sigma_{ij} = \sigma_{\lvert i-j\rvert+1}$
VC	Variance components	z	$\sigma_{ij} = \sigma_k^2 1(i=j)$ and i corresponds to the k^{th} effect
SP(POW)	Power spacial	2	$\sigma^2\rho^{dij}$
SP(EXP)	Exponential spatial	2	$\sigma^2 \exp\frac{-d_{ij}}{\theta}$
SP(GAU)	Gaussian spatial	2	$\sigma^2 \exp\frac{-d_{ij}}{\rho^2}$

(i) Independence Structure

Is the standard variance component which is the default. It assumes repeated measures are uncorrelated and the corresponding covariance structure for four observations per subject is given by:

$$\begin{pmatrix} \sigma_A^2 & 0 & 0 & 0 \\ 0 & \sigma_B^2 & 0 & 0 \\ 0 & 0 & \sigma_C^2 & 0 \\ 0 & 0 & 0 & \sigma_D^2 \end{pmatrix}.$$

However such a covariance and therefore correlation structure is highly improbable in longitudinal data studies. Better structures should be considered.

(ii) Compound Symmetry

This structure assumes the covariances are homogeneous. The correlation between two separate measurements is assumed to be constant no matter how far apart the measurements are. Its written as:

$$\begin{pmatrix} \sigma^2+\sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma^2+\sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma^2+\sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma^2+\sigma_1^2 \end{pmatrix}.$$

However such a structure may be unrealistic with longitudinal studies in that closer observations are more correlated than observations which are far apart.

(iii) Autoregressive Order One

This covariance structure depends on two parameters $q = 2$ and the covariance for two time points i.e. j and j' equals

$$\sigma_{jj'} = \sigma^2 \rho^{|j-j'|}$$

where ρ is the AR(1) parameter and σ^2 is the error variance thus the covariance structure is given by:

$$\sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}.$$

(iv) Unstructured

This is the most liberal covariance structure because it allows every term to be different. It is expressed as:

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 \end{pmatrix}.$$

(v.) Toeplitz

Toeplitz and AR(1) are similar in that all correlations at the same distance have

the same correlation. But with Toeplitz there is no known function relating the ρ values to the distance. The AR(1) correlation model can be estimated with a single parameter while Toeplitz model has as many parameters as there are distances. Toeplitz and AR(1) are both appropriate for evenly spaced observations where there is no possibility of error structure changing with time (Littell et al., 2002). It is given by:

$$\begin{pmatrix} \sigma^2 & \sigma_1 & \sigma_2 & \sigma_3 \\ \sigma_1 & \sigma^2 & \sigma_1 & \sigma_2 \\ \sigma_2 & \sigma_1 & \sigma^2 & \sigma_1 \\ \sigma_3 & \sigma_2 & \sigma_1 & \sigma^2 \end{pmatrix}.$$

Note: The subsequent structures discussed below relax the assumption of requiring that the observations within an individual be evenly spaced.

(vi.) The Power Spatial Covariance Structure

For this case, correlations decline with increasing spacing between points. It is expressed as:

$$\sigma^2 \begin{pmatrix} 1 & \rho^{d_{i12}} & \rho^{d_{i13}} & \rho^{d_{i14}} \\ \rho^{d_{i21}} & 1 & \rho^{d_{i23}} & \rho^{d_{i24}} \\ \rho^{d_{i31}} & \rho^{d_{di32}} & 1 & \rho^{d_{i34}} \\ \rho^{d_{i41}} & \rho^{d_{i42}} & \rho^{d_{i43}} & 1 \end{pmatrix}.$$

The notation d_{ijk} means observations Y_{ij} and Y_{ik} are actually d_{ijk} units apart and $0 < \rho < 1$.

(vii.) The Exponential Spatial Covariance Structure

The exponential spatial covariance structure is expressed as:

$$\sigma^2 \begin{pmatrix} 1 & exp(\frac{-d_{12}}{\rho}) & exp(\frac{-d_{13}}{\rho}) & exp(\frac{-d_{14}}{\rho}) \\ exp(\frac{-d_{21}}{\rho}) & 1 & exp(\frac{-d_{23}}{\rho}) & exp(\frac{-d_{24}}{\rho}) \\ exp(\frac{-d_{31}}{\rho}) & exp(\frac{-d_{32}}{\rho}) & 1 & exp(\frac{-d_{34}}{\rho}) \\ exp(\frac{-d_{41}}{\rho}) & exp(\frac{-d_{42}}{\rho}) & exp(\frac{-d_{43}}{\rho}) & 1 \end{pmatrix}.$$

(viii.) The Gaussian Spatial Covariance Structure

For Gaussian spatial covariance structure correlation decline with increasing distance between points. It's given by:

$$\sigma^2 \begin{pmatrix} 1 & exp(\frac{-d_{12}^2}{\rho^2}) & exp(\frac{-d_{13}^2}{\rho^2}) & exp(\frac{-d_{14}^2}{\rho^2}) \\ exp(\frac{-d_{21}^2}{\rho^2}) & 1 & exp(\frac{-d_{23}^2}{\rho^2}) & exp(\frac{-d_{24}^2}{\rho^2}) \\ exp(\frac{-d_{31}^2}{\rho^2}) & exp(\frac{-d_{32}^2}{\rho^2}) & 1 & exp(\frac{-d_{34}^2}{\rho^2}) \\ exp(\frac{-d_{41}^2}{\rho^2}) & exp(\frac{-d_{42}^2}{\rho^2}) & exp(\frac{-d_{43}^2}{\rho^2}) & 1 \end{pmatrix}.$$

We should note that, the major advantage of the spatial type structures over the standard AR structure (which do assume equal spaced observations) is that, they make use of the actual distance between observations which allows the modeller to be in a position to deal with unequally spaced observations within and between subjects. Toeplitz and the autoregressive order allows observations that are far apart to be less strongly correlated and the correlation between two observations is a function of the separation between the observations (Verbeke and Molenberghs, 2009).

3.7.4 Selection of Covariance Structure

Design and analysis of a study are very important stages to ensure validity of the final results and conclusions. A poorly designed and analysed study can result into poor and misleading results. As a result of this a lot of efforts is usually needed to decide on the suitable covariance structure for the data at the beginning of a statistical analysis. The commonly used criteria to help in guidance so as to came up with the best model of choice includes: (1) Akaike Information Criteria (AIC), (2) The Schwarz Bayesian Information Criteria (BIC), (3) Bozdogan Corrected Akaike Information Criterion (CAIC) and (4) Hannan and Quinn Information Criterion among others. Studies have shown that the performance of these (AIC, BIC, CAIC and HQIC) in selection of the covariance structure are not always successful in arriving to the true covariance structure. Therefore, possible effects of misspecification of the covariance structure on statistical properties of the inference needs to be considered (Ali, 2007). Ali (2007) carried simulation studies and they found out that, CAIC and BIC are competitive in terms of their ability to identify the correct covariance structure and showed outstanding performance (Barnett et al., 2010).

3.8 Selection of Random effects

For valid conditional and marginal distributions of the data the positivity constraint on Ω is required. Consider the case where $H_0 : \theta_i = 0$ and $\Omega = (0, \infty)$. The value of $\theta_i = 0$ under the null hypothesis is on the boundary of the parameter space and the distribution of the likelihood ratio test statistic λ is therefore non standard. This is a mixture of chi-squared distributions. According to Verbeke and Molenberghs (2009) testing hypothesis such as the need for random effects uses the likelihood ratio test statistic which has an asymptotic mixture of chi-square distributions under the null hypothesis rather than the classical single chi-squared distribution (Verbeke and Molenberghs, 2009).

3.9 Model Diagnostics

After fitting a model before performing inferences it is advisable to check whether all the necessary model assumptions are valid to avoid wrong inference. Some of the areas which can make a model to be incorrect includes: the linear predictor, the variance function, the link function and the data since it may contain outliers or inferential observations. After the assumptions, model diagnostic procedures which do involve both graphical methods and formal statistical tests allows the exploration to find out whether the assumptions are valid and to see whether we can base our argument on the subsequent inference results. For further details about model diagnostics see Kutner et al. (2004).

The next chapter shows the application of linear mixed model to the Sinikithemba study data.

Chapter 4

Applications to Longitudinal HIV Bio-marker Data with Genetic Covariate Information

4.1 Introduction

Generally to come-up with the best fitting model for the data is not an easy task. Model selection is one of the major challenges faced in data analysis. Model building will be done using the stepwise procedure in R version 2.15.3 using the lme4 library. Square root CD4 count and log viral load will be used as the response variables. Modelling repeated measurements or longitudinal data requires appropriate specification of the mean and the covariance structure as also discussed in Verbeke and Molenberghs (2009). This is because misspecification of the covariance structure for repeated measures in longitudinal analysis may lead to biased estimates of the model parameters.

4.2 Random Effects

The variability in the subject-specific intercepts and slopes may not be completely explained by the covariates in the model but can be solved using random effects (Verbeke and Molenberghs, 2009). Random effects model helps in accounting for the extra variability due to individual to individual heterogeneity.

In order to choose the best model for our analysis, the random intercept and slope model, the random intercept model and the model with no random effects for square root CD4 count and log viral load data were compared. As illustrated in Table 4.1 and

Table 4.2 the random intercept model is the best fitting model for our analysis for the square root CD4 count and the log viral load data, (AIC=14136.08) and (AIC=7753.59) respectively.

TABLE 4.1: Fit criteria for CD4 count comparing random effects

	-2logLikelihood	AIC	BIC
Random Intercept	1737.26	14136.08	14842.37
Random Intercept and Slope	1880.3	15280.39	15986.68
No Random Effects	1971.36	16006.88	16712.08

TABLE 4.2: Fit criteria for log viral load comparing random effects

	-2logLikelihood	AIC	BIC
Random Intercept	1891.37	7753.59	8402.03
Random Intercept and Slope	1881.9	7791.49	8364.13
No Random effects	1936.33	7969.29	8581.77

4.3 Square Root CD4 Count as the Response

The model (linear mixed model) uses square root CD4 count as the response. A linear mixed model is a model that contain both the fixed and the random effects term. Fixed effects are the effects attributable to a finite set of levels of a factor that occur in the data while random effects are attributable to an infinite level of a factor, of which only a random sample are allowed to occur in the data (McCulloch, 2006; Verbeke and Molenberghs, 2005). CD4 count is transformed because square root CD4 count better approximates the normal distribution compared to the original CD4 count. The independent variables to be included in the model are: APOBEC3G, HLA-B*0801, HLA-B*1510, HLA-B*4201, HLA-B*5801, HLA-B*5802, $PPIA_{1650}$, $PSIP_{rs12339417}$, TNPO3, IL-10(-3575, -1082, -592), gender, time and the two way interactions between the covariates. The single nucleotide polymporphisms (APOBEC3G, TNPO3, $PPIA_{1650}$ and $PSIP_{rs12339417}$) have three categories namely: (CC, CG, GG), (AA, AG, GG), (AA, AG, GG) and (CC, CT, TT) respectively. Interleukin-10(-3575, -592, -1082) also has three categories each namely: (AA, TA, TT), (AA, AG, GG) and (AA,CA, CC) respectively.

Our objective in this chapter is to come-up with a model that best fits and describes the data. The best covariance structure is chosen depending on the AICs' (Akaike Information Criterion) for the different models. The model with the smallest AIC is considered the best. Compound symmetry (CS), spatial exponential (sp(exp)) and Gaussian (sp(gau)) are some of the covariance structures whose AIC will be compared. Restricted maximum likelihood method (REML) and maximum likelihood (ML) are the estimation methods that will be used. The difference in the value of model fit statistics between REML and ML increases as the number of fixed effects in the model increases. The AIC

and BIC for REML and ML are different because REML takes into account degrees of freedom used in estimating fixed effects mean parameters while ML does not.

4.3.1 Covariance Structure

There is a high probability of ending up with a more complicated covariance structure if the mean structure is not properly chosen. Therefore, its advisable to choose the best covariance structure. This is because (1) If the chosen covariance structure is simple then, there would be an increase in the rate of type I error while if it is too complex then, there would be decrease in power and efficiency (Bentler and Bonett, 1980). To choose the best covariance structure we fitted a linear mixed model with square root CD4 count as the response and APOBEC3G, $PPIA_{1650}$, $PSIP_{rs12339417}$, TNPO3, all the HLA-B types, IL-10(-3575), IL-10(-592), IL-10(-1082), time, gender and all the two way interactions between the covariates as the independent variables. Higher order interactions were not included to avoid increased model complexity already encountered by having two way interactions and main effects.

Maximum likelihood (ML) and restricted maximum likelihood method (REML) are the estimation methods to be used. Independence structure, autoregressive and Toeplitz covariance structures were inappropriate for our data because: (1) Independence structure assumes that repeated measures are uncorrelated which is unrealistic with a longitudinal study. (2) Toeplitz structure assumes that correlations between equally distant points is constant which most likely not the case with our study given we had fairly long individual sequence of observations. (3) Autoregressive structure assumes that the measurements are equally spaced which is not the case with our study because of miss-timed visits and missing values which exacerbated the coarseness of the data. We compared the AIC's of compound symmetry, spatial Gaussian and spatial exponential covariance structures. We used AIC instead of the likelihood ratio test because the chosen structures were unnested within each other and also the likelihood ratio test would not be valid under REML. Table 4.3 and Table 4.4 gives a summary of logLikelihood, AIC and BIC for the chosen structures using REML and ML as methods of estimation respectively. As illustrated in Table 4.3 and Table 4.4, spatial exponential structure is the best covariance structure for our analysis under both REML and ML methods of estimation with AIC=13,439.24 for REML and 13670.21 for ML.

Table 4.3: Model fit by REML for CD4 count

	CS	SP(EXP)	SP(GAU)
-2logLikelihood	3474.81	3299.81	3349.61
AIC	14138.08	13439.24	13.638.43
BIC	14850.31	14151.46	14350.66

TABLE 4.4: Model fit by ML for CD4 count

	CS	SP(EXP)	SP(GAU)
-2logLikelihood	3529.39	3357.55	3349.61
AIC	14357	13670.21	13638.43
BIC	15074.7	14387.36	14350.66

4.3.2 Mean Structure

To come-up with the mean structure, we fitted a full model with maximum likelihood method as the method of estimation. What happens in this step is, from the Type III effects terms are continually removed from the model in a systematic manner starting from the least significant term. The new simple model is then compared to the original model using likelihood ratio test. The original model is used over the new model if it's not significantly different from the new model (p-value>0.05) Bryan (2011). This process continues and only stops if the remaining terms in the model are all significant. Terms which are found to be insignificant but their interactions are significant are included back in the final model Bryan (2011). The final analysis is then done using the best covariance structure and mean structure. Table 4.5 shows the final model with square root CD4 count as the response, including APOBEC3G, HLA-B types, $PPIA_{1650}$, $PSIP_{rs12339417}$, time, gender, TNPO3, IL-10(-3575, -592, -1082), and the two way interactions between the covariates.

TABLE 4.5: Type III tests of fixed effects for the final model with square root CD4 count as the response

	numDF	denDF	F value	p-value		numDF	denDF	F value	p-value
(Intercept)	1	2648	20.244391	<.0001	HLA-B*0801:IL-10(-592)	2	146	7.854490	0.0006
APOBEC3G	2	146	8.395758	0.0004	HLA-B*1510:PPIA(1650)	2	146	3.530057	0.0318
HLA-B*0801	1	146	0.196582	0.6582	HLA-B*1510:PSIP(rs12339417)	2	146	2.553794	0.0813
HLA-B*1510	1	146	0.747948	0.3885	HLA-B*1510:time	1	2648	0.934748	0.3337
HLA-B*4201	2	146	0.378467	0.6856	HLA-B*1510:IL-10(-3575)	2	146	0.018580	0.9816
HLA-B*5801	2	146	3.247277	0.0417	HLA-B*1510:IL-10(-592)	2	146	0.771113	0.4644
HLA-B*5802	2	146	0.978292	0.3784	HLA-B*1510:IL-10(-1082)	2	146	0.093316	0.9110
PPIA(1650)	2	146	1.982185	0.1415	HLA-B*5802:time	2	2648	3.802323	0.0224
PSIP(rs12339417)	2	146	2.653502	0.0738	PPIA(1650):PSIP(rs12339417)	4	146	0.753139	0.5575
Time	1	2648	5.830202	0.0158	PPIA(1650):time	2	2648	1.560449	0.2102
Gender	1	146	3.544032	0.0617	PPIA(1650):IL-10(-592)	4	146	4.105113	0.0035
TNPO3	2	146	2.395751	0.0947	PPIA(1650):IL-10(-1082)	4	146	1.696247	0.1540
IL-10(-3575)	2	146	4.374254	0.0143	PSIP(rs12339417):gender	2	146	1.719469	0.1828
IL-10(-592)	2	146	1.155152	0.3179	PSIP(rs12339417):IL-10(-3575)	4	146	2.965596	0.0216
IL-10(-1082)	2	146	2.272595	0.1067	PSIP(rs12339417):IL-10(-592)	4	146	0.447877	0.7738
APOBEC3G:HLA-B*0801	2	146	0.023573	0.9767	PSIP(rs12339417):IL-10(-1082)	4	146	2.176684	0.0745
APOBEC3G:PSIP(rs12339417)	4	146	3.069618	0.0183	time:TNPO3	2	2648	0.794155	0.4521
APOBEC3G:gender	2	146	0.720092	0.4884	time:IL-10(-3575)	2	2648	11.125808	<.0001
APOBEC3G:IL-10(-592)	4	146	0.752768	0.5577	time:IL-10(-592)	2	2648	2.660453	0.0701
APOBEC3G:IL-10(-1082)	4	146	1.333309	0.2603	time:IL-10(-1082)	2	2648	1.991641	0.1367
HLA-B*0801:HLA-B*1510	1	146	10.310074	0.0016	gender:TNPO3	2	146	1.645871	0.1964
HLA-B*0801:PPIA(1650)	2	146	7.745759	0.0006	gender:IL-10(-3575)	2	146	2.034496	0.1344
HLA-B*0801:PSIP(rs12339417)	2	146	2.311127	0.1028	gender:IL-10(-592)	2	146	0.488078	0.6148
HLA-B*0801:time	1	2648	0.634662	0.4257	gender:IL-10(-1082)	2	146	0.615321	0.5419
HLA-B*0801:gender	1	146	11.534964	0.0009	TNPO3:IL-10(-3575)	4	146	2.006605	0.0966
HLA-B*0801:TNPO3	2	146	4.641411	0.0111	TNP03:IL-10(-1082)	4	146	1.313501	0.2677

As indicated in Table 4.5, we see that, at 5% significant level, APOBEC3G (p=0.000), HLA-B*5801 (p=0.042), time (p=0.016) and IL-10(-3575) (p=0.014) are significantly associated with mean square root CD4 count. The two way interactions between APOBEC3G and $PSIP_{rs12339417}$ (p=0.018), HLA-B*0801 and HLA-B*1510 (p=0.002), HLA-B*0801

and $PPIA_{1650}$ (p=0.001), HLA-B*0801 and gender (p=0.001), HLA-B*0801 and IL-10(-592) (p=0.001), HLA-B*0801 and TNPO3 (p=0.011), $PPIA_{1650}$ and IL-10(-592) (p=0.004), $PSIP_{rs12339417}$ and IL-10(-3575) (p=0.022) and lastly time and IL-10(-3575) (p<.000) are also significantly associated with mean square root CD4 count. The significant two way interaction involving time implies that the slope or increase in mean square root CD4 count differs per unit time for the different levels of the categorical factor.

As illustrated in Table 4.6, the reference categories are: gender:female, APOBEC3G: CC genotype, $PPIA_{1650}$: AA genotype, $PSIP_{rs12339417}$: CC genotype, TNPO3: AA genotype, IL-10(-3575): AA genotype, IL-10(-592): AA genotype and IL-10(-1082): AA genotype and absent for the HLA-B alleles.

As indicated in Table 4.6 individuals with APOBEC3G (CG) genotype have a mean square root CD4 count 11.275 units higher compared to those with the CC genotype (p=0.002). This implies that, APOBEC3G (CG) genotype could be associated with a slow progression of HIV. The time estimate shows that for every unit increase in time the mean square root CD4 count decreases by 0.072 units. The two way interaction between APOBEC3G and $PSIP_{rs12339417}$ shows that, the effect of APOBEC3G depends on the level of $PSIP_{rs12339417}$ and vice-versa. Individuals with interaction between APOBEC3G (CG) genotype and $PSIP_{rs12339417}$ CT genotype have a mean square root CD4 count 8.227 units lower (p=0.010) compared to those with APOBEC3G CC genotype interacting with $PSIP_{rs12339417}$ CC genotype. The interaction between APOBEC3G (CG) genotype and $PSIP_{rs12339417}$ TT genotype also showed a decrease in the mean square root CD4 count by 7.260 units (p=0.024) compared to those individuals with APOBEC3G CC genotype interacting with $PSIP_{rs12339417}$ CC genotype.

The interaction between HLA-B*0801 and HLA-B*1510 showed a decrease in the mean square root CD4 count by 26.292 units (p=0.014) compared to those without. The interaction between HLA-B*0801 and $PPIA_{1650}$ shows that, the effect of HLA-B*0801 depends on the level of $PPIA_{1650}$ and vice-versa. Individuals who experienced an interaction between HLA-B*0801 and $PPIA_{1650}$ AG genotype showed a decrease in the mean square root CD4 count by 11.723 units (p=0.006). The interaction between HLA-B*0801 and gender showed that, the presence of the HLA-B*0801 allele in males lead to decrease in the mean square root CD4 count by 17.074 units (p=0.010). Therefore, HLA-B*0801 allele in males could be associated with a fast progression of HIV. The interaction between HLA-B*0801 allele and TNPO3 shows that, the effect of HLA-B*0801 depends on the level of TNPO3 and vice-versa. Interaction between HLA-B*0801 and TNPO3 AG genotype showed a decrease in the mean square root CD4 count by 11.830 units (p=0.020). The interaction between HLA-B*0801 allele and IL-10(-592) show that

the effect of HLA-B*0801 depends on the level of IL-10(-592) and vice-versa. The interaction between HLA-B*0801 allele and IL-10(-592) CA genotype showed an increase in the mean square root CD4 count by 13.533 units (p=0.030).

Interaction between HLA-B*1510 and $PPIA_{1650}$ show that the effect of HLA-B*1510 depends on the level of $PPIA_{1650}$ and vice-versa. Individuals with HLA-B*1510 interacting with $PPIA_{1650}$ GG genotype showed an increase in the mean square root CD4 count by 6.814 units (p=0.036). Absence of HLA-B*5802 allele with time showed a decrease in the mean square root CD4 count by 0.233 units (p=0.001). Interaction between $PSIP_{rs12339417}$ and IL-10(-3575) shows that, the effect of $PSIP_{rs12339417}$ depends on the level of IL-10(-3575) and vice-versa. Individuals with interaction between $PSIP_{rs12339417}$ CT genotype and IL-10(-3575) TT genotype showed an increase in the mean square root CD4 count by 16.133 units (p=0.048). The interaction between IL-10(-592) CA genotype with time showed an increase in the mean square root CD4 count by 0.044 units (p=0.031). This implies that as time elapsed individuals with IL-10(-592) CA genotype are likely to progress to HIV at a slower rate compared to those with IL-10(-592) AA genotype.

4.3.3 Diagnostic Analysis of the Model with Random Intercept Term

The histogram on Figure 4.1 indicates normality for the data. The quantile-quantile plots shows some slight but systematic and symmetrical deviation from a straight line. Suggesting that, the random intercept model is appropriate for our analysis.

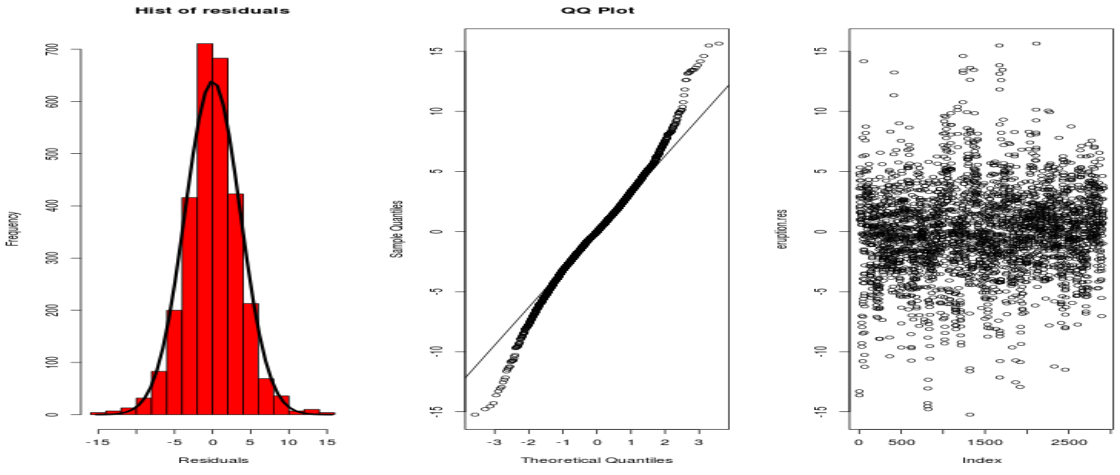

FIGURE 4.1: Model diagnostics for square root CD4 count as the response

4.4 Log Viral Load

The same procedure used to analyse square root CD4 count was used to analyse the log viral load.

4.4.1 Covariance Structure

To identify the best covariance structure, we used the full model with log viral load as the response. Gender, time, HLA-B types, SNPs (APOBEC3G, TNPO3, $PPIA_{1650}$, $PSIP_{rs12339417}$, IL-10(-592, -1082, -3575) and the two way interactions as the covariates. We compared the AIC between spatial exponential, spatial Gaussian and compound symmetry covariance structures. Restricted maximum likelihood method (REML) and maximum likelihood (ML) were the estimation methods used. We identified spatial exponential covariance structure as the best covariance structure for our data (AIC=7297.81) with maximum likelihood method as the method of estimation. See Table 4.9 and Table 4.10.

4.4.2 Mean Structure

The same procedure used to determine the mean structure for square root CD4 count was applied to determine the mean structure for the log viral load. The mean structure was chosen using the Type III tests for fixed effects. The final mean structure model includes: APOBEC3G, all HLA-B types, $PPIA_{1650}$, $PSIP_{rs12339417}$, time, gender, TNPO3, IL-10(-3575), IL-10(-592), IL-10(-1082) and the two way interactions between the covariates. See Table 4.11.

It can be seen from Table 4.11 that, at 5% level of significance APOBEC3G (p=0.060), time (p=< .000), gender (p=0.006) and IL-10(-592) (p=0.001) are significantly associated with log viral load. The two way interactions between APOBEC3G and HLA-B*0801 (p=0.037), $PPIA_{1650}$ and $PSIP_{rs12339417}$ (p=0.045), gender and IL-10(-3575) (p=0.002) and lastly between gender and IL-10(-592) (p=0.006) are also significantly associated with log viral load.

As indicated in Table 4.12 gender is observed to be associated with log viral load (p=0.008). The estimated value (-6.476) shows that, females have a predicted log viral load that is 6.476 units lower compared to males. This implies that, females possibly progress to AIDS slightly at a lower rate compared to males. Individuals with APOBEC3G CG genotype have a predicted log viral load which is 2.378 units lower compared to those with the CC genotype (p=0.041). Individuals with HLA-B*5802 allele have a predicted log viral load of 0.883 units higher compared to those without the

same allele (p=0.040). Therefore, individuals with the HLA-B*5802 allele could progress to AIDS faster compared to those without the same allele. Individuals with IL-10(-592) CA genotype have a predicted log viral load of 5.522 units lower (p=0.000) compared to those with IL-10(-592) AA genotype. Individuals with IL-10(-592) CC genotype also have a predicted log viral load of 4.961 units lower (p=0.003) compared to those with IL-10(-592) AA genotype. Male individuals with APOBEC3G CG genotype have a predicted log viral load of 2.074 units lower (p=0.021) compared to those without. Implying that, APOBEC3G (CG) genotype could be a good controller of HIV in males.

The two way interaction between APOBEC3G CG genotype and HLA-B*0801 allele shows that, the presence of APOBEC3G CG genotype and HLA-B*0801 allele interactions lead to decrease of the predicted log viral load by 2.251 units (p=0.030). The interaction between $PPIA_{1650}$ and IL-10(-592) shows that, the effect of $PPIA_{1650}$ depends on the level of IL-10(-592) and vice-versa. The interaction between $PPIA_{1650}$ GG genotype and IL-10(-592) CA genotype showed an increase in the predicted log viral load by 2.543 units (p=0.030) higher compared to those with $PPIA_{1650}$ AA genotype interacting with IL-10(-592) AA genotype. Individuals with interactions between $PSIP_{rs12339417}$ and IL-10 (-592) shows that, the effect of $PSIP_{rs12339417}$ depends on the level of IL-10(-592). Those individuals with $PSIP_{rs12339417}$ CT genotype interacting with IL-10(-592) CC genotype have a predicted log viral load of 3.530 units higher (p=0.020) compared to those with $PSIP_{rs12339417}$ CC genotype interacting with IL-10(-592) AA genotype. The interaction between IL-10(-3575) TA genotype and gender showed that, male individuals with such interactions have a predicted log viral load of 4.362 units higher (p=0.020) compared to the females. The interaction between IL-10(-3575) TT genotype and gender showed that, male individuals with such interactions have a predicted log viral load of 6.377 units higher (p=0.001) compared to the females. Male individuals with IL-10(-592) CA genotype have a predicted log viral load of 3.528 units higher (p=0.007) compared to the females. The individuals who progress to AIDS faster are referred to as fast progressors while those who progress to AIDS at a lower rate are referred to as elite controllers.

4.4.3 Diagnostic Analysis for the model Including the Random Intercept Term

The histogram and the quantile-quantile plot for the residuals of the fitted model in Figure 4.3 shows that, the fitted model is the best for our data. This is because, the histogram does not show much skewness and the quantile-quantile plot of the residual only shows a slight deviation from the straight line.

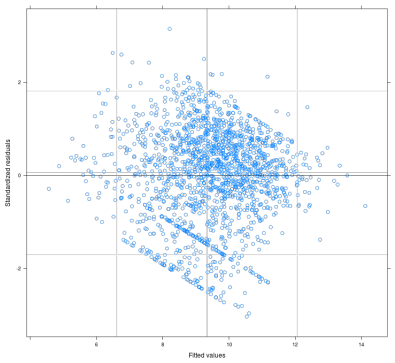

FIGURE 4.2: Residual plot

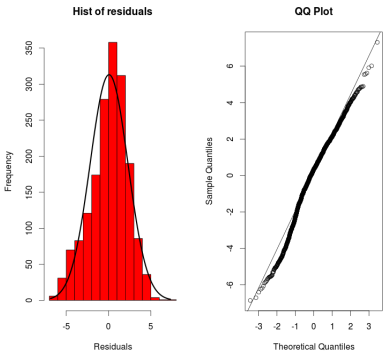

FIGURE 4.3: QQ and the histogram plots for residuals

In this chapter, we started by identifying the best model for our data. We went ahead and identified the best covariance and mean structures for our data. Then, we fitted the random intercept model with our chosen covariance and mean structure. We conclude the chapter by carrying out model diagnostics for our chosen model. We did all these so as to reaffirm previous studies that have shown that, immuno-genetic factors could play a role in susceptibility, transmission, progression and even response to antiviral therapy. Which was illustrated in these chapter from the findings we achieved.

As stated earlier a prevalent and common problem encountered with longitudinal or repeated measurement studies is that of drop-out or missing values in general. This problem if not handled properly can lead to biased effect estimates and inference. In the next chapter we demonstrate the use of multiple imputation as a powerful tool to deal with this problem.

TABLE 4.6: Solution for fixed effects with square root CD4 count as the response

	Value	Std. Error	df	t-value	p-value
(Intercept)	25.658745	7.611542	2648	3.371031	0.0008
APOBEC3G					
CC	Ref				
CG	11.274612	3.635709	146	3.101077	0.0023
GG	1.404015	6.398505	146	0.219429	0.8266
HLA-B					
0801(absent)	Ref				
0801	-2.492310	7.122443	146	-0.349923	0.7269
1510(absent)	Ref				
1510	-8.815121	12.724679	146	-0.692758	0.4896
4201(absent)	Ref				
4201	2.243414	3.304191	146	0.678960	0.4982
5801(absent)	Ref				
5801	-1.576985	1.431460	146	-1.101662	0.2724
5802(absent)	Ref				
5802	-0.288599	1.113713	146	-0.259132	0.7959
PPIA(1650)					
AA	Ref				
AG	-0.965303	3.967628	146	-0.243295	0.8081
GG	7.431028	5.190007	146	1.431795	0.1543
PSIP(rs12339417)					
CC	Ref				
CT	-15.222616	9.217205	146	-1.651544	0.1008
TT	-5.223732	8.730656	146	-0.598321	0.5506
time	-0.072254	0.031274	2648	-2.310370	0.0209
Gender					
Female	Ref				
Male	14.591454	10.080466	146	1.447498	0.1499
TNPO3					
AA	Ref				
AG	-10.616059	6.834190	146	-1.553375	0.1225
GG	-12.223412	14.724912	146	0.830118	0.4078
IL-10(-3575)					
AA	Ref				
TA	-5.223965	5.523744	146	-0.945729	0.3458
TT	-11.729954	6.200654	146	-1.891728	0.0605
IL-10(-592)					
AA	Ref				
CA	-2.052666	4.646096	146	-0.441804	0.6593
CC	1.436878	4.863856	146	0.295420	0.7681
IL-10(-1082)					
AA	Ref				
AG	-3.701354	3.377718	146	-1.095815	0.2750
GG	-8.856794	5.605192	146	-1.580105	0.1162
APOBEC3G:HLA-B*0801					
CC:HLA-B*0801(absent)	Ref				
CG:HLA-B*0801	0.511706	3.844880	146	0.133088	0.8943
GG:HLA-B*0801	0.764798	6.068091	146	0.126036	0.8999
APOBEC3G:PSIP(rs12339417)					
CC:CC	Ref				
CG:CT	-8.226570	3.153083	146	-2.609056	0.0100
GG:CT	-2.935321	4.318715	146	-0.679675	0.4978
CG:TT	-7.259722	3.179098	146	-2.283579	0.0238
GG:TT	-0.848703	4.514648	146	-0.187989	0.8511
APOBEC3G:Gender					
CC:Female	Ref				
CG:male	-0.988960	2.676843	146	-0.369450	0.7123
GG:male	2.910994	3.580181	146	0.813086	0.4175
APOBEC3G:IL-10(-592)					
CC:AA	Ref				
CG:CA	-3.576917	2.691488	146	-1.328974	0.1859
GG:CA	-1.014625	6.049333	146	-0.167725	0.8670
CG:CC	-2.713960	3.027243	146	-0.896512	0.3715
GG:CC	-0.161092	6.574955	146	-0.024501	0.9805
APOBEC3G:IL-10(-1082)					
CC:AA	Ref				
CG:AG	-0.671825	1.884822	146	-0.356439	0.7220
GG:AG	-1.111505	3.020351	146	-0.368005	0.7134
CG:GG	-5.092057	3.122687	146	-1.630665	0.1051
GG:GG	-1.023992	8.684694	146	-0.117908	0.9063

TABLE 4.7: Solution for fixed effects with square root CD4 count as the response (Continuation of Table 4.6)

	Value	Std. Error	df	t-value	p-value
HLA-B*0801:HLA-B*1510					
HLA-B*0801:HLA-B*1510(absent)	Ref				
HLA-B*0801:HLA-B*1510	-26.291995	10.610336	146	-2.477961	0.0144
HLA-B*0801:PPIA(1650)					
HLA-B*0801(absent):AA	Ref				
HLA-B*0801:AG	-11.723796	4.210339	146	-2.784525	0.0061
HLA-B*0801:GG	-4.245608	5.429383	146	-0.781969	0.4355
HLA-B*0801:PSIP(rs12339417)					
HLA-B*0801(absent):CC	Ref				
HLA-B*0801:CT	6.128774	4.938860	146	1.240929	0.2166
HLA-B*0801:TT	0.639590	5.336409	146	0.119854	0.9048
HLA-B*0801:time	-0.010231	0.016853	2648	-0.607109	0.5438
HLA-B*0801:Gender					
HLA-B*0801(absent):female	Ref				
HLA-B*0801:male	17.073820	6.571002	146	2.598359	0.0103
HLA-B*0801:TNPO3(AG)					
HLA-B*0801(absent):AA	Ref				
HLA-B*0801:AG	-11.829535	5.027773	146	-2.352838	0.0200
HLA-B*0801:GG	-1.069339	9.363622	146	-0.114201	0.9092
HLA-B*0801:IL-10(-592)					
HLA-B*0801(absent):AA	Ref				
HLA-B*0801:CA	13.533450	6.158832	146	2.197405	0.0296
HLA-B*0801:CC	2.660494	7.294973	146	0.364702	0.7159
HLA-B*1510:PPIA(1650)					
HLA-B*1510(absent):AA	Ref				
HLA-B*1510:AG	2.696730	2.847278	146	0.947126	0.3451
HLA-B*1510:GG	6.814130	3.215543	146	2.119123	0.0358
HLA-B*1510:PSIP(rs12339417)					
HLA-B*1510(absent):CC	Ref				
HLA-B*1510:CT	8.577531	5.504864	146	1.558173	0.1214
HLA-B*1510:TT	9.660170	5.503397	146	1.755310	0.0813
HLA-B*1510:time	-0.013203	0.015557	2648	-0.848737	0.3961
HLA-B*1510:IL-10(-3575)					
HLA-B*1510(absent):AA	Ref				
HLA-B*1510:TA	-1.317443	9.565703	146	-0.137726	0.8906
HLA-B*1510:TT	-1.338797	10.108063	146	-0.132448	0.8948
HLA-B*1510:IL-10(-592)					
HLA-B*1510(absent):AA	Ref				
HLA-B*1510:CA	0.325000	5.674339	146	0.057275	0.9544
HLA-B*1510:CC	-2.269404	6.307898	146	-0.359772	0.7195
HLA-B*1510:IL-10(-1082)					
HLA-B*1510(absent):AA	Ref				
HLA-B*1510:AG	-1.349636	3.400919	146	-0.396844	0.6921
HLA-B*1510:GG	-0.965155	7.310386	146	-0.132025	0.8951
HLA-B*5802:time	-0.017827	0.017446	2648	-1.021847	0.3069
PPIA(1650):PSIP(rs12339417)					
AA:CC	Ref				
AG:CT	-2.057613	2.835491	146	-0.725664	0.4692
GG:CT	-2.839094	4.277255	146	-0.663765	0.5079
AG:TT	-3.332023	2.849508	146	-1.169333	0.2442
GG:TT	-2.459026	4.408811	146	-0.557753	0.5779
PPIA(1650):time					
AA:time	Ref				
AG:time	-0.018309	0.012208	2648	-1.499781	0.1338
GG:time	-0.016052	0.013046	2648	-1.230452	0.2186
PPIA(1650):IL-10(-592)					
AA:AA	Ref				
AG:CA	6.206435	3.667708	146	1.692183	0.0927
GG:CA	-3.696112	3.631227	146	-1.017868	0.3104
AG:CC	2.096196	3.835001	146	0.546596	0.5855
GG:CC	-2.710289	3.742090	146	-0.724272	0.4701
PPIA(1650):IL-10(-1082)					
AA:AA	Ref				
AG:AG	0.444817	1.905650	146	0.233420	0.8158
GG:AG	-3.391112	2.189075	146	-1.549107	0.1235
AG:GG	2.724601	3.100708	146	0.878703	0.3810
GG:GG	2.817028	7.691356	146	0.366259	0.7147

TABLE 4.8: Solution of fixed effects with square root CD4 count as the response (Continuation of Table 4.6)

	Value	Std. Error	df	t-value	p-value
PSIP(rs12339417):Gender					
PSIP(rs12339417)CC:Female	Ref				
CT:male	2.287070	3.278342	146	0.697630	0.4865
TT:male	-0.822195	3.323050	146	-0.247422	0.8049
PSIP(rs12339417):IL-10(-3575)					
CC:AA	Ref				
CT:TA	9.063041	7.611045	146	1.190775	0.2357
TT:TA	2.493441	6.792622	146	0.367081	0.7141
CT:TT	16.133496	8.100716	146	1.991613	0.0483
TT:TT	5.419592	7.340281	146	0.738336	0.4615
PSIP(rs12339417):IL-10(-592)					
CC:AA	Ref				
CT:CA	2.736245	4.640692	146	0.589620	0.5564
TT:CA	2.573761	4.962902	146	0.518600	0.6048
CT:CC	2.053439	4.894507	146	0.419539	0.6754
TT:CC	0.724341	5.180713	146	0.139815	0.8890
PSIP(rs12339417):IL-10(-1082)					
CC:AA	Ref				
CT:AG	6.039493	3.464668	146	1.743167	0.0834
TT:AG	4.965290	3.373936	146	1.471661	0.1433
CT:GG	10.625176	6.496142	146	1.635613	0.1041
TT:GG	10.361931	5.621231	146	1.843356	0.0673
time:TNPO3					
time:AA	Ref				
time:AG	0.003459	0.012158	2648	0.284532	0.7760
time:GG	-0.027272	0.024610	2648	-1.108184	0.2679
time:IL-10(-3575)					
time:AA	Ref				
time:TA	-0.028140	0.022702	2648	-1.239535	0.2153
time:TT	0.037050	0.026684	2648	1.388490	0.1651
time:IL-10(-592)					
time:AA	Ref				
time:CA	0.043568	0.020147	2648	2.162554	0.0307
time:CC	0.030721	0.020601	2648	1.491225	0.1360
time:IL-10(-1082)					
time:AA	Ref				
time:AG	0.027772	0.014827	2648	1.873120	0.0612
time:GG	0.034201	0.023104	2648	1.480320	0.1389
Gender:TNPO					
Female:AA	Ref				
male:AG	-3.251797	2.387810	146	-1.361833	0.1753
male:GG	1.425966	6.848622	146	0.208212	0.8354
Gender:IL-10(-3575)					
Female:AA	Ref				
male:TA	-13.019287	8.871321	146	-1.467570	0.1444
male:TT	-11.388263	9.518929	146	-1.196381	0.2335
Gender:IL-10(-592)					
Female:AA	Ref				
male:CA	-1.550495	3.310529	146	-0.468353	0.6402
male:CC	0.052177	3.622757	146	0.014403	0.9885
Gender:IL-10(-1082)					
Female:AA	Ref				
male:AG	-1.442543	2.902552	146	-0.496991	0.6199
male:GG	-4.773543	5.943686	146	-0.803128	0.4232
TNPO:IL-10(-3575)					
AA:AA	Ref				
AG:TA	10.331271	6.132561	146	1.684659	0.0942
GG:TA	3.121368	8.372489	146	0.372812	0.7098
AG:TT	13.053418	6.723517	146	1.941457	0.0541
GG:TT	14.274675	13.199758	146	1.081435	0.2813
TNPO:IL-10(-1082)					
AA:AA	Ref				
AG:AG	0.924407	2.728187	146	0.338836	0.7352
GG:AG	-2.951881	6.589517	146	-0.447966	0.6548
AG:GG	3.611688	5.155994	146	0.700483	0.4847
GG:GG	12.799510	13.432964	146	0.952843	0.3422

TABLE 4.9: Model fit by REML for log viral load

	CS	SP(EXP)	SP(GAU)
logLikelihood	-3782.74	-3538.34	-3600.02
AIC	7793.49	7304.67	7428.04
BIC	8409.43	7920.62	8043.98

TABLE 4.10: Model fit by ML for log viral load

	CS	SP(EXP)	SP(GAU)
logLikelihood	-3773.66	-3534.91	-3592.05
AIC	7775.31	7297.81	7412.09
BIC	8393.72	7921.22	8035.50

TABLE 4.11: Type III tests of fixed effects for log viral load as the response

	numDF	denDF	F value	p-value		numDF	denDF	F value	p-value
(Intercept)	1	1472	36.46782	<.0001	HLA-B*0801:time	1	1472	0.01757	0.8946
APOBEC3G	2	169	2.86118	0.0600	HLA-B*0801:gender	1	169	0.10700	0.7440
HLA-B*0801	1	169	2.15296	0.1442	HLA-B*0801:TNPO3	2	169	1.39873	0.2498
HLA-B*1510	1	169	0.26605	0.6067	HLA-B*0801:IL-10(-592)	2	169	1.33493	0.2659
HLA-B*4201	2	169	1.86444	0.1582	HLA-B*1510:PSIP(rs12339417)	2	169	1.83939	0.1621
HLA-B*5801	2	169	1.70658	0.1846	HLA-B*1510:time	1	1472	0.00411	0.9489
HLA-B*5802	2	169	2.29350	0.1040	HLA-B*1510:gender	1	169	0.11492	0.7350
PPIA(1650)	2	169	0.30467	0.7378	HLA-B*1510:IL-10(-3575)	2	169	0.12404	0.8834
PSIP(rs12339417)	2	169	0.59322	0.5537	HLA-B*1510:IL-10(-592)	2	169	0.46556	0.6286
Time	1	1472	17.15192	<.0001	HLA-B*1510:IL-10(-1082)	2	169	0.28991	0.7487
Gender	1	169	7.79243	0.0059	HLA-B*4201:time	2	1472	0.67031	0.5117
TNPO3	2	169	0.65697	0.5197	HLA-B*5802:time	2	1472	2.59375	0.0751
IL-10(-3575)	2	169	0.09084	0.9132	PPIA(1650):PSIP(rs12339417)	4	169	2.49126	0.0451
IL-10(-592)	2	169	7.05669	0.0011	PPIA(1650):time	2	1472	1.19791	0.3021
IL-10(-1082)	2	169	0.08436	0.9191	PPIA(1650):IL-10(-592)	4	169	1.65242	0.1634
APOBEC3G:HLA-B*1510	2	169	0.64798	0.5244	PPIA(1650):IL-10(-1082)	4	169	0.29824	0.8788
APOBEC3G:PSIP(rs12339417)	4	169	1.23555	0.2977	PSIP(rs12339417):IL-10(-592)	4	169	1.70630	0.1508
HLA-B*1510:PPIA(1650)	2	169	0.42547	0.6542	Time:gender	1	1472	0.78414	0.3760
APOBEC3G:time	2	1472	0.06022	0.9416	Time:TNPO3	2	1472	1.30849	0.2705
APOBEC3G:gender	2	169	2.91170	0.0571	Time:IL-10(-1082)	2	1472	0.14900	0.8616
APOBEC3G:IL-10(-592)	4	169	1.48970	0.2075	Gender:TNPO3	2	169	0.66666	0.5148
APOBEC3G:HLA-B*0801	2	169	3.35556	0.0372	Gender:IL-10(-3575)	2	169	6.64257	0.0017
APOBEC3G:IL-10(-1082)	4	169	0.26266	0.9016	Gender:IL-10(-592)	2	169	5.24481	0.0062
HLA-B*0801:HLA-B*1510	1	169	0.16657	0.6837	PSIP(rs12339417):IL-10(-1082)	4	169	0.13610	0.9688
HLA-B*0801:PPIA(1650)	2	169	0.64119	0.5279	PSIP(rs12339417):IL-10(-3575)	4	169	0.55661	0.6945
HLA-B*0801:PSIP(rs12339417)	2	169	0.77955	0.4603	PSIP(rs12339417):gender	2	169	0.49640	0.6096

TABLE 4.12: Solution of fixed effects for final model with log viral load as the response

	Value	Std. Error	df	t-value	p-value
(Intercept)	14.607412	2.499374	1472	5.844429	0.0000
APOBEC3G					
CC	Ref				
CG	-2.377775	1.153450	169	-2.061446	0.0408
GG	-3.047504	2.009332	169	-1.516675	0.1312
HLA-B					
0801(absent)	Ref				
0801	3.427881	2.413906	169	1.420056	0.1574
1510(absent)	Ref				
1510	2.238049	4.483297	169	0.499197	0.6183
4201(absent)	Ref				
4201	0.615903	0.374142	169	1.646175	0.1016
5801(absent)	Ref				
5801	0.763442	0.466407	169	1.636857	0.1035
5802(absent)	Ref				
5802	0.882953	0.426961	169	2.067994	0.0402
PPIA(1650)					
AA	Ref				
AG	-0.927312	1.469911	169	-0.630863	0.5290
GG	-0.825637	1.618229	169	-0.510210	0.6106
PSIP(rs12339417)					
CC	Ref				
CT	-1.909094	2.929346	169	-0.651713	0.5155
TT	-2.968520	2.835120	169	-1.047053	0.2966
Time	-0.020109	0.005017	1472	-4.008147	0.0001
Gender					
Female	Ref				
male	-6.475880	2.397039	169	-2.701616	0.0076
TNPO3					
AA	Ref				
AG	-0.066813	0.408128	169	-0.163705	0.8702
GG	1.067490	0.986119	169	1.082517	0.2806
IL-10(-3575)					
AA	Ref				
TA	0.043369	1.775987	169	0.024420	0.9805
TT	0.436763	1.925426	169	0.226840	0.8208
IL-10(-592)					
AA	Ref				
CA	-5.521973	1.519090	169	-3.635053	0.0004
CC	-4.961327	1.656206	169	-2.995597	0.0032
IL-10(-1082)					
AA	Ref				
AG	0.358637	1.045283	169	0.343101	0.7319
GG	0.597402	1.712726	169	0.348802	0.7277
APOBEC3G(CG):HLA-B*1510					
CC:HLA-B*1510(absent)	Ref				
CG:HLA-B*1510	-0.425754	0.748532	169	-0.568785	0.5703
GG:HLA-B*1510	1.099631	1.596517	169	0.688769	0.4919
APOBEC3G:PSIP(rs12339417)					
CC:CC	Ref				
CG:CT	1.381543	0.928825	169	1.487409	0.1388
GG:CT	0.831081	1.371145	169	0.606122	0.5452
CG:TT	1.136626	0.913026	169	1.244901	0.2149
GG:TT	-0.810770	1.424627	169	-0.569110	0.5700
HLA-B*1510:PPIA(1650)					
HLA-B*1510(absent):AA	Ref				
HLA-B*1510:AG	-0.069746	0.989719	169	-0.070470	0.9439
HLA-B*1510:GG	0.919938	1.105515	169	0.832135	0.4065
APOBEC3G:time					
CC:time	Ref				
CG:time	0.001202	0.004556	1472	0.263819	0.7920
GG:time	0.002271	0.00871	1472	0.260585	0.7944
APOBEC3G:Gender					
CC:Female	Ref				
CG:male	-2.074103	0.889732	169	-2.331155	0.0209
GG:male	-0.106324	1.149689	169	-0.092480	0.9264
APOBEC3G:IL-10(-592)					
CC:AA	Ref				
CG:CA CG:CA	1.584637	0.904140	169	1.752645	0.0815
GG:CA	3.152430	1.970033	169	1.600192	0.1114
CG:CC	1.587988	0.969753	169	1.637518	0.1034
GG:CC	3.414311	2.070346	169	1.649150	0.1010

TABLE 4.13: Solution of fixed effects for final model with log viral load as the response (Continuation of Table 4.12)

	Value	Std. Error	df	t-value	p-value
APOBEC3G:HLA-B*0801					
CC:HLA-B*0801(absent)	Ref				
CG:HLA-B*0801	-2.250705	1.024827	169	-2.196180	0.0294
GG:HLA-B*0801	-3.348634	1.842304	169	-1.817634	0.0709
APOBEC3G:IL-10(-1082)					
CC:AA	Ref				
CG:AG	-0.306395	0.570990	169	-0.536604	0.5922
GG:AG	0.030661	1.030392	169	0.029757	0.9763
CG:GG	0.088675	0.953315	169	0.093017	0.9260
GG:GG	-2.586353	3.247694	169	-0.796366	0.4269
HLA-B*0801:HLA-B*1510					
HLA-B*0801(absent):HLA-B*1510(absent)	Ref				
HLA-B*0801:HLA-B*1510	1.319101	3.339574	169	0.394991	0.6933
HLA-B*0801:PPIA(1650)					
HLA-B*0801(absent):AA	Ref				
HLA-B*0801:AG	1.267345	1.289980	169	0.982454	0.3273
HLA-B*0801:GG	0.298753	1.680725	169	0.177752	0.8591
HLA-B*0801:PSIP(rs12339417)					
HLA-B*0801(absent):CC	Ref				
HLA-B*0801:CT	0.700135	1.309654	169	0.534595	0.5936
HLA-B*0801:TT	-0.780332	1.600713	169	-0.487490	0.6265
HLA-B*0801:time					
HLA-B*0801(absent):time	Ref				
HLA-B*0801:time	-0.001029	0.008021	1472	-0.128283	0.8979
HLA-B*0801:Gender					
HLA-B*0801(absent):Female	Ref				
HLA-B*0801:male	-0.587673	1.856315	169	-0.316581	0.7520
HLA-B*0801:TNPO3(AG)					
HLA-B*0801(absent):AA	Ref				
HLA-B*0801:AG	0.080113	1.349314	169	0.059373	0.9527
HLA-B*0801:GG	3.750799	2.328659	169	1.610712	0.1091
HLA-B*0801:IL-10(-592)					
HLA-B*0801(absent):AA	Ref				
HLA-B*0801:CA	-3.566143	2.256565	169	-1.580342	0.1159
HLA-B*0801:CC	-2.977862	2.283428	169	-1.304119	0.1940
HLA-B*1510:PSIP(rs12339417)					
HLA-B*1510:CC					
HLA-B*1510:CT	-0.190833	1.675435	169	-0.113901	0.9095
HLA-B*1510:TT	-1.517498	1.638582	169	-0.926105	0.3557
HLA-B*1510:time					
HLA-B*1510(absent):time	Ref				
HLA-B*1510:time	0.000403	0.006490	1472	0.062032	0.9505
HLA-B*1510:Gender					
HLA-B*1510(absent):Female	Ref				
HLA-B*1510:male	-0.488307	1.488380	169	-0.328079	0.7433
HLA-B*1510:IL-10(-3575)					
HLA-B*1510(absent):AA	Ref				
HLA-B*1510:TA	-1.168814	3.435210	169	-0.340245	0.7341
HLA-B*1510:TT	-0.737063	3.618228	169	-0.203708	0.8388
HLA-B*1510:IL-10(-592)					
HLA-B*1510(absent):AA	Ref				
HLA-B*1510:CA	0.187646	2.020350	169	0.092878	0.9261
HLA-B*1510:CC	1.116688	2.193627	169	0.509060	0.6114
HLA-B*1510:IL-10(-1082)					
HLA-B*1510(absent):AA	Ref				
HLA-B*1510:AG	-0.271911	1.105165	169	-0.246037	0.8060
HLA-B*1510:GG	1.146630	2.195806	169	0.522191	0.6022
HLA-B*4201:time					
HLA-B*4201(absent):time	Ref				
HLA-B*4201:time	0.004725	0.00539	1472	0.876551	0.3809
HLA-B*5802(absent):time	Ref				
HLA-B*5802:time	-0.012001	0.007651	1472	-1.568605	0.1170
PPIA(1650):PSIP(rs12339417)					
AA:CC	Ref				
AG:CT	-0.813426	0.986127	169	-0.824869	0.4106
GG:CT	-0.500984	1.311488	169	-0.381997	0.7029
AG:TT	1.054753	0.972175	169	1.084941	0.2795
GG:TT	0.153008	1.316095	169	0.116259	0.9076

TABLE 4.14: Solution of fixed effects for final model with log viral load as the response (Continuation of Table 4.12)

	Value	Std. Error	df	t-value	p-value
PPIA(1650):time					
AA:time	Ref				
AG:time	-0.001522	0.005063	1472	-0.300686	0.7637
GG:time	-0.008759	0.006026	1472	-1.453587	0.1463
PPIA(1650):IL-10(-592)					
AA:AA	Ref				
AG:CA	1.476154	1.341192	169	1.100628	0.2726
GG:CA	2.542516	1.164745	169	2.182895	0.0304
AG:CC	1.529898	1.411843	169	1.083618	0.2801
GG:CC	1.660356	1.192178	169	1.392708	0.1655
PPIA(1650):IL-10(-1082)					
AA:AA	Ref				
AG:AG	-0.214747	0.617114	169	-0.347986	0.7283
GG:AG	-0.168968	0.701563	169	-0.240846	0.8100
AG:GG	-0.888465	0.990876	169	-0.896646	0.3712
GG:GG	0.135077	1.685572	169	0.080137	0.9362
PSIP(rs12339417):IL-10(-592)					
CC:AA	Ref				
CT:CA	3.529642	1.502736	169	2.348811	0.0200
TT:CA	3.057450	1.589961	169	1.922971	0.0562
CT:CC	2.429424	1.662017	169	1.461732	0.1457
TT:CC	2.111525	1.728410	169	1.221658	0.2235
time:Gender					
time:Female	Ref				
time:male	0.004872	0.005685	1472	0.857007	0.3916
time:TNPO3					
time:AA	Ref				
time:AG	0.008458	0.005403	1472	1.565451	0.1177
time:GG	0.001314	0.011765	1472	0.111683	0.9111
time:IL-10(-1082)					
time:AA	Ref				
time:AG	-0.001787	0.004946	1472	-0.361189	0.7180
time:GG	0.001521	0.006283	1472	0.242117	0.8087
Gender:TNPO					
Female:AA	Ref				
male:AG	0.532181	0.748989	169	0.710532	0.4784
male:GG	1.775859	1.970439	169	0.901250	0.3687
Gender:IL-10(-3575)					
Female:AA	Ref				
male:TA	4.361901	1.847788	169	2.360607	0.0194
male:TT	6.376486	1.909731	169	3.338944	0.0010
Gender:IL-10(-592)					
male:AA	Ref				
male:CA	1.642347	1.118443	169	1.468422	0.1438
male:CC	3.528345	1.281586	169	2.753109	0.0065
PSIP(rs12339417):IL-10(-1082)					
CC:AA	Ref				
CT:AG	-0.296982	1.117880	169	-0.265665	0.7908
TT:AG	-0.007064	1.113044	169	-0.006346	0.9949
CT:GG	-0.587812	2.058497	169	-0.285554	0.7756
TT:(-1082)GG	-0.811700	1.859279	169	-0.436567	0.6630
PSIP(rs12339417):IL-10(-3575)					
CC:AA	Ref				
CT:TA	-0.830969	2.343341	169	-0.354609	0.7233
TT:TA	0.458199	2.080889	169	0.220194	0.8260
CT:TT	-1.697710	2.489042	169	-0.682073	0.4961
TT:TT	-0.859177	2.276841	169	-0.377355	0.7064
PSIP(rs12339417):Gender					
CT:female	Ref				
CT:male	-0.427892	1.133608	169	-0.377460	0.7063
TT:male	0.227230	1.138007	169	0.199673	0.8420

Chapter 5

Dealing with Missing Data using Log Viral Load as the Response

5.1 Introduction

Drop-out is one of the major challenges encountered in statistical studies mostly with longitudinal data, regardless of how well they are designed or executed. Our aim in this chapter is to investigate the influence of drop-out on the evolution of the HIV Bio-marker, viral load, in a model including selected genetic factors as covariates.

Data are said to be missing if some variables are having no measurement. The missing pattern is termed monotone if the event that a variable Y_j is missing for an individual implies that all subsequent variables $Y_k, k > j$, are all missing for the individual (Yuan, 2010), and whenever the missing pattern is not monotone then it's arbitrary.

Drop-outs do lead to serious problems because, most statistical procedures automatically remove cases with missing values leading to loss of information and hence making the subsequent statistical analysis to be less reliable if the missingness depends on the outcome of interest, observed or unobserved. Such strategies in standard statistical practice are problematic because they may cause misleading results by introducing bias in the analysis. Therefore there is need to deal with this problem.

Missing data in general can be as a result of non-compliance to planned visits by some individuals who leave the study and thereafter lost to follow-up, sickness, equipment failure, poor data entry, weather conditions and death among other reasons.

5.2 Drop-out Mechanisms

Different methods have been suggested to deal with drop-out. In this study we review some of these but a compressive review of these methods is found in Schafer (2003). Suppose that N individuals are to be observed at n occasions. Then for the i^{th} individual $(i = 1, 2, \ldots, N)$ we can form a $(n \times 1)$ vector $Y_i = (Y_{i1}, \ldots, Y_{in})'$, where Y_{ij} is the j^{th} outcome for individual i, which is continuous or discrete depending on the study problem. However, because of missing values missing intermittently or due to drop out the actual data is such that an individual i ends up with $n_i \leq n$ observations instead of exactly n because not all observations were collected as intended. Each individual has a $(n_i \times p)$ covariate matrix X_i. The covariates may be both time stationary and time varying. Now, suppose R_i is a $(n \times 1)$ random vector for the i^{th} individual, whose j^{th} component R_{ij} equals 1 if Y_{ij} is fully observed and 0 if not observed. The purpose of the random vector R_i is to aid in modelling the missingness process. Therefore, the full data information for the i^{th} individual are given jointly by Y_i and R_i, with a joint distribution that can be expressed as:

$$f_{y,r}(y_i, r_i|X_i, \theta, \gamma) = f_r(r_i|y_i, X_i, \gamma) f_y(y_i|X_i, \theta), \tag{5.1}$$

where θ and γ are used to parameterize the joint distribution. The "missing data mechanism", $f_r(r_i|y_i, X_i, \gamma)$ is parametrized by γ while the repeated measurement process is parametrized by θ. In general, the missing data mechanism can depend on the full vector of responses, Y_i (including possibly unobserved component of Y_i) and the matrix of covariates X_i. We denote the observed and missing components of Y_i by Y_i^o and Y_i^m respectively. According to Rubin (1976, 1987) drop-out mechanisms can be classified into three basic categories:

(1) Missing completely at random (MCAR) which implies that, the missing process does not depend on Y_i expressed as,

$$P(R_i|Y_i^o, Y_i^m, X_i, \gamma) = P(R_i|X_i, \gamma)$$

(2) Missing at random (MAR) implying that, the missingness process depends on the unobserved response expressed as,

$$P(R_i|Y_i^o, Y_i^m, X_i, \gamma) = P(R_i|Y_i^o, X_i, \gamma)$$

and

(3) Missing not at random (MNAR) which allows the missing process to depend on the unobserved responses in probability terms expressed as,

$$P(R_i|Y_i^o, Y_i^m, X_i, \gamma) = P(R_i|Y_i^m, X_i, \gamma).$$

In the context of likelihood based analysis, an MCAR mechanism is a special case of MAR, these two mechanisms are referred to as "ignorable". A mechanism is referred to as "ignorable" if, the parameters that govern the missing data process are unrelated to the parameters to be estimated. This implies θ and γ are orthogonal. In contrast, an MNAR mechanism is often referred to as "non ignorable" because the missingness mechanism is not independent of the measurement process. That is the orthogonality condition above fails. Clearly it is easier to handle ignorability compared to non-ignorability. Our focus in this chapter is to deal with missing data as a result of subject drop-outs assuming the missing mechanism is MAR. Given the values of the vector R_i take the value 0 or 1 depending on whether the outcome is missing or observed then drop-out variable for the i^{th} individual can be defined as follows:

$$D_i = 1 + \sum_{t=1}^{T} R_{it}, \tag{5.2}$$

hence, the model for missing data or drop-out process can be rewritten as:

$$f(r_i|y_i, X_i, \gamma) = P(D_i = d_i|y_i, X_i, \gamma), \tag{5.3}$$

where d_i is a realization of the variable D_i. From Equation 5.2 we assume that, all subjects are observed on the first instance so that D_i takes values between 2 and $(n+1)$. The maximum value $(n+1)$ corresponds to a complete measurement sequence. Using Equation 5.2, in the case of drop-out missing completely at random (MCAR) the model reduces to $P(D_i = d_i|Y_i, X_i, \gamma) = P(D_i = d_i|X_i, \gamma)$ while for drop-out missing at random (MAR) the model is given by $P(D_i = d_i|Y_i, X_i, \gamma) = P(D_i = d_i|Y_i^o, X_i, \gamma)$ (Ali and Talukder, 2005).

5.3 Methods of Handling Missing Data

In this section, we discuss some of the approaches developed in the literature to deal with missing data. These includes: deletion methods; imputation methods; the missing

data models; weighting methods; and the likelihood based approaches.

5.3.1 Deletion Methods

Deletion methods are techniques which reduces the available data to a dataset with no missing values. The two main deletion methods are: complete case deletion; and available case analysis.

Complete Case Deletion

In complete case deletion method, only subjects (individuals) with complete data on all the variables are analysed while those with incomplete data are dropped from further analysis.

Advantages of Complete Case Deletion Method

(1) Can be applied for any kind of statistical analysis, from structural equation modelling to log linear analysis.

(2) Do not require any special computation methods.

Disadvantage of Complete Case Deletion Method

(1) It reduces the sample size, which reduces the precision of estimates and therefore can result into biased results.

Available Case Analysis

Available case analysis methods uses the available information to estimate means and covariances by incorporating vectors of repeated measures of unequal length in the analysis. One of the most common methods of available case analysis is the pair-wise deletion method. In pair-wise deletion method, a covariance (or correlation) matrix is computed where each element is based on the full number of cases with complete data for each pair of variables. For example, to compute the covariance structure between two variables X and Y, we only use the cases with data in both X and Y.

Advantages of Available Case Analysis Method

(1) It's more efficient than complete case analysis because it incorporates partial information obtained from those who are missing.

(2) If the sample size is large, it results into parameter estimates that are consistent and approximately unbiased.

Disadvantages of Available Case Analysis Method

(1) The sample base changes from variable to variable according to the pattern of missing. Which posses challenges when it comes to determining the sample size, degrees of freedom and combining different pieces of information.

(2) This method leads to biased estimates of treatment comparisons if missing data is not MCAR.

(Nakai and Ke, 2011).

5.3.2 Imputation Methods

Imputation, is the practice of "filling in" missing data with plausible values. There are two types of imputation (1) Single imputation and (2) Multiple imputation.

Single Imputation

Single imputation involves replacing a missing value with a single imputed value based on an estimate of the true value of the unobserved variable (Nakai and Ke, 2011). It is commonly used because it's easy to analyze once imputation has been done. The only challenge with this method is, imputing a single value considers that value as known, therefore without special adjustments single imputation fails to show sampling variability under one model for non-response or uncertainty about the correct model for non-response (Nakai and Ke, 2011). The three main common simple imputation approaches are: the last observation carried forward; mean imputation; and hot-deck imputation.

The last observation carried forward (LOCF) is one of the most common methods used in longitudinal analysis. In this method, every missing observation is replaced with the last observed value from the same subject assuming that the response value remains unchanged after missing. This is quite unrealistic and only applies when missing data is due to recovery or cure (Nakai and Ke, 2011).

The other approach is the mean imputation method. In this method, missing values are filled with the mean of the observed values. This method assumes that, the mean of the variable is the best estimate for any observation that has missing value on the variable. For unbiased estimates, however, the missing values needs to be MCAR. This method seems easy but can really alter the variable distribution thereafter interfering with the summary measures. It should be noted that, the distribution of the imputed values using this method is different from the observed values (Nakai and Ke, 2011).

Finally, the hot-deck imputation is another common single imputation approach. This method replaces missing values with values from similar responding units in the sample. The imputed values do not alter the distribution of the sampled values. The only challenge with this method is, it interferes with correlations and covariances (Nakai and Ke, 2011).

Multiple Imputation (MI)

This is one of the best methods of dealing with missing values on data sets and has become popular among social science researchers due to its simplicity and generality (Peng and Zhu, 2008). The technique involves imputing missing values multiple times and replacing each missing value with two or more acceptable values (Rubin, 1987). The MI technique is preferred over other methods because it is more powerful under MAR and the fact that it accounts for between imputation uncertainty. Some of the limitations of this method includes: (1) The missing value individuals are allowed to have varying probability. (2) MI requires a lot of time for the creation of imputations and analysis of the results (Nakai and Ke, 2011).

Under the MAR setting, MI requires the analyst to specify which variables are to be used as regressors in the imputation model. The MI inference assumes that the analysis model is the same as the model used to impute missing values (the imputation model). Practically, the two models might not be the same (Meng, 1994; Schafer, 1997). The quality of the imputation model influences the quality of the results from the analysis model. Therefore, it is advisable to consider the design of the imputation model.

Rubin's (1987) imputation procedure works in a way that, missing values are filled multiple times in order to construct multiple complete data sets in three steps namely: imputation; analysis and pooling. In the imputation stage, each of the missing values is replaced by $M \geq 2$ simulated values. Each of these sets of plausible values can be used to fill in the missing values and create a complete data set assuming the missing data is ignorable. In the analysis stage, each of the M complete data sets is then analysed using standard methods, such as linear mixed models (LMM), generalized linear mixed model (GLMM), among other methods. The number of imputations (M) need not to be very

large since, in practice, $3-10$ imputations often provide satisfactory results (Schafer and Olsen, 1998). In the third stage, analyses from M complete data sets are combined for inference using methods that allow for uncertainty in imputation to be taken into account (Yuan, 2010).

The MI technique uses Y_i^o to fill in Y_i^m leading to complete data $Y_i = (Y_i^o, Y_i^m)$. Suppose we know the distribution of Y_i^m, with parameter vector γ, then we could impute Y_i^m by drawing from the conditional distribution $f(Y_i^m|Y_i^o, \gamma)$. However, since γ is unknown, we estimate it from the observed data yielding $\hat{\gamma}$ and use the distribution $f(Y_i^m|Y_i^o, \hat{\gamma})$. Since $\hat{\gamma}$ is a random variable, we must take its variability into account in drawing imputations. In Bayesian terms, γ is a random variable where the distribution depends on the data. So, we first obtain the posterior distribution of γ from the data, a distribution which is a function of $\hat{\gamma}$. After formulating the posterior distribution of γ, the following imputation algorithm can be used: (1) Draw γ^* from the posterior distribution of γ, $f(\gamma \mid X_i, Y_i^o)$. We approximate this posterior distribution by the normal distribution. (2) Draw Y_i^m from $f(Y_i^m \mid X_i, Y_i^o, \gamma^*)$. (3) Use the complete data Y_i and the model to estimate the parameter of interest (β^*) and its variance ($\Sigma(\beta^*)$) called the within-imputation variance. The three stages described earlier are repeated independently M times, resulting in β_k^*, $\Sigma(\beta_k^*)$, $k = 1, \ldots, M$. Stage 1 and 2 are referred to as the imputation task and stage 3 is the estimation task. Finally, we combine the estimates obtained after M imputations. The overall estimated parameter vector is the average of all individual estimates:

$$\beta^* = \frac{1}{M} \sum_{k=1}^{M} \beta_k^*.$$

We obtain the variance as a weighted sum of the within-imputation variance and the between-imputation variability:

$$\Sigma^* = W + \left(\frac{M+1}{M}\right) B, \tag{5.4}$$

where

$$W = \frac{1}{M} \sum_{k=1}^{M} \Sigma(\beta_k^*) \tag{5.5}$$

is the average of the within-imputation variances and

$$B = \frac{1}{M-1}\sum_{k=1}^{M}(\beta_k^* - \beta^*)(\beta_k^* - \beta^*)' \tag{5.6}$$

is the between-imputations variance (Little and Rubin, 1987). According to Verbeke and Molenberghs (2000), γ is an easily estimated set of parameters characterizing the distribution of Y_i. In contrast, β is difficult to estimate in the presence of non-response and obtaining a correct estimate for the variability is non-trivial. The overall standard errors are the square roots of the diagonal elements of Σ^*. Confidence intervals are obtained by taking the overall estimate plus or minus a number of standard errors, where that number is a quantile of Student's t-distribution with degrees of freedom

$$df = (M-1)\left(1 + \frac{MW}{(M+1)\beta}\right)^2. \tag{5.7}$$

A significance test of the null hypothesis β_i^*=0 is performed by comparing the ratio $t = \beta^*/\sqrt{\Sigma_{i,i}^*}$ to the same t-distribution. More methods for pooling the results from multiply imputed data can be found in (Schafer, 1997).

5.3.3 Types of Missing Data Analysis Models

Some of the examples of missing data analysis models includes: (1) Generalized estimation equations (GEE), (2) Weighted generalized estimation equations (WGEE), (3) Generalized linear mixed models (4) MNAR models among others.

(i) Generalized Estimation Equations (GEE)

Let $P_{it}(\beta) = E(Y_i|X_i, \beta)$ be the vector of probabilities of responses for the complete data case, where P_{1i} is the marginal probability of response for the i^{th} subject (individual) at time point 1 and β is the vector of regression parameters. Partitioning P_{1i} into P_{1i}^o and P_{1i}^m to denote the observed and the missing part, respectively. Then, the GEEs is expressed as:

$$U_\beta(\hat{\beta}) = \sum_{i=1}^{N} \hat{D_{1i}}' \hat{V_{1i}^{-1}}\{Y_i^o - P_{1i}^o(\hat{\beta})\} = 0$$

as proposed by Liang and Zeger (1986) and Prentice (1988), where $D_{1i} = \partial P_{1i}^o(\beta)/\partial\beta$ and V_{1i} is a "working" covariance matrix. The GEE estimate of β and its covariance matrix Y_i^o can be obtained using the PROC GENMOD procedure in SAS.

We should note that, in-case there is missingness of data, the GEE method only provides consistent estimates of the model parameters under the MCAR assumption. Which is very restrictive in many longitudinal studies for example in clinical trials (Ali and Talukder, 2005).

(ii) Weighted Generalized Estimation Equations (WGEE)
Fitzmaurice and Laird (2000) came-up with a formulation of the WGEE method based on the Robins et al. (1995). The estimation equation under this improvement is given by:

$$U_\beta(\hat{\beta}) = \sum_{i=1}^{N} \frac{1}{v_{id}} \hat{D_{1i}}' \hat{V_{1i}^{-1}} \{Y_i^o - P_{1i}^o(\hat{\beta})\} = 0, \tag{5.8}$$

as a result of modifying

$$U_\beta(\hat{\beta}) = \sum_{i=1}^{N} \hat{D_{1i}}' \hat{V_{1i}^{-1}} \{Y_i^o - P_{1i}^o(\hat{\beta})\} = 0$$

where, v_{id} is the probability of drop-out of the i^{th} subject at time d i.e. $v_{id} = pr(D_i = d|Y_i, X_i, \gamma)$. Provided v_{id} are correctly specified, the WGEE method provides consistent estimates of the model parameters under the MAR mechanism (Robins et al., 1995). The parameter estimates for Equation 5.8 can also be determined using PROC GENMOD procedure in SAS.

(iii) Generalized Linear Mixed Models (GLMM)
This method was proposed by Fitzmaurice and Laird (2000) as a generalization of the conditional linear model developed by Wu and Bailey (1989) for dealing with informative (MNAR) drop-outs. The difference between the model developed by Wu and Bailey (1989) to that proposed by Fitzmaurice and Laird (2000) is, the Wu and Bailey model, only handles continuous response variables while the model proposed by Fitzmaurice and Laird (2000) can be used for both continuous and discrete cases. Fitzmaurice and Laird (2000) model for Y_i conditional on the drop-out time D_i is expressed as:

$$g(E[Y_{it}|D_i, X_{it}]) = Z_{it}'\beta + U_i,$$

where $g(.)$ is a known link function, X_{it} the covariates, Z_{it} is the design vector which depends on the drop-out time and U_i is a subject specific random effect. If the data is binary then, $g()$ will be the logit transformation $\log\left(\frac{P_{it}}{1-P_{it}}\right)$, $P_{it} = pr(Y_{it} = 1|X_i, \beta) = E(Y_{it}|X_i), \beta$. In the GLMM models the aim of statistical analysis is to estimate the marginal mean of repeated outcomes and the

comparisons between treatment groups. This marginal means are averaged over the distribution of drop-out times as shown:

$$E(Y_{it}|X_{it}) = \mu_{it} = \sum_{l=2}^{T+1} \pi_l g^{-1}(Z'_{it}\beta) \tag{5.9}$$

where, Z_{it} depends on the drop-out patterns and X_{it}, π_l depends on some subset X_i. The GEE formulation for estimating β in Equation 5.9, so that the model parameters could be estimated using PROC GENMOD of SAS was described by Fitzmaurice and Laird (2000). The multinormial probabilities, π_l are estimated by the sample proportional at each drop-out time while the marginal mean at occasion t is estimated by:

$$\hat{\mu_{it}} = \sum_{i=2}^{T+1} \hat{\pi}_l g^{-1}(Z'_{it}\hat{\beta}). \tag{5.10}$$

The standard errors of $\hat{\mu_{it}}$ and the difference $\hat{\mu_{it}} - \hat{\mu_{jt}}$ can be obtained using the δ method. Note, for any function $h(\beta, \pi)$, the asymptotic covariance matrix of $N^{\frac{1}{2}}[h(\hat{\beta}, \hat{\pi}) - h(\beta, \pi)]$ can be approximated by $\Phi' V \Phi'$ where Φ is the Jacobian evaluated at β, π and

$$V = \begin{pmatrix} V_{\hat{\beta}} & 0 \\ 0 & V_{\hat{\pi}} \end{pmatrix},$$

where $V_{\hat{\beta}}$ denotes the covariance matrix of $\hat{\beta}$, an estimate that can be obtained from the GEE algorithm and $V_{\hat{\pi}}$ denotes the covariance matrix of $\hat{\pi}$ which can be estimated by $\frac{1}{n}[diag(\hat{\pi}) - \hat{\pi}\hat{\pi}']$ (Ali and Talukder, 2005).

(iv) MNAR Models

Data is referred to as non-ignorable if it is MNAR. According to Rubin, the propensity for missing data is a random variable that has a distribution. In practical terms this implies that, each variable potentially yields a pair of scores: an underlying Y value that may or may not be observed and a corresponding R value that denotes whether Y is observed or is missing i.e. R=0 if Y is observed and R=1 if Y is missing. Under an MNAR mechanism, the data and the probability of missingness have a joint distribution:

$$f(Y, R|\theta, \alpha),$$

where f denotes a probability distribution, Y is the outcome variable, R is the corresponding missing data indicator, θ is a set of parameters that describes the

distribution of Y and α contains parameters that describe the propensity for missing data on Y. Collectively, the parameters of the joint distribution dictate the mutual occurrence of different Y values and missing data (Enders, 2011).

We should note that, an MNAR mechanism requires an analysis model that includes all parameters of the joint distribution, not just those that are of substantive interest. Selection and pattern mixture models both incorporate a model for R into the analysis, but they do so in different ways (Enders, 2011).

(i) Selection Model

Selection model specifies the model for both longitudinal and missing process simultaneously (Fitzmaurice, 2003). Both models depends on random subject effects, most or all of which are shared by both models. In the model, the complete data model is used for the longitudinal response and the probability of missingness is modelled conditional on the possibly unobserved response. Let R denote the missingness process variable and Y the longitudinal outcome variable. The joint distribution of R and Y is given by:

$$f(R, Y|\theta, \alpha) = f(Y|\theta)f(R|Y, \alpha), \tag{5.11}$$

where $f(Y|\theta)$ is the distribution of Y in the population, $f(R|Y, \alpha)$ models the incidence of missing value as a function of Y, θ and α are unknown vector parameters and both are distinct (Little and Rubin, 1987).

For identification, the set of outcomes is usually restricted in some way. With this model, identification comes from unverifiable models for the dependence of the drop-out probabilities on the unobserved response. It is usually difficult to be able to identify the restrictions that needs to be inserted in the model. Even though the selection model is easy to formulate, the hypotheses about drop-out process is computationally intractable because it is difficult to infer how assumptions on drop-out process translate into assumptions about distribution of unobserved response (Nakai and Ke, 2011).

(ii) Pattern Mixture Model

The pattern mixture model uses missing value pattern as between subject variable in the longitudinal model. The main idea behind this model is to explicitly model the missing data distribution by first identifying different patterns of missing data and then including parameters in the response model that capture this effect (Hedeker and Gibbons, 1997) and (Nakai and Ke, 2011). The model is expressed as:

$$f(R, Y|\xi, \varphi) = f(Y|R, \xi)f(R|\varphi), \tag{5.12}$$

where $f(Y|R, \xi)$ denotes the distribution of Y in the strata defined by different patterns of missing data R and $f(R|\varphi)$ models the incidence of the different patterns. ξ and φ are unknown distinct vector parameters. This method is simple however, it makes an implicit assumptions about distribution of unobserved response since the missingness process is not immediately transparent (Nakai and Ke, 2011).

5.3.4 Weighting Methods

The basic idea behind this method is, constructing weights for complete cases in order to reduce or remove bias. Little and Schenker (1995) illustrates the basic concept of weighting adjustments in the sample survey setting. Heyting et al. (1992) described the method as follows, consider a situation where each patient belongs to a subgroup in the patient population in which all patients have a similar baseline and response profile. A proportion within each subgroup is destined to complete the clinical trial while the remainder are designed to drop early. Those "completer" patients with a very low probability of completing can certainly have an overly strong influence on the results. Heyting et al. (1992) came up with a particular weighing method where the evaluation of the mean treatment differences at the end of the study is not only of primary interest but also easy to explain in nature. Robins et al. (1995) introduced a weighting method which allows generalized estimation equation (GEE) analyses to be correct under the MAR assumption. This method can be done in softwares that allows for weights such as Stata, SUDAAN and SAS. This method is more appropriate with single missing predictor and becomes difficult to deal with, with multiple missing variables if they are not monotone (Horton and Kleinman, 2007).

5.3.5 Likelihood Based Approaches

Maximum likelihood method assumes that the missingness is MAR. It is one of the most recommended techniques because it produces unbiased estimates with both MCAR and MAR data. The only challenge with this method is, it does lead to biased estimates when the data is MNAR. The aim of this method is to estimate the parameter of the joint distribution of the data such as the mean vector and covariance matrix. The joint distribution is maximized using the Expectation Maximization (EM) algorithm. The EM algorithm due to Dempster et al. (1977) made it possible to workout ML estimates

in many missing data problems instead of deleting or filling in incomplete cases. The ML estimates, considers missing data as random variables to be removed from the likelihood function as if they were never sampled (Schafer and Graham, 2002). The EM algorithm is an iterative algorithm that finds the parameters which maximize the log likelihood when there are missing values. This method capitalizes on the relationship between missing data and the unknown parameters of a data model (Nakai and Ke, 2011; Nelwamondo et al., 2007). The only disadvantage of this method is that the rate of convergence can be very slow, if the missing values are many. Each iteration of EM consists of an E step (expectation step) and the M step (maximization step). Given a set of parameter estimates, the E-step calculates the conditional expectation of the complete data log likelihood given the observed data and the parameter estimates. Let θ^t be the estimate of θ. Then,

$$H(\theta|\theta^t) = \int g(\theta|Y) f(Y^R|Y^o, \theta = \theta^t) dY^R, \tag{5.13}$$

where $g(\theta|Y)$ denotes complete data log likelihood. Given a complete data log likelihood, the M step finds the parameter estimates to maximize the complete data log likelihood from E step,

$$H(\theta^{(t+1)}|\theta^t) \geq H(\theta|\theta^t) \quad \forall\theta. \tag{5.14}$$

These two steps (M and E) only stops when the iteration converges (Nakai and Ke, 2011).

We should not that, EM algorithm is not the only method that can be used to maximize likelihood for incomplete data. Methods such as Newton-Raphson and Fisher scoring can also be used.

5.4 Applications of Multiple Imputation Method to the Sinikithemba Data

5.4.1 Model Formulation

Having described some of the approaches used to deal with missing data, we now employ one of the methods in the modelling of viral load. We consider the following linear model for the log viral load for subject i at time j as the response. See model 5.15 below.

$$log(VL_{ij}) = \beta_0 + \beta_1 APOBEC3G + \beta_2 PSIP_{rs2339417} + \beta_3 TNP03 +$$
$$\beta_4 PPIA_{1650} + \beta_5 IL-10(-592) + \beta_6 IL-10(-1082) + \beta_7 IL-10(-3575) +$$
$$\beta_8 Gender + \beta_9 HLA-B*0801 + \beta_{10} HLA-B*5801 + \beta_{11} HLA-B*1503 +$$
$$\beta_{12} HLA-B*5802 + \beta_{13} HLA-B*1510 + \beta_{14} HLA-B*4201 + \epsilon_{ij}, \quad (5.15)$$

where ϵ_{ij} is an unknown independent distributed normal random error, with mean 0 and variance σ^2. This is the same best model arrived in chapter 4 but we ignore interaction terms since our aim here is to show the potential of MI in handling missing data.

We carried out an application study to investigate the potential influence that drop-outs might have on the HIV Bio-marker data. This study (Sinikithemba study) consists of missing observations in the response variable and in some of the covariates. All the individuals have observed values for all the covariates at time point 1 except for (APOBEC3G, $PSIP_{rs12339417}$, TNPO3 and $PPIA_{1650}$). The missing rates in the covariates were: 10.20 %, 11.31%, 28.38% and 12.20% for APOBEC3G, $PSIP_{rs12339417}$, TNPO3 and $PPIA_{1650}$ respectively; while in the response variables the missing rates were: 43.06% and 60.35% for VL and CD4 count respectively.

Due to these missing observations, the design matrix of covariates X can be split into two parts, $X = (X^o, X^m)$; where X^o represents the part of the design matrix X with covariates that are completely observed and X^m is the subset of X with explanatory variables which have at least one value that is not observed. In this MI application we distinguished between two scenarios: (1) The missing outcome and (2) The missing outcome and covariates. In the first instance, we fitted model 5.15 while accounting for missing values in the outcome ignoring all the missing values in the covariates analysis. Thereafter, we fitted the same model (5.15) and analysed it taking into account for both missing outcome (Y^m) and missing covariates (X^m) using the MI method.

5.4.2 Inference under MI

MI was conducted using SAS PROC MI to fill in all the missing observations in the above dataset. The (MI) technique uses Markov Chain Monte Carlo (MCMC) sampling to draw imputations. We used a burn in of period of 200 iterations, 100 iterations between each step and five imputations. The defaults of the SAS MI algorithm are used for the MCMC computations, namely a Jeffreys prior and initial sampler values from the EM posterior mode was also used. The linear mixed model was then fitted to each imputed dataset, and thereafter, the results combined for final inference. We used PROC MIXED to set up effect parameterizations for the class variables as well as

ODS STATEMENT to create an output datasets that matches PROC MIANALYZE for combining the estimates from the 5 completed datasets. Thereafter, each dataset that was produced consisted of 6396 outcomes $\log(VL)$ with complete data at all the fourteen time points. Then we used PROC MIANALYZE to combine the estimates from PROC MIXED into a single inference.

5.4.3 Imputation model

The imputation model was fitted using all individuals included in the data. The imputation model is based on model 5.15 which assumes multivariate normality of the variables. In the imputation model, the variables that are included in Y_i^o should be those that causes VL to be missing at random. To make the MAR assumption more plausible as well as improving the accuracy and efficiency of the imputation, we used all the available data including the outcome variable (VL) as recommended by Van Buuren et al. (1999) who suggested the following variables to be included in the imputation model: variables in the analysis model, variables associated with missingness of the imputed variable, and variables correlated with the imputed variable.

5.4.4 Number of Imputations (M)

After the imputation model had been specified, PROC MI was then applied to generate M=5 complete datasets. Note that, the choice of M=5 was considered adequate and the efficiency of the parameter estimate based on imputation given by:

$$(1 + \frac{\xi}{M})^{-1}), \tag{5.16}$$

where ξ is the rate of missing data (Rubin, 1987). Formula 5.16 shows that, the relative efficiency of the MI inference is related to the missingness rate (ξ) in combination with the number of imputations (M). The estimated missing data rate (ξ) is $\xi = \frac{z+2/(df+3)}{z+1}$, where $z = \frac{(1+M^{-1})\beta}{W}$ is the relative increase in variance due to non-response. Rubin (1987) demonstrated using simulations that the number of imputations can be restricted to less than 10. Many statistical practices tend to support Rubin's heuristics of (3-10) imputations. However, Schafer and Olsen (1998); Peng et al. (2007) recommended the use of M=5 before the results are combined. By, this rationale for the missing outcome as well as for both missing outcome and covariates models, we achieve at least 93% and 95% efficiency, respectively.

5.5 Results

Table 5.1 shows the results for our analysis.

5.5.1 Results for Handling Missing Outcome

TABLE 5.1: Parameter estimates, standard errors and p-values of the covariates before and after handling drop-outs (interaction terms are not shown)

	Incomplete data			MI		
effect	parameters	errors	p-value	parameters	errors	p-value
intercept	3.7853	0.3745	< 0.0001	3.3157	0.3518	< 0.0001
time	-0.0841	0.0070	< 0.0001	-0.0974	0.0059	< 0.0001
APOBEC3G	-0.0563	0.0817	0.4907	-0.0237	0.0759	0.7549
PSIP(rs2339417)	0.0993	0.0782	0.2041	0.0585	0.0726	0.4200
TNP03	0.1647	0.1036	0.1121	0.1875	0.0963	0.0518
PPIA(1650)	0.1353	0.0707	0.0562	0.1736	0.0660	0.0086
IL-10(-592)	0.1480	0.0900	0.1004	0.2595	0.0830	0.0018
IL-10(-1082)	0.0604	0.1210	0.6175	0.0467	0.1136	0.6808
IL-10(-3575)	0.0185	0.1384	0.8934	0.2050	0.1271	0.1070
Gender	-0.2762	0.1314	0.0357	-0.1149	0.1197	0.0372
HLA-B*0801	0.0336	0.1911	0.8605	0.0242	0.1812	0.8935
HLA-B*5801	0.0769	0.1678	0.6466	0.0374	0.1577	0.8123
HLA-B*1503	-0.0085	0.1383	0.9506	-0.2549	0.1305	0.0509
HLA-B*5802	0.2168	0.1426	0.1286	0.2037	0.1270	0.1090
HLA-B*1510	0.3576	0.1496	0.0169	0.4362	0.1382	0.0016
HLA-B*4201	0.1917	0.1344	0.1538	0.2739	0.1209	0.0235

Table 5.1 shows the results of parameter estimates, standard errors and p values of the genetic factors on HIV Bio-marker data before and after handling drop-outs in the outcome variable ($\log(VL)$). Comparing the results obtained from the MI method with those obtained from the incomplete datasets, we see that, the parameter estimates produced from the MI method were not much different from those obtained from the original dataset as observed in IL-10(-592), IL-10(-3575), HLA-B*1503 and HLA-B*4201. However, the standard errors from the original dataset were generally larger compared to those from the MI method. The results obtained from the original data analysis were associated with the inflated standard errors, compared to the results obtained from the MI method. Inflation of the standard errors as a result of incomplete observations is caused by the missing observations that tend to create inflated artificial values than expected. Inflation of the standard errors as a result of incomplete observations is well explained in (De et al., 2003). The results in general showed that there are systematic differences in either slope estimates and standard errors.

The analysis in Table 5.1 depicts that, time, gender and HLA-B*1510 effects have significant p-values for incomplete and MI method, indicating a rejection of the null hypothesis. The $p-$value for gender under MI (0.0372) was slightly higher, compared to that of the

incomplete data analysis (0.0357). Moreover, both the MI and incomplete data analysis provide strong evidences of significance for the intercept and time effects as their $p-$values was (<0.0001). The MI method produced significant $p-$values for $PPIA_{1650}$ and IL-10(-592) effects which was not the case under the incomplete data analysis, showing one of the major challenges of drop-out, as incomplete data may lead to non-rejection of the null hypothesis than would be the case if the data was complete. For IL-10(-1082) and HLA-B*5802, the results obtained from the MI method and those obtained from incomplete dataset were not that different (they were closer to one another). Evidently, there can be extreme differences between the MI and incomplete original analysis for the other effects.

5.5.2 Results for Handling Missing Outcome and Covariates

TABLE 5.2: Parameter estimates, standard errors and p-values of the covariates before and after handling drop-outs (interaction terms not shown)

	Incomplete data			MI		
effect	parameters	errors	p-value	parameters	errors	p-value
intercept	3.7853	0.3745	< 0.0001	3.8947	0.2761	< 0.0001
time	-0.0841	0.0070	< 0.0001	-0.0989	0.0042	< 0.0001
APOBEC3G	-0.0563	0.0817	0.4907	-0.0530	0.0579	0.3597
PSIP(rs2339417)	0.0993	0.0782	0.2041	0.0297	0.0555	0.5927
TNP03	0.1647	0.1036	0.1121	0.0896	0.0739	0.2258
PPIA(1650)	0.1353	0.0707	0.0562	0.1677	0.0502	0.0008
IL-10(-592)	0.1480	0.0900	0.1004	0.1672	0.0626	0.0077
IL-10(-1082)	0.0604	0.1210	0.6175	0.0843	0.0879	0.3374
IL-10(-3575)	0.0185	0.1384	0.8934	0.1157	0.0976	0.2363
Gender	-0.2762	0.1314	0.0357	-0.0904	0.0927	0.0329
HLA-B*0801	0.0336	0.1911	0.8605	0.0495	0.1325	0.7081
HLA-B*5801	0.0769	0.1678	0.6466	0.2082	0.1261	0.0988
HLA-B*1503	-0.0085	0.1383	0.9506	-0.1213	0.0999	0.2251
HLA-B*5802	0.2168	0.1426	0.1286	0.3303	0.0993	0.0009
HLA-B*1510	0.3576	0.1496	0.0169	0.0.1731	0.1082	0.1097
HLA-B*4201	0.1917	0.1344	0.1538	0.1700	0.0910	0.0620

Table 5.2 gives the results for parameter estimates, standard errors and p-values of the genetic factors on HIV Bio-marker data for handling missing values in the outcome and the covariates. From Table 5.2 we can clearly see that, the results obtained from the MI method are much different from those obtained from analysing the incomplete datasets. The difference in magnitude for TNPO3, IL-10(-3575), gender, HLA-B*5801, HLA-B*1503 and HLA-B*1510 were a lot in-terms of parameter estimates compared to those obtained from incomplete datasets. However, the difference was not that large for other covariates. Just as observed in Table 5.1, the standard errors associated with the MI method were much lower compared to those obtained from the incomplete dataset. These were much lower than the standard errors shown in Table 5.1 for handling only missing

outcome. Indicating that, missing values can lead to inefficient results if missingness is not accounted for in both the outcome and the covariates.

Looking at the $p-$values in Table 5.2, we can clearly see that the MI method produced statistically different $p-$values (though not always), except for few exceptions where the the MI method produced the same results as those given by the incomplete dataset i.e. for the intercept, time and gender. In these cases, the $p-$values were less than 0.05 and would result into a rejection of the null hypothesis at the 5% significance level. From Table 5.2 we can see that, the MI method gave non-significant $p-$value for HLA-B*1510, which was not the case as shown in Table 5.1. This indicates that, the difference is likely to be as a result of correcting for missing values in both the outcome and covariates. The $p-$values for $PPIA_{1650}$ and IL-10(-592) effects were highly significant under the MI method. However, they were non-significant under the incomplete data results. Which is expected as both effects have missing measurements for some individuals.

In this chapter, we defined drop-out, some of it's cause, the pattern it can take, it's mechanisms and some of the techniques of dealing with it. We should note that, there are several other strategies that can be used to deal with incomplete longitudinal data with continuous outcome under ignorability assumption, valid under MAR assumption; however, the scope of this chapter is limited to application of only MI method. For instance, the direct likelihood approach which makes use of the observed data without the need of deletion or imputation, the EM algorithm by Dempster et al. (1977) which is an alternative method to use for handling incomplete longitudinal data, as well as the mixed effects model repeated measures analysis (MMRM) by Mallinckrodt et al. (2001) which is a particular form of a linear mixed model fitted within direct likelihood analysis.

We concentrated more on the use of multiple imputation (MI) method for handling incomplete longitudinal count data due to drop-out. Comparing the results obtained from LOCF to those obtained from MI would have been a good idea, since it would have highlighted some of the strengths of MI, but this was not the scope of our study and that is why we also did not apply GEE to our data. We illustrated the application and compared results of analysis based on the original dataset with those based on the MI method using a longitudinal clinical trial data from the Sinikithemba study. MI was selected for its solid foundation on the MAR drop-out mechanism and being that it can be applied in longitudinal studies. Our objective for the chapter was to investigate the influence of drop-out on the evolution of HIV Bio-marker (log viral load) in a model including genetic factors as covariates. We wanted to see whether missingness of information has an effect on statistical analysis and inference. For this reason, we did not consider all the HIV Bio-markers (VL and CD4 count) available in the Sinikithemba

study but only used the data with log viral load as the response so see whether drop-out has an influence on statistical analysis and inference. We investigated the influence on inference that might be caused on the data by the drop-out process. The results from the incomplete dataset were found to be analogous to those from the MI method. Therefore, missingness has an effect on statistical analysis and inference. The next chapter is a summary of the whole dissertation.

Chapter 6

Conclusion

Since the discovery of AIDS among the gay men in 1981 in the United States of America. It has became a major world pandemic with over 40 million infected world wide. Numerous studies have been carried out to understand the pathogenesis and the dynamics of this deadly disease but, still its pathogenesis is poorly understood. Further understanding of the disease is still needed so as to reduce the rate of its acquisition. Researchers have atleast come-up with statistical and mathematical models which help in understanding and predicting the progression of the disease better so as to find ways in which its acquisition can be prevented and controlled.

In this dissertation, we used the longitudinal clinical trial data from the Sinikithemba study at Nelson Mandela Medical School, University of Kwazulu Natal in Durban to bring out the effects and contribution of the immuno genetic factors (HLA types, SNPs and IL-10) in the disease pathogenesis using HIV Bio-markers as indicators for HIV progression. The study cohort consisted of 451 HIV naive individuals of which 79.06 % were females and 20.4 % were males. The individuals in the study were followed after every three and six months for their CD count and viral load to be measured respectively. They were counselled, guided and referred to the government medical sector for ARV initiation as per the national guidance given by South African government whenever their CD4 count fell below 350 cells for more conservative visits. They were also advised to start on highly active antiretroviral therapy (HAART) once their CD4 count fell below 200 cells.

We plotted the graphs for the males and females for their CD4 count and viral load data respectively. From the two graphs, we could easily see that, the behaviour patterns in males and females for the CD4 count and viral load data were similar, except for some few outliers. The original CD4 count and the viral load data failed to satisfy the normality distribution assumptions. Therefore, there was need to perform transformations on the CD4 count and viral load data. Square root CD4 count and log viral load

were considered the best for our further analysis after performing a transformation. We also performed Kruskal-Wallis rank sum test to see whether there existed any relationship or variability between the different HLA subtypes alleles, and found there was. We calculated allele frequencies for IL-10(-592, -1082 and -3575) and for the SNPs ($PPIA_{1650}$, $PSIP_{rs12339417}$, TNPO3 and APOBEC3G) to identify the major and minor alleles and also checked whether the alleles were in Hardy Weinberg Equilibrium (HWE). The Kruskal-Wallis test between the median CD4 count and the various IL-10 showed no significance even with the pairwise comparison test using wilcoxon test. For the viral load data, there was no observable significance for Kruskal-Wallis test but pairwise comparison test using wilcoxon test showed some significance for IL-10(-3575)"AA","TA" genotypes.

We identified linear mixed model as the most appropriate model for our study since the study involved missing values, inconsistent timed observations and incomplete data with random intercept model as the best model for our analysis. Indicating that there were a lot of variability among the individuals due to individual to individual heterogeneity. We identified spatial exponential structure as the best covariance structure for log viral load and square root CD4 count data with ML and REML as the estimation methods, respectively. We fitted a random intercept model with our best covariance and mean structure and found that, individuals with APOBEC3G (CG) genotype have a significantly high mean square root CD4 count and a low log viral load compared to those with the CC genotype. Implying that, APOBEC3G (CG) genotype could be a good controller of HIV. Individuals with HLA-B*0801 allele and HLA-B*1510 allele have a significantly low mean square root CD4 count compared to those without the two alleles. Therefore, HLA-B*0801 and HLA-B*1510 alleles were possibly associated with fast progression of HIV-1. Reaffirming previous studies carried out by Turnbull et al. (2006) which showed that, HLA-B*0801 allele could be associated with fast HIV-1 progression in its acute phase.

Individuals with interactions between APOBEC3G (CG) genotype and $PSIP_{rs12339417}$ CT genotype, APOBEC3G (CG) genotype and $PSIP_{rs12339417}$ TT genotype, HLA-B*0801 allele and $PPIA_{1650}$ AG genotype and lastly HLA-B*0801 allele and TNPO3 (AG) genotype have a significantly low mean square root CD4 count. Male individuals having HLA-B*0801 showed a significantly low mean square root CD4 count. Indicating that, HLA-B*0801 allele in males could be associated with fast progression of HIV-1. Individuals with interactions between HLA-B*0801 allele and IL-10(-592) CA genotype, HLA-B*1510 allele and $PPIA_{1650}$ GG genotype, $PSIP_{rs12339417}$ CT genotype and IL-10(-3575) TT genotype have a significantly high mean square root CD4 count and therefore such genetic factors are possibly considered as relative controllers of HIV-1. Individuals with IL-10(-592) CA genotype with time showed a significantly high

mean square root CD4 count and a low mean log viral load implying that IL-10(-592) CA genotype could be a good controller of HIV. Gender was significantly associated with low log viral load. Showing that, females possibly progress to HIV at a slower rate compared to males. Reaffirming previous studies carried out by Farzadegan et al. (1998) which showed that, HIV-1 viral load was substantially lower in women than in men at a common threshold of CD4 cell count.

Individuals with HLA-B*5802 allele were found to be significantly associated with a high log viral load. Therefore, HLA-B*5802 allele could be associated with a fast progression of HIV-1. Male individuals with APOBEC3G CG genotype have a significantly low mean log viral load. Therefore, there was a possibility that they progressed to HIV-1 at a slower rate. Individuals with APOBEC3G CG genotype and HLA-B*0801 allele have a significantly low log viral load. Implying that, individuals with such genetic factors possibly progress to HIV-1 slightly slower. Those individuals with $PPIA_{1650}$ GG genotype and IL-10(-592) CA genotype have a significantly high mean log viral load. Therefore, they possibly progressed to HIV-1 at a faster rate. Male individuals with the IL-10(-3575) TA and IL-10(-592) CA genotype have a significantly high mean log viral load. Indicating that, they could progress to HIV faster. Reaffirming previous studies carried out by Shin et al. (2000) which showed that, individuals with the -592A promoter allele were possibly associated to progression of HIV-1.

Drop-out is one of the major challenges commonly encountered by longitudinal studies, regardless of how well they are designed and executed. In the Sinikithemba study, the outcome VL and $CD4$ count and some covariates (APOBEC3G, $PSIP_{rs12339417}$, TNPO3 and $PPIA_{1650}$) contained drop-outs may be as a result of non-response by some individuals who leave the study and thereafter lost to follow-up, sickness, equipment failure, poor data entry, weather conditions and death among others.

In this dissertation, we assumed the missing mechanism was MAR, since its the most plausible and used the MI method in our analysis since it prevents the reduction of the sample size, helps in avoiding biased estimates and also appears to be the most attractive and powerful technique for handling drop-out in longitudinal studies. We illustrated the application and compared results of analysis of parameter estimates, standard errors and p-values based on the original dataset with those based on the MI method using a longitudinal clinical trial data from the Sinikithemba study with the log viral load as the response. Our aim was to investigate the influence of drop-out on the evolution of HIV-Bio marker, in a model including genetic factors as covariates.

We dealt with two scenarios: (1) Missing outcome and (2) Missing outcome and covariates. Investigating the influence on inference that might be caused on the data as a result of drop-out. In the first instance, we observed that, the parameter estimates associated with the MI method were different from those obtained from the original incomplete

dataset. The standard errors from the incomplete dataset were generally larger compared to those obtained from the MI method. This was because, the results obtained from the incomplete datasets were associated with the inflated standard errors. The same scenario was observed when we handled the missing outcome and covariates. The parameter estimates obtained from the MI method were much different from those obtained from the incomplete datasets. Just as observed earlier when dealing with missing outcome, the parameter estimates were much smaller in size compared to those obtained from the incomplete dataset. Indicating that, there exists extreme difference between the results obtained from the original datasets and those obtained from the MI method. Suggesting that, missing values can lead to inefficient results if not accounted for. Based on the application results we reached the following conclusions:

(i) Findings in general suggested that the conclusions obtained under both analyses for the HIV Bio-markers data practically were not in agreement as they lead to different results. Indicating that, drop-outs have a substantial impact on HIV Bio-markers data. Therefore, the need to address drop-outs problems in order to avoid biased results. We need to note that, there is no universal technique for handling all the drop-out situations, however, there are some rules that can be considered. In this dissertation, we focussed more on the MI method, as this study pays attention to modern procedures that can be useful under most circumstances in missing data problems. There are other ways of dealing with drop-outs i.e. Expectation Maximization (EM) algorithm Dempster et al. (1977) however, we used the MI method for our study. Furthermore, since drop-out processes are often unverifiable, one of the recommendation in many settings is to apply multiple strategies or models, such as selection and pattern mixture models to the same HIV Bio-markers data in order to investigate the impact of assumption on drop-out.

(ii) In this dissertation we emphasized on the MAR missingness mechanism because it is the most plausible assumption, as the MCAR assumption is not easy to justify in that, it is restrictive to generally hold (Molenberghs and Kenward, 2007). MNAR modelling techniques can also be used to deal with non-random drop-out by explicitly modelling the assumption that caused the drop-out and incorporating this additional complexity into the model. Here, we agreed that it's impossible to distinguish which underlying missingness mechanisms are in play, unless one knows the motivation for and individual dropping out. This problem is discussed further in Molenberghs et al. (2008), they show that a formal distinction between MAR and MNAR is not possible. This is because for any MNAR model there exists an MAR model that fits the data equally well, but they differ in the prediction of what is unobserved. Hence, it is broadly agreed that, the role of such MNAR

model is in sensitivity analysis; that is, if the assumptions are changed, the conclusions from the primary (typically MAR) analysis are also changed. According to Molenberghs and Verbeke (2005), sensitivity analysis is defined as an analysis in which several statistical models are considered simultaneously under different missing data scenarios.

(iii) The major conclusion drawn from our study was that, there is need to obtain further insight into the HIV Bio-markers data by comparing various models in the presence of drop-out. Knowing the reasons why some of the HIV Bio-markers data are missing is also helpful and very important in choosing the right statistical procedures to approximate missingness.

Therefore, we conclude our dissertation by saying that, missingness of data poses a very serious challenge in longitudinal studies that needs to be accounted for. This is because, it can lead to inappropriate statistical inference and analysis. Hence, we conclude by suggesting that, there is need also to compare various sensitivity analysis frameworks, shared parameters, selection and pattern mixtures models.

Linear Mixed Models

**********Random Intercept Model for Square Root CD4 count**********

Table 1: Linear mixed effects model fit by REML, Data: pp0x

AIC	BIC)	logLik
13439.24	14151.46	-6599.618

Random effects: Formula: 1|PID (Intercept) Residual StdDev: 4.383928 2.694175

Correlation Structure: Exponential spatial correlation Formula: 1|PID Parameter estimate(s): range 1.74549
Fixed effects: sqrt(CD) = apobec + HLA-B*0801 + HLA-B*1510 + HLA-B*4201 + HLA-B*5801 + HLA-B*5802 + PPIA(1650) + PSIP(rs12339417) + time + Gender + TNPO3 + IL-10(-3575) + IL-10(-592) + IL-10(-1082) + APOBEC3G:HLA-B*0801 + APOBEC3G:PSIP(rs12339417) + APOBEC3G:Gender + APOBEC3G:IL-10(-592) + APOBEC3G:IL-10(-1082) + HLA-B*0801:HLA-B*1510 + HLA-B*0801:PPIA(1650) + HLA-B*0801:PSIP(rs12339417) + HLA-B*0801:time + HLA-B*0801:Gender + HLA-B*0801:TNPO3 + HLA-B*0801:IL-10(-592) + HLA-B*1510:PPIA(1650) + HLA-B*1510:PSIP(rs12339 + HLA-B*1510:time + HLA-B*1510:IL-10(-3575) + HLA-B*1510:IL-10(-592) + HLA-B*1510:IL-10(-1082) + HLA-B*5802:time + PPIA(1650):PSIP(rs12339417) + PPIA(1650):time + PPIA(1650):IL-10(-592) + PPIA(1650):IL-10(-1082) + PSIP(rs12339417):Gender + PSIP(rs12339417):IL-10(-3575) + PSIP(rs12339417):IL-10(-592) + PSIP(rs12339417):IL-10(-1082) + time:TNPO3 + time:IL-10(-3575) + time:IL-10(-592) + time:IL-10(-1082) + Gender:TNPO3 + Gender:IL-10(-3575) + Gender:IL-10(-592) + Gender:IL-10(-1082) + TNPO3:IL-10(-3575) + TNPO3:IL-10(-1082)

**********Random Intercept Model for Log Viral Load**********

Table 2: Linear mixed effects model fit by ML, Data: bbn

AIC	BIC)	logLik
7297.807	7921.218	-3534.904

Random effects: Formula: 1 |PID (Intercept) Residual StdDev: 0.0003518797 2.258973

Correlation Structure: Exponential spatial correlation Formula: 1|PID Parameter estimate(s): range 2.17669
Fixed effects: log(VL) =APOBEC3G + HLA-B*0801 + HLA-B*1510 + HLA-B*4201 + HLA-B*5801 + HLA-B*5802 + PPIA(1650) + PSIP(rs12339417) + time + Gender + TNPO3 + IL-10(-3575) + IL-10(-592) + IL-10(-1082) + APOBEC3G:HLA-B*1510 + APOBEC3G:PSIP(rs12339417) + HLA-B*1510:PPIA(rs12339417)+ APOBEC3G:time + APOBEC3G:Gender + APOBEC3G:IL-10(-592) + APOBEC3G:HLA-B*0801 + APOBEC3G:IL-10(-1082) + HLA-B*0801:HLA-B*1510 + HLA-B*0801:PPIA(1650) + HLA-B*0801:PSIP(rs12339417) + HLA-B*0801:time + HLA-B*0801:Gender + HLA-B.0801:TNPO3 + HLA-B*0801:IL-10(-592) + HLA-B*1510:PSIP(rs12339417) + HLA-B*1510:time + HLA-B*1510:Gender + HLA-B*1510:IL-10(-3575) + HLA-B*1510:IL-10(-592) + HLA-B*1510:IL-10(-1082) + HLA-B*4201:time + HLA-B*5802:time + PPIA(1650):PSIP(rs12339417) +:time + PPIA(rs12339417):IL-10(-592) + PPIA(1650):IL-10(-1082) + PSIP(rs12339417):IL-10(-592) + time:Gender + time:TNPO3 + time:IL-10(-1082) + Gender:TNPO3 + Gender:IL-10(-3575) + Gender:IL-10(-592)+ PSIP(rs12339417):IL-10(-1082) + PSIP(rs12339417):IL-10(3575) + PSIP(rs12339417):Gender

Bibliography

Al-Jabri, A. A. (2007). Mechanisms of host resistance against HIV infection and progression to AIDS. *Sultan Qaboos University Medical Journal*, 7(2):82.

Ali, H. A. (2007). The new approach to guide the selection of the covariance structure in mixed model. *Research journal of medicine and medical sciences*, 2(2):88–97.

Ali, M. W. and Talukder, E. (2005). Analysis of longitudinal binary data with missing data due to dropouts. *Journal of Biopharmaceutical Statistics*, 15(6):993–1007.

An, P., Li, R., Wang, J. M., Yoshimura, T., Takahashi, M., Samudralal, R., O'Brien, S. J., Phair, J., Goedert, J. J., Kirk, G. D., et al. (2011). Role of exonic variation in chemokine receptor genes on AIDS: CCRL2 F167Y association with pneumocystis pneumonia. *PLoS genetics*, 7(10):e1002328.

An, P., Wang, L. H., Hutcheson-Dilks, H., Nelson, G., Donfield, S., Goedert, J. J., Rinaldo, C. R., Buchbinder, S., Kirk, G. D., O'Brien, S. J., et al. (2007). Regulatory polymorphisms in the cyclophilin A gene, PPIA, accelerate progression to AIDS. *PLoS pathogens*, 3(6):e88.

Asadullah, K., Eskdale, J., Wiese, A., Gallagher, G., Friedrich, M., and Sterry, W. (2001). Interleukin-10 promoter polymorphism in psoriasis. *Journal of investigative dermatology*, 116(6):975–978.

Barker, E., Mackewicz, C. E., Reyes-Terán, G., Sato, A., Stranford, S. A., Fujimura, S. H., Christopherson, C., Chang, S., and Levy, J. (1998). Virological and immunological features of long-term human immunodeficiency virus–infected individuals who have remained asymptomatic compared with those who have progressed to acquired immunodeficiency syndrome. *Blood*, 92(9):3105–3114.

Barnett, A. G., Koper, N., Dobson, A. J., Schmiegelow, F., and Manseau, M. (2010). Using information criteria to select the correct variance–covariance structure for longitudinal data in ecology. *Methods in Ecology and Evolution*, 1(1):15–24.

Bentler, P. M. and Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological bulletin*, 88(3):588–606.

Boldman, K. G. and Van, V. L. D. (1991). Derivative-free restricted maximum likelihood estimation in animal models with a sparse matrix solver. *Journal of dairy science*, 74(12):4337–4343.

Borghans, J. A. M., Mølgaard, A., De Boer, R. J., and Keşmir, C. (2007). Hla alleles associated with slow progression to AIDS truly prefer to present HIV-1 p24. *PLoS One*, 2(9):e920.

Brennan, P. A. and Kendrick, K. M. (2006). Mammalian social odours: attraction and individual recognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1476):2061–2078.

Brookes, A. J. (1999). The essence of SNPS. *Gene*, 234:177–186.

Bryan, S. R. (2011). Modelling longitudinally measured outcome HIV bio-makers with immuno genetic parameters. Master's thesis, Univesity of KwaZulu-Natal.

Carlson, C. S., Eberle, M. A., Rieder, M. J., Yi, Q., Kruglyak, L., and Nickerson, D. A. (2004). Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *The American Journal of Human Genetics*, 74(1):106–120.

Carrington, M. and O'Brien, S. J. (2003). The influence of HLA genotype on AIDS*. *Annual review of medicine*, 54(1):535–551.

Chatterjee, K. (2010). Host genetic factors in susceptibility to HIV-1 infection and progression to AIDS. *Journal of genetics*, 89(1):109–116.

De, L. E. D., Hox, J. J., and Huisman, M. (2003). Prevention and treatment of item nonresponse. *Journal of Official statistics*, 19:153–176.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:1–38.

Diaz-Griffero, F. (2012). The role of TNPO3 in HIV-1 replication. *Molecular biology international*, 2012.

Diggle, P. J., Heagerty, P., Liang, K. Y., and Zeger, S. L. (2002). *Analysis of longitudinal data*. New York: Oxford University Press.

Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. *Psychological Methods*, 16:1–16.

Farzadegan, H., Hoover, D. R., Astemborski, J., Lyles, C. M., Margolick, J. B., Markham, R. B., Quinn, T. C., and Vlahov, D. (1998). Sex differences in HIV-1 viral load and progression to AIDS. *The Lancet*, 352:1510–1514.

Fauci, A. S. et al. (2003). HIV and AIDS: 20 years of science. *Nature Medicine*, 9(7):839–843.

Feero, W. G., Guttmacher, A. E., and Collins, F. S. (2010). Genomic medicine. *Journal of Medicine*, 362:2001–2011.

Fitzmaurice, G. M. (2003). Methods for handling dropouts in longitudinal clinical trials. *Statistica Neerlandica*, 57(1):75–99.

Fitzmaurice, G. M. and Laird, N. M. (2000). Generalized linear mixture models for handling nonignorable dropouts in longitudinal studies. *Biostatistics*, 1:141–156.

Foulkes, A. S. (2009). *Applied statistical genetics with R: for population-based association studies*. Springer.

Frahm, N., Adams, S., Kiepiela, P., Linde, C. H., Hewitt, H. S., Lichterfeld, M., Sango, K., Brown, N. V., Pae, E., Wurcel, A. G., et al. (2005). Hla-b63 presents HLA-B57/b58-restricted cytotoxic T-lymphocyte epitopes and is associated with low human immunodeficiency virus load. *Journal of virology*, 79(16):10218–10225.

Gao, X., Nelson, G. W., Karacki, P., Martin, M. P., Phair, J., Kaslow, R., Goedert, J. J., Buchbinder, S., Hoots, K., Vlahov, D., et al. (2001). Effect of a single amino acid change in MHC class i molecules on the rate of progression to AIDS. *New England Journal of Medicine*, 344(22):1668–1675.

Gillespie, G. M. A., Kaul, R., Dong, T., Yang, H., Tim, R., Bwayo, J. J., Kiama, P., Peto, T., Plummer, F. A., McMichael, A. J., et al. (2002). Cross-reactive cytotoxic T lymphocytes against a HIV-1 p24 epitope in slow progressors with B* 57. *Aids*, 16(7):961–972.

Hedeker, D. and Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, 2(1):64.

Heyting, A., Tolboom, J. T. B. M., and Essers, J. G. A. (1992). Statistical handling of drop-outs in longitudinal clinical trials. *Statistics in medicine*, 11(16):2043–2061.

Horton, N. J. and Kleinman, K. P. (2007). Much ado about nothing. *The American Statistician*, 61(1).

Huang, X., Ling, H., Feng, L., Ding, X., Zhou, Q., Han, M., Mao, W., and Xiong, H. (2009). Human leukocyte antigen profile in HIV-1 infected individuals and AIDS patients from Chongqing, China. *Microbiology and immunology*, 53(9):512–523.

Jin, X., Wu, H., and Smith, H. (2007). APOBEC3G levels predict rates of progression to AIDS. *Retrovirology*, 4(1):20.

Joint United Nations Programme on HIV |AIDS, U. G. H. r. e. u. and health sector progress towards universal access: progress report (2011). Technical report, World health organisation and others.

Jones, R. H. (1993). *Longitudinal data with serial correlation: a state-space approach*, volume 47. Chapman & Hall.

Kanki, P. J., Hamel, D. J., Sankalé, J., Hsieh, C., Thior, I., Barin, F., Woodcock, S. A., Guèye-Ndiaye, A., Zhang, E., Montano, M., et al. (1999). Human immunodeficiency virus type 1 subtypes differ in disease progression. *Journal of Infectious Diseases*, 179(1):68–73.

Kaur, G. and Mehra, N. (2009). Genetic determinants of HIV-1 infection and progression to AIDS: susceptibility to HIV infection. *Tissue antigens*, 73(4):289–301.

Klein, M. R., Van der Burg, S. H., Hovenkamp, E., Holwerda, A. M., Drijfhout, J. W., Melief, C. J., and Miedema, F. (1998). Characterization of HLA-B57-restricted human immunodeficiency virus type 1 Gag-and RT-specific cytotoxic T lymphocyte responses. *Journal of general virology*, 79(9):2191–2201.

Kumar, S. and Subramanian, S. (2002). Mutation rates in mammalian genomes. *Proceedings of the National Academy of Sciences*, 99(2):803–808.

Kutner, M. H., Nachtsheim, C., and Neter, J. (2004). Applied linear regression models.

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

Levsky, J. M. and Singer, R. H. (2003). Fluorescence in situ hybridization: past, present and future. *Journal of Cell Science*, 116(14):2833–2838.

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

Litt, M. and Luty, J. A. (1989). A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *American journal of human genetics*, 44(3):397.

Littell, R. C., Stroup, W. W., and Freund, R. J. (2002). *SAS for linear models*. SAS Institute.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical analysis with missing data*, volume 539. Wiley New York.

Little, R. J. A. and Schenker, N. (1995). Handbook of statistical methods—missing data.

Mallinckrodt, C. H., Clark, W. S., and David, S. R. (2001). Accounting for dropout bias using mixed-effects models. *Journal of biopharmaceutical statistics*, 11:9–21.

Martin, M. P., Qi, Y., Gao, X., Yamada, E., Martin, J. N., Pereyra, F., Colombo, S., Brown, E. E., Shupert, W. L., Phair, J., et al. (2007). Innate partnership of HLA-B and KIR3DL1 subtypes against HIV-1. *Nature genetics*, 39(6):733–740.

McCulloch, C. E. (2006). *Generalized linear mixed models*. Wiley Online Library.

Meier, B. P. and Robinson, M. D. (2005). The metaphorical representation of affect. *Metaphor and Symbol*, 20(4):239–257.

Meng, X. (1994). Multiple imputation inferences with uncongenial sources of input. *Statistical Science*, pages 538–558.

Molenberghs, G., Beunckens, C., Sotto, C., and Kenward, M. (2008). Every missing is not at random model has got a missing at random counterpart with equal fit. *Journal of Royal Statistics Society Series*, 70:371–388.

Molenberghs, G. and Kenward, M. (2007). *Missing data in clinical studies*. John Wiley & Sons.

Molenberghs, G. and Verbeke, G. (2005). *Models for discrete longitudinal data*. Springer.

Naicker, D. D. (2011). The role of interleukin-10 promoter polymorphisms in HIV-1 susceptibility and primary HIV-1 pathogenesis.

Naicker, D. D., Wang, B., Losina, E., Zupkosky, J., Bryan, S., Reddy, S., Jaggernath, M., Mokgoro, M., Goulder, P. J. R., Kaufmann, D. E., et al. (2012). Association of il-10-promoter genetic variants with the rate of CD4 T-cell loss, IL-10 plasma levels, and breadth of cytotoxic T-cell lymphocyte response during chronic HIV-1 infection. *Clinical Infectious Diseases*, 54(2):294–302.

Naicker, D. D., Werner, L., Kormuth, E., Passmore, J., Mlisana, K., Karim, S. A., and Ndung'u, T. (2009). Interleukin-10 promoter polymorphisms influence HIV-1 susceptibility and primary HIV-1 pathogenesis. *Journal of Infectious Diseases*, 200(3):448–452.

Nakai, M. and Ke, W. (2011). Review of the methods for handling missing data in longitudinal data analysis. *International Journal of Mathematical Analysis*, 5(1):1–13.

Nelwamondo, F. V., Mohamed, S., and Marwala, T. (2007). Missing data: A comparison of neural network and expectation maximisation techniques. *arXiv preprint arXiv:0704.3474*.

Peng, C. J., Harwell, M., Liou, S., and Ehman, L. H. (2007). Advances in missing data methods and implications for educational research. *Real data analysis*, pages 31–78.

Peng, C. J. and Zhu, J. (2008). Comparison of two approaches for handling missing covariates in logistic regression. *Educational and Psychological Measurement*, 68(1):58–77.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, pages 1033–1048.

Rajapaksa, U. S., Li, D., Peng, Y., McMichael, A. J., Dong, T., and Xu, X. (2012). HLA-B may be more protective against HIV-1 than HLA-A because it resists negative regulatory factor (nef) mediated down-regulation. *Proceedings of the National Academy of Sciences*, 109(33):13353–13358.

Ramroop, S. (2008). *Analysis of longitudinal binary data: An application to a disease process*. PhD thesis, University of KwaZulu-Natal.

Reddy, K., Winkler, C. A., Werner, L., Mlisana, K., Abdool Karim, S. S., Ndung'u, T., et al. (2010). APOBEC3G expression is dysregulated in primary HIV-1 infection and polymorphic variants influence CD4+ T-cell counts and plasma viral load. *Aids*, 24(2):195.

Reddy, T., Mwambi, H., and Ndung'u, T. (2011). Modelling HIV progression using multistate models. *Proceedings of 53rd Annual Conference of the South Africa Statistical Association*, pages 100–117.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121.

Roger, M. (1998). Influence of host genes on HIV-1 disease progression. *The FASEB journal*, 12(9):625–632.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63:581–592.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. *New York, USA: John Willey & Sons*.

Schafer, J. L. (1997). Analysis of incomplete multivariate data. *Chapman &Hall, London*.

Schafer, J. L. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica*, 57(1):19–35.

Schafer, J. L. and Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological methods*, 7:147–177.

Schafer, J. L. and Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate behavioral research*, 4:545–571.

Shin, H. D., Winkler, C., Stephens, J. C., Bream, J., Young, H., Goedert, J. J., O'Brien, T. R., Vlahov, D., Buchbinder, S., Giorgi, J., et al. (2000). Genetic restriction of HIV-1 pathogenesis to AIDS by promoter alleles of IL10. *Proceedings of the National Academy of Sciences*, 97(26):14467–14472.

Silva, E. and Stumpf, M. P. H. (2004). Hiv and the ccr5-Δ32 resistance allele. *FEMS microbiology letters*, 241(1):1–12.

South Africa National Department of Health, N. A. T. G. P. S. A. (2004). Technical report, Ministry of health.

Trachtenberg, E. A. and Erlich, H. A. (2001). A review of the role of the human leukocyte antigen (HLA) system as a host immunogenetic factor influencing HIV transmission and progression to AIDS. *HIV Molecular Immunology. Los Alamos, New Mexico. USA*.

Turnbull, E. L., Lopes, A. R., Jones, N. A., Cornforth, D., Newton, P., Aldam, D., Pellegrino, P., Turner, J., Williams, I., Wilson, C. M., et al. (2006). Hiv-1 epitope-specific CD8+ T cell responses strongly associated with delayed disease progression cross-recognize epitope variants efficiently. *The Journal of Immunology*, 176:6130–6146.

Van Buuren, S., Boshuizen, H. C., Knook, D. L., et al. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in medicine*, 18:681–694.

Verbeke, G. and Molenberghs, G. (2000). Linear mixed models for longitudinal data Springer-Verlag. *Springer series. New York.*

Verbeke, G. and Molenberghs, G. (2005). *Linear mixed models for longitudinal data.* Springer series. New York.

Verbeke, G. and Molenberghs, G. (2009). *Linear mixed models for longitudinal data.* Springer Verlag. New York.

Walker, M. B. (2004). Assessing the barriers to universal antiretroviral treatment access for HIV/AIDS in South Africa. *Duke J. Comp. & Int'l L.*, 15:193.

Wang, Z. and Moult, J. (2001). SNPs, protein structure, and disease. *Human mutation*, 17(4):263–270.

Weber, J. L. and May, P. E. (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *American journal of human genetics*, 44(3):388.

Wener, L. (2009). Modell acute HIV-1 infection using longitudinally measured bio-maker data including informative drop-out. *Masters dissartation, University of KwaZulu Natal*.

Wolfe, K. H., Sharp, P. M., and Li, W. (1989). Mutation rates differ among regions of the mammalian genome.

Wu, M. C. and Bailey, K. R. (1989). Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics*, 3:939–955.

Yuan, Y. C. (2010). Multiple imputation for missing data: Concepts and new development (version 9.0). *SAS Institute Inc, Rockville, MD*.

Ziegler, A., König, I. R., and Pahlke, F. (2010). *A Statistical approach to genetic epidemiology: Concepts and applications, with an e-learning platform*. Wiley-VCh.