MODELLING LONGITUDINALLY MEASURED OUTCOME HIV BIOMARKERS WITH IMMUNO GENETIC PARAMETERS

by

SUSAN RUTH BRYAN

Submitted in fulfillment of the academic

requirements for the degree of

Master of Science

in

Statistics

in the

School of Statistics and Actuarial Science,

University of KwaZulu-Natal

Pietermaritzburg

2011

Declaration

I, Susan Ruth Bryan, declare that

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

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

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

4. This dissertation does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:
   a. Their words have been re-written but the general information attributed to them has been referenced
   b. Where their exact words have been used, then their writing has been placed in italics and inside quotation marks, and referenced.

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

Signed: ……

As the candidate's supervisor I have approved this dissertation for submission.

Name: _____ Signed: _____

Name: _____ Signed: _____

Acknowledgments

I would like to thank my supervisor Professor Henry Mwambi and my co-supervisor Dr Shaun Ramroop for their support and encouragement. Without them this would not have been possible.

I would like to thank Professor Thumbi Ndung'u for his advice and the HIV Pathogenesis Programme for the use of their data as well as the South African Centre for Epidemiology Modeling and Analysis for funding this research.

Thank you also to David, Irene, Sarah, Stephen and Kelly for always being there for me and to Bruce for his help.

Abstract

According to the Joint United Nations Programme against HIV/AIDS 2009 AIDS epidemic update, there were a total of 33.3 million (31.4 million–35.3 million) people living with HIV worldwide in 2009. The majority of the epidemic occurs in Sub-Saharan Africa. Of the 33.3 million people living with HIV worldwide in 2009, a vast majority of 22.5 million (20.9 million-24.2 million) were from Sub-Saharan Africa. There were 1.8 million (1.6 million-2.0 million) new infections and 1.3 million (1.1 million-1.5 million) AIDS-related deaths in Sub-Saharan Africa in 2009 (UNAIDS, 2009).

Statistical models and analysis are required in order to further understand the dynamics of HIV/AIDS and in the design of intervention and control strategies. Despite the prevalence of this disease, its pathogenesis is still poorly understood. A thorough understanding of HIV and factors that influence progression of the disease is required in order to prevent the further spread of the virus. Modelling provides us with a means to understand and predict the progression of the disease better.

Certain genetic factors play a key role in the way the disease progresses in a human body. For example HLA-B types and IL-10 genotypes are some of the genetic factors that have been independently associated with the control of HIV infection. Both HLA-B and IL-10 may influence the quality and magnitude of immune responses and IL-10 has also been shown to down regulate the expression of certain HLA molecules. Studies are therefore required to investigate how HLA-B types and IL-10 genotypes may interact to affect HIV infection outcomes.

This dissertation uses the Sinikithemba study data from the HIV Pathogenesis Programme (HPP) at the Medical School, University of KwaZulu-Natal involving 450 HIV positive and treatment naive individuals to model how certain outcome biomarkers (CD4+ counts and viral loads) are associated with immuno genetic parameters (HLA-B types and IL-10 genotypes). The work also seeks to exploit novel longitudinal data methods in Statistics in order to efficiently model longitudinally measured HIV

outcome data. Statistical techniques such as linear mixed models and generalized estimating equations were used to model this data. The findings from the current work agree quite closely with what is expected from the biological understanding of the disease.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

According to the Joint United Nations Programme against HIV/AIDS 2009 AIDS epidemic update, there were a total of 33.3 million (31.4 million–35.3 million) people living with HIV worldwide in 2009. This far exceeded the number reported in 1999, reflecting a 27% increase. A total of 2.6 million (2.3 million – 2.8 million) people were newly infected with HIV, coinciding with 1.8 million (1.6 million–2.1 million) AIDS-related deaths in 2009 (UNAIDS, 2009). As of December 2009, the majority of the epidemic occurred in Sub-Saharan Africa where 22.5 million (20.9 million-24.2 million) of the 33.3 million people living with HIV reside. There were 1.8 million (1.6 million-2.0 million) incident cases and 1.3 million (1.1 million-1.5 million) AIDS-related deaths in Sub-Saharan Africa in 2009 (UNAIDS, 2009). Despite the numerous studies about the disease, its pathogenesis is still poorly understood. A thorough understanding of HIV and the factors that influence the progression of the disease is required in order to prevent further spread of the virus. CD4+ (cluster of differentiation 4) count and viral load are the most common biomarkers of HIV used to monitor its progression in humans.

This dissertation will use the Sinikithemba cohort study conducted by the HIV Pathogenesis Programme (HPP) at the Nelson Mandela Medical School in the University of Kwa-Zulu Natal, Durban, consisting of 450 HIV positive and treatment naive individuals, to model how certain outcome biomarkers (CD4+ counts and viral loads) associate with immuno genetic parameters (HLA-B types and IL-10 genotypes). It will seek to exploit novel longitudinal data methods in order to efficiently model longitudinally measured HIV outcome data including information and effects of certain immune response genes on the pathogenesis of the disease.

1.1 Human Immunodeficiency Virus

AIDS (acquired immune deficiency syndrome) was first identified in the early 1980s (Averting HIV and AIDS, 2009). The virus causing this deadly disease, known as HIV (human immunodeficiency virus), attacks the most vulnerable part of the human body - the immune system. The virus is passed on from one person to the next via bodily fluids, i.e.: semen and blood. This can occur during sexual contact (anal, vaginal or oral), from the practice of sharing needles, prevalent amongst some drug addicts, through tattooing and body piercing practices which use unsterilized needles, accidental needle pricks, unsafe blood or blood products, or mother to child transmission (either during pregnancy, during delivery or through breast feeding). A person who contracts the virus (HIV) may live for many years before developing full blown AIDS. AIDS is not a specific illness, but rather a collection of different symptoms or conditions that manifest in the human body due to the weakened immune system of an infected person.

1.2 HIV Outcome Biomarkers

An individual's immune system contains different types of cells that help protect the body from infection. The T-cells or CD4+ cells are one of these specialized cells, known as "helper" cells. They are a type of white blood cell, which play an important role in the body's immune system, and therefore in fighting infections. HIV is a retrovirus, meaning that it needs cells from a "host" to replicate. CD4+ cells act as a host for HIV. HIV attaches to the CD4+ cells thus allowing the virus to enter and infect other/further infect CD4+ cells. During this process the CD4+ cells are damaged, leaving a weakened immune system.

A CD4+ count is a blood test which determines how well an individual's immune system is working by measuring the number of functioning CD4+ cells in the individual's body. A CD4+ count is measured as cells per cubic milliliter of blood. The lower the CD4+ count, the weaker the individual's immune system and the higher the risk to the individual for opportunistic infections.

The CD4+ count for an individual can fall in one of the following categories:

CD4+ count for a healthy adult	600–1200	Healthy HIV negative individual
CD4+ count for a HIV+ adult	350-600	Considered very good
CD4+ count for a HIV+ adult	200 – 350	Immune System is weakened and therefore the individual may be at increased risk for infection and illness
CD4+ count for a HIV+ adult	<200	Classified as having AIDS

Another imperative biomarker to consider is viral load. This is a blood test that measures the amount of active HIV in an individual's blood. This is stated as the number of HIV copies per milliliter of blood and is tested using reverse transcriptase polymerase chain reaction or PCR.

1.3 Natural History of HIV

Over several years, HIV infection reduces the number of T helper cells available to help fight disease and therefore damages the body's immune system. This process occurs in four stages, namely primary infection, clinically asymptomatic stage, symptomatic HIV infection, and progression from HIV to AIDS. The first stage is primary HIV infection. This normally lasts a few weeks and often includes flu-like symptoms. This may include a rash, fever, headaches, diarrhea and vomiting. Opportunistic infections are not seen at this stage. This stage is normally not severe and individuals rarely seek medical consultation. Besides possibly obtaining swollen glands, an individual will not experience any major symptoms during the asymptomatic stage of HIV infection. This stage typically lasts on average ten years. The individual remains healthy and unaware of the disease, with a CD4+ count above 500 cells per cubic milliliter of blood. In early symptomatic HIV disease, the immune system begins to fail and symptoms include fever, unexplained weight loss, recurrent diarrhea, fatigue and headaches. The individual's CD4+ count drops to below 200 cells per cubic milliliter of blood at this

stage, and usually anti retroviral therapy (ARV) is started. This stage is typically caused by the emergence of opportunistic infections. The individuals CD4+ count will continue to drop if the patient is not started on ARV's. The individual then progresses from HIV to AIDS and ultimately death. These stages can be seen graphically in Figure 1. Treatment strategies have recently recommended a CD4+ count of 200 as the cut off to begin ARV treatment (National Department of Health South Africa, 2010).

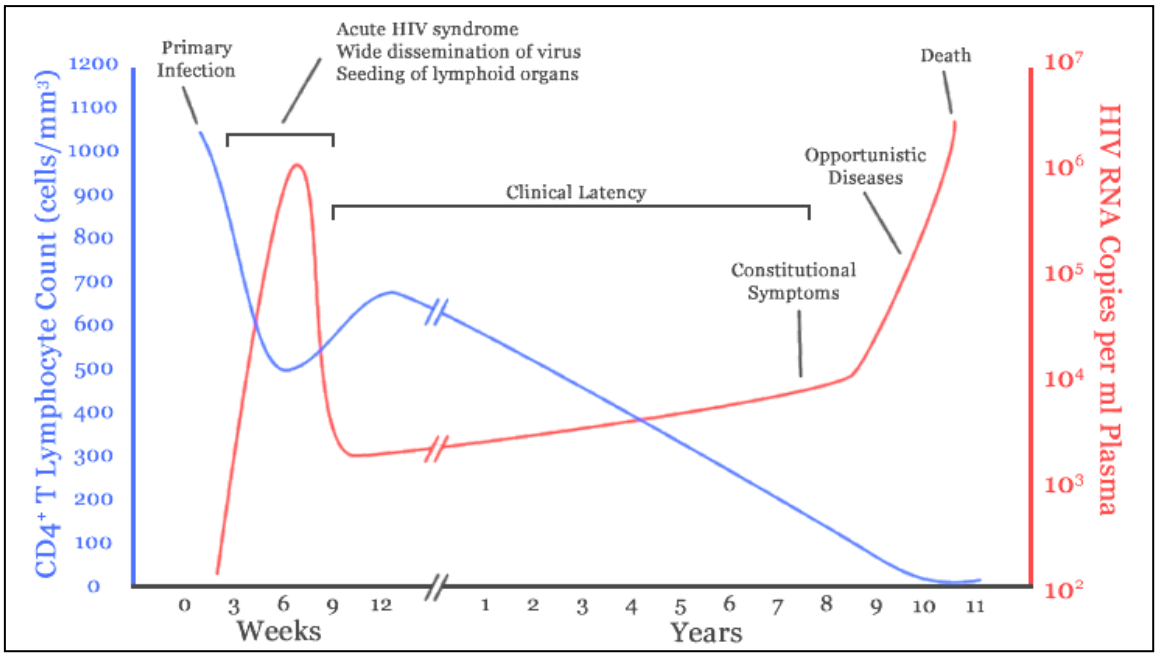


Figure 1: HIV progression over time

1.4 Immuno Genetic Parameters

The major histocompatibility complex (MHC) in humans is called the human leukocyte antigen system (HLA). These are a class of proteins which are often found on the surface membrane of cells. HLA's can be broken into categories, namely class I, class II and class III HLA. Class I HLA's include HLA-A, B and C while the class II HLA's includes DP, DQ, and DR. The super locus contains a large number of genes which are

found on chromosome 6, which is where the genes that produce HLA antigens are inherited. These genes are related to immune system function in humans. For the purpose of this dissertation we will be concentrating on the class I HLA-B types. Figure 2 shows the human chromosome 6 and the genes which are found on this chromosome.

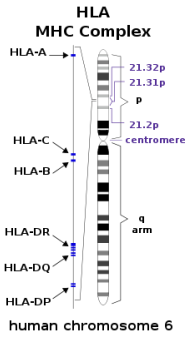


Figure 2: Human Chromosome 6

The roles of the different HLA classes are similar. The class I HLA antigens present peptides from inside the cell, whereas class II HLA antigens present antigens from outside of the cell to T-lymphocytes. Class III HLA antigens encode components of the complement system. There is a groove on each of the antigens, which attach bits of proteins or other antigens. These proteins can either be exterior to the cell, or created by the cell itself. The HLA then moves to the outside of the cell, where a T-cell can identify it. The T-cell checks whether the protein is foreign or not, and will either pass over or destroy the foreign cell.

Other important roles include disease defense. The antigens may be the cause of organ transplant rejections, and also may protect against or fail to protect against cancers. They may also mediate auto-immune disease, such as type I diabetes. Diversity of HLA in human population is one aspect of disease defense, and therefore the chance of two unrelated individuals having identical HLA molecules on all loci is very low.

Interleukin 10 (IL-10) is an important immunoregulatory cytokine in humans (Eskdale et al, 1998). It is involved in the regulation of inflammatory responses and in the pathology of human auto-immune disease. IL-10 suppresses T-cell immune responses,

which are the defense 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 disease (Eskdale et al., 1998). In this dissertation the focus will be on the IL-10 genotypes that are present on the -592 and -1082 loci. The genotypes on the -592 loci includes: AA, CA and CC. The genotypes present on the -1082 loci includes: AA, AG and GG.

1.5 Cohort Studies

A cohort study is a longitudinal (prospective) study in contrast to a cross-sectional one. In medical science, studies are often designed to investigate changes in the outcome of interest which is measured repeatedly over time in the participating subjects (Verbeke et al., 1998).

Advantages of a cohort or prospective study include:

1) Multiple diseases and outcomes can be observed at the same time during the study.

2) Risk and relative risk, odds and odds ratio can be calculated depending on the objectives of the study. These are all measures of disease association of a study disease and exposure to a risk factor(s).

3) It is more accurate than a case-control study since researchers take the data themselves rather than relying on records or the subjects recall information.

Disadvantages of a cohort or prospective study include:

1) A large sample size is needed (exposed and unexposed samples) compared to a case-control study.

2) Time consuming (individuals are usually followed over a long period of time).

3) It is expensive because of time and lengthy follow up period.

The current study cannot be classified as a cohort study in the strict sense of the word but rather an observational type cohort study. Information obtained from such a study can still be useful in assessing association between a study disease and a risk factor or exposure. The strict requirement of a specific sample size is not a pre-requisite.

1.6 Longitudinal Data

Longitudinal data is defined as data which is collected on subjects who are measured repeatedly over time (followed on two or more occasions), regarding a specific outcome or outcomes of interest in investigation. In this study, experimental units are repeatedly observed and measurements made over time. Experimental units can include patients or subjects in a clinical trial, or plots in agricultural experiments and other generalizations. Subjects are regarded as a random sample from a bigger population and hence, any effects that are not constant for all subjects are regarded as random. Such longitudinal studies are useful in determining individual changes over time, as well as the study of factors likely to influence change (Verbeke et al., 1998).

Longitudinal data analysis can be based on a balanced or unbalanced design. The former occurs when every individual has an equal number of observations taken at the same time points and it is assumed that there are no missing observations (Jones, 1993). Analysis based on balanced designs can be performed using classical multivariate analysis of variance methods (Verbeke et al., 1998). The balanced model assumptions are usually violated in observational studies, since the data is often unbalanced due to factors that cannot be controlled by the researcher, such as individuals entering and withdrawing from the study at different times (Laird & Ware, 1982; Verbeke et al., 1998). This results in individuals being observed a different number of times and the intervals between observations may also differ (Laird & Ware, 1982; Verbeke et al., 1998). The resulting unbalanced data sets cannot be analyzed using these methods applicable to balanced designs. Therefore, the need for extensions such as Linear Mixed Models (Laird & Ware, 1982), Generalized Estimating Equations and their extensions

(Liang & Zeger, 1986), and other techniques that may become necessary depending on the complexity of the data pattern.

Longitudinal data analysis can be advantageous in comparison to a cross-sectional study because the analysis can

(1) Increase the precision by allowing for within subject information
(2) Observe the change or evolution of a process over time
(3) Be able to separate age and cohort effects
(4) Account for individual to individual heterogeneity, a task which is not possible under cross-sectional studies
(5) Under such a study both time dependent and baseline covariates can be accounted for simultaneously

Measurements made on the same subjects are likely to be more similar than measurements made on different subjects. That is, repeated measurements are correlated. Therefore subject to subject variability and within subject correlation needs to be accounted for in the analysis. Overall there is within subject correlation and between subject variability, both contained in the data structure. Thus longitudinal data possesses a more informative structure compared to a cross-sectional data for the same problem.

1.7 Missing Data

Missing data is a frequent complication of any real-world study and can cause bias or lead to inefficient analyses. Missing observations is one of the many complications of analyzing longitudinal data and can result in subjects being measured a different number of times (Jones, 1993). Missing data may be due to design, chance or unforeseen circumstances. Missing data by design may include some variables not being collected or measured from all the subjects, subjects refusing to provide certain data as well as subjects or investigators leaving out information to insure confidentiality. Missing data can be classified as missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). MCAR refers to missing data that is completely

independent of the response. MAR is less severe, but missing data may depend on the response, however only on the observed responses. MNAR refers to missing data that is dependent on the unobserved responses (Molenberghs & Kenward, 2007). Missing data is a problem frequently encountered in longitudinal studies, and therefore needs to be taken into account during the analysis of longitudinal data.

1.8 Objectives of the Study

According to Kipiela et al. (2004) there is enough evidence to justify the importance of investigating the involvement of HLA-B in influencing HIV disease outcome and progression. It was found that the rate of disease progression is strongly associated with certain HLA-B, but not HLA-A allele expression ($P < 0.0001$ and $P = 0.91$ respectively). More specifically, it was found that there was a significant association between B*57 and B*5801 with low viral load (non-progression), and between B*18 and B*5802 with high viral load (progression). The paper by Kipiela et al. (2004) suggests the dominant role the HLA-B alleles play in the successful or unsuccessful immune containment of HIV infection and in the control of human pathogens.

According to Shin et al. (2000), individuals carrying the IL-10 -592A promoter allele could possibly have a higher risk of HIV infection than the IL-10 -592 CC genotype. These individuals progress to AIDS more rapidly after infection, especially in the later stages. It is suggested that IL-10 -592A facilitates HIV replication in vivo, and thus accelerates the progression to AIDS (Shin et al., 2000). Long-term non-progressors are classified as individuals who avoid clinical AIDS for 10 or more years after HIV infection (Shin et al., 2000).

1.8.1 Specific Objective of the Study

(1) To understand and apply methods relevant in modelling longitudinal data using a range of statistical techniques
(2) To understand the disease pathogenesis using biomarkers such as CD4+ counts and viral loads
(3) To understand the role of immuno genetic parameters HLA-B types and IL-10 genotypes in disease progression
(4) To investigate the interaction between HLA-B types and IL-10 genotypes

Chapter 2

Exploratory Data Analysis

2.1. Data Description

The Sinikithemba Study (SK Study) was initiated by the HIV Pathogenesis Program (HPP) in 2005 at the Nelson Mandela Medical School in the University of Kwa-Zulu Natal, Durban. This cohort enrolled a sample of 450 HIV positive individuals, consisting of 93 (20.67%) males and 357 (79.33%) females, some of whom were recruited from a previous study. Baseline characteristics such as demographics, HLA typing, cellular immunology, CD4+ count and viral load were collected. Participants were then seen every three months after enrolment. CD4+ counts were taken at every visit (every three months) and viral loads taken every second visit (every six months). Other information, such as anti-retroviral treatment (ART) status and TB status, was continually updated throughout the study. For the purpose of this research, the focus pertains to participants while they were ART naive.

Study participants were offered ongoing clinical care and monitoring. This was inclusive of regular feedback on the status of their HIV infection through the monitoring of their CD4+ levels and viral loads. Free counseling and support from dedicated and experienced doctors, nursing staff and counselors was also made available to all study participants. Once sufficient evidence to warrant treatment was gathered, participants were referred to the government sector medical care centers for initiation of antiretroviral therapy according to the national guidelines given by the South African government.

2.2 Study Design

The data that will be used in this research can be classified as both a longitudinal study and an observational cohort study, as measurements were repeatedly conducted on each individual over time at regular, but not necessarily equal, intervals as time elapsed. As already stated, such studies may lead to unbalanced data which will need to be accounted for during the modelling process and will henceforth be discussed in subsequent chapters.

2.3 Preliminary Plots and Descriptive Analysis

2.3.1 CD4+ Count

After excluding participants who did not have information on HLA-B typing or IL-10 genotyping, and excluding information of an individual after transferring onto treatment, the data set used includes 426 participants with a total of 4016 CD4+ count observations between August 2003 and January 2010. Figure 3 shows CD4+ counts over time in days using observations for all individuals. From this graph it is evident that there is a clear decrease in CD4+ count over time. There is however an outlier whose CD4+ count increases over time.

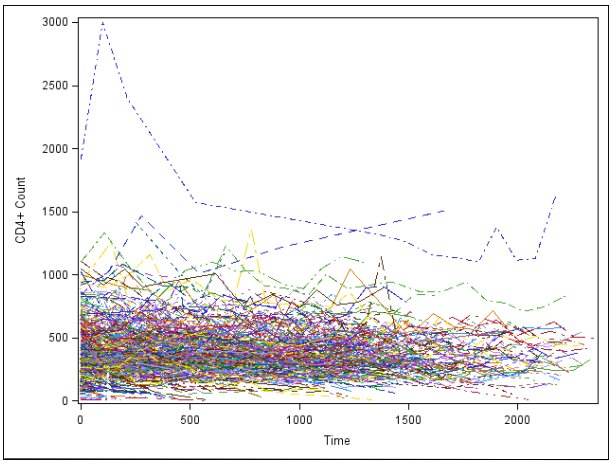


Figure 3: CD4+ counts of all the individuals over time in days

Figure 4 shows CD4+ counts over time for males and females. Although it is clear that there are less males then females, their CD4+ counts show the same decreasing pattern. Outlying patients can be seen in both male and female plots. The data also shows evidence of individual to individual variability over time.

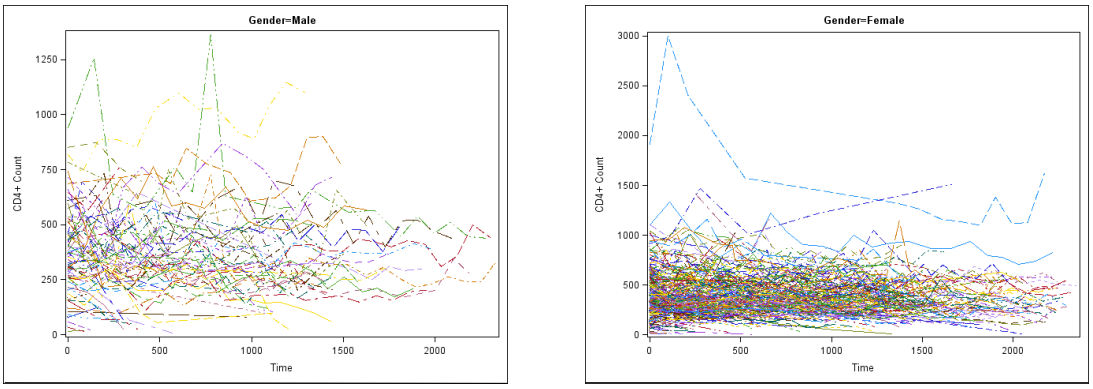


Figure 4: CD4+ counts for males and females over time in days

In Figure 5 where mean CD4+ counts for males and females are plotted, it is evident that there is a slight difference in the mean CD4+ count at each time point between males and females; however the trend is the same. Males seem to have a slightly lower mean CD4+ count than females in most of the time points except towards time points 23 and 24. The last time points (time points 24 and 25) show an increase in CD4+ count. These are however less precise than the early time points, since there are fewer observations toward the end due to dropout and death.

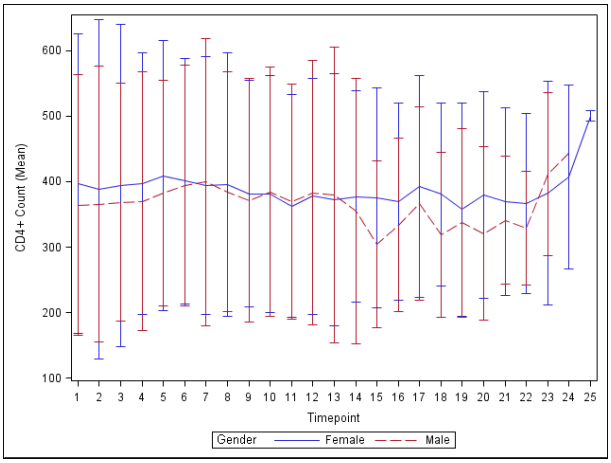


Figure 5: Mean CD4+ counts for males and females at each time point with standard deviations

A summary of the repeated measurement for CD4+ counts can be found in Table 1. It can be seen that the highest percentage of participants have only one observation for CD4+ count (11.50%). The lowest percentage of participants has twenty five CD4+ count observations (0.47%). This can be explained by the fact that by this measurement value most participants have already been transferred onto treatment, withdrawn from the study or succumbed to death. This can be seen graphically in Figure 6.

Table 1: Summary of repeated measurements for CD4+ counts

Number of CD4+ Observations per Participants	Number of Participants	% of Participants (Cumulative %)	Total number of CD4+ count observations (%)
1	49	11.5 (11.50)	49 (1.21)
2	39	9.15 (20.65)	78 (1.92)
3	29	6.81 (27.46)	87 (2.14)
4	37	8.69 (36.15)	148 (3.64)
5	20	4.69 (40.84)	100 (2.46)
6	11	2.58 (43.42)	66 (1.63)
7	13	3.05 (46.47)	91 (2.24)
8	12	2.82 (49.29)	96 (2.36)
9	15	3.52 (52.81)	135 (3.32)
10	6	1.41 (54.22)	60 (1.48)
11	7	1.64 (55.86)	77 (1.90)
12	19	4.46 (60.32)	228 (5.61)
13	22	5.16 (65.48)	286 (7.04)
14	25	5.87 (71.35)	350 (8.62)
15	28	6.57 (77.92)	420 (10.34)
16	26	6.10 (84.02)	416 (10.24)
17	11	2.58 (86.60)	187 (4.60)
18	12	2.82 (89.42)	216 (5.32)
19	4	0.94 (90.36)	76 (1.87)
20	12	2.82 (93.18)	240 (5.91)
21	6	1.41 (94.59)	126 (3.10)
22	9	2.11 (96.70)	198 (4.88)
23	7	1.64 (98.34)	161 (3.96)
24	5	1.17 (99.51)	120 (2.95)
25	2	0.47 (99.98)*	50 (1.23)

*Does not equal 100% due to rounding off

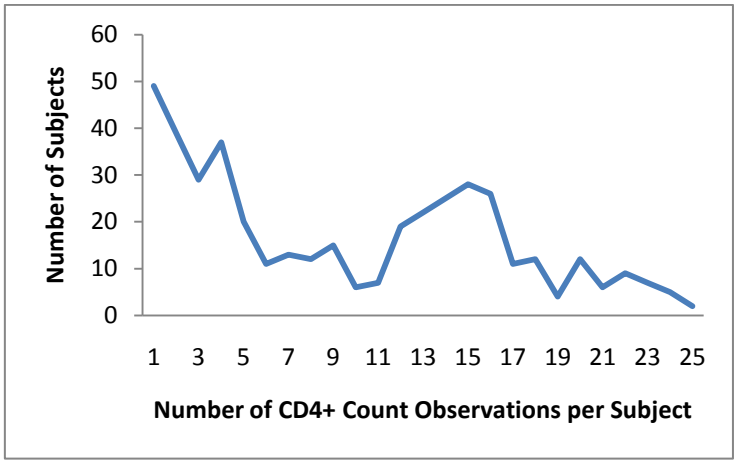


Figure 6: Number of CD4+ count observations

Tests for normality were performed on the CD4+ count data. Although results are not shown, it was seen that the CD4+ count data violates the normality conditions and therefore some kind of transformations will have to be made on the data. This is confirmed by the histogram in Figure 7.

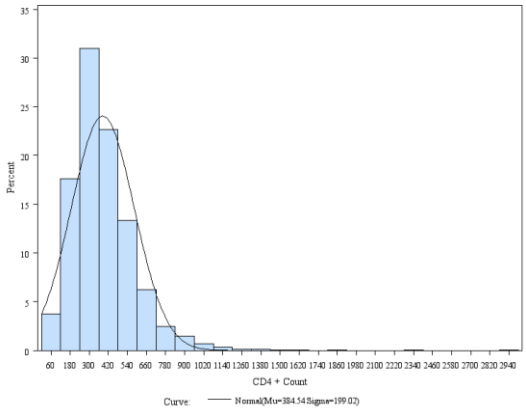


Figure 7: Histogram for CD4+ count

In an attempt to improve the assumption of normality, a log transformation was applied to the CD4+ count data. This led to an improvement as compared to the original CD4+ count data; however, the log CD4+ count data still violates the normality conditions, and therefore alternative transformations need to be conducted. This is confirmed by the histogram in Figure 8.

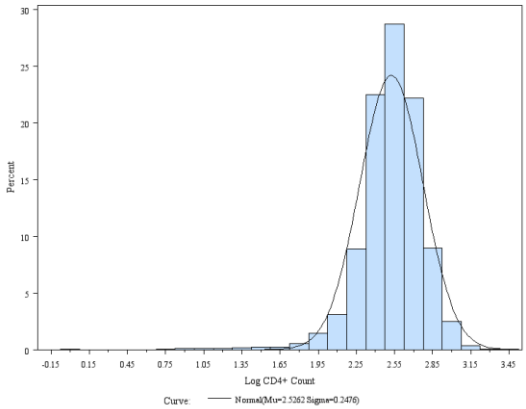


Figure 8: Histogram for log CD4+ count

By taking the square root of the CD4+ count data, the normality conditions showed a notable improvement over the log CD4+ count and the raw CD4+ count data. Although the normality test still shows significance as seen in Table 2, this may be due to the test detecting the longer than normal tails. The skewness and kurtosis are only slightly different to that of a normal distribution as seen in Table 3. The histogram in Figure 9 shows that this is approximately normally distributed. Therefore the square root transformation has more normality features than the log transformation and will be used in further analysis. Descriptive statistics for the square root CD4+ count can be seen in Table 3.

Table 2: Test for normality for square root CD4+ count

Test	Statistic	P-Value
Kolmogorov-Smirnov	0.038738	<0.0100
Cramer-von Mises	1.996975	<0.0050
Anderson-Darling	12.22675	<0.0050

Table 3: Summary descriptive statistics for square root CD4+ count

Mean	19.00
Median	18.68
Mode	16.76
Standard Deviation	4.84
Variance	23.48
Range	53.75
Interquartile Range	5.92
Skewness	0.41
Kurtosis	2.02

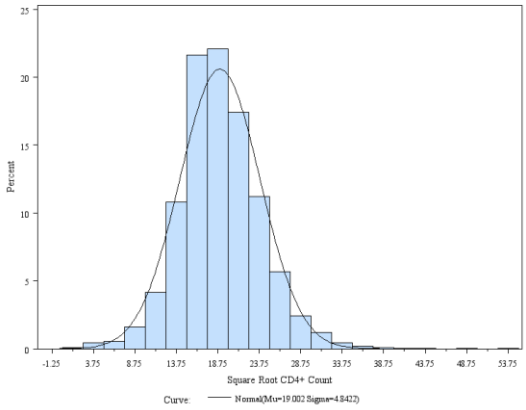


Figure 9: Histogram for square root CD4+ count

It is however premature to check the normality assumptions before accounting for measured or observed covariates and any other variation that can be accounted for, since normality is a conditional property on the outcome.

2.3.2 Viral Load

After excluding participants who did not have information on HLA-B typing or IL-10 genotyping, and individual information after they switched onto treatment, the data set used includes 426 participants with a total of 2007 viral load observations between August 2003 and January 2010. Figure 10 shows log viral loads over time for all individuals. From this graph it appears that log viral load remains constant over time. In

Figure 11 it can be seen that log viral loads for males and females remain constant over time.

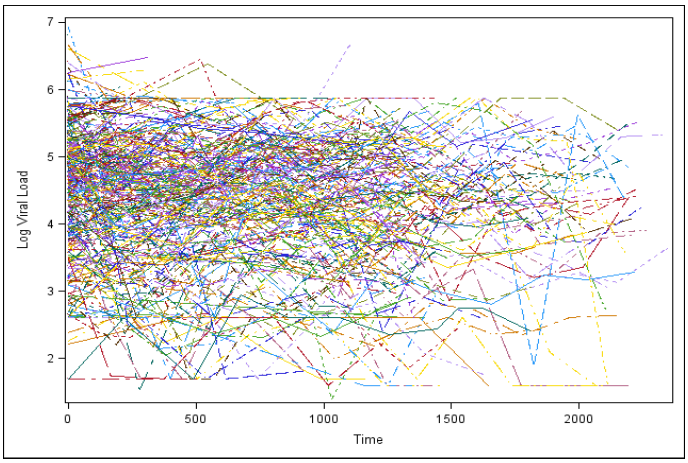


Figure 10: Data on the log viral loads of participants over time in days

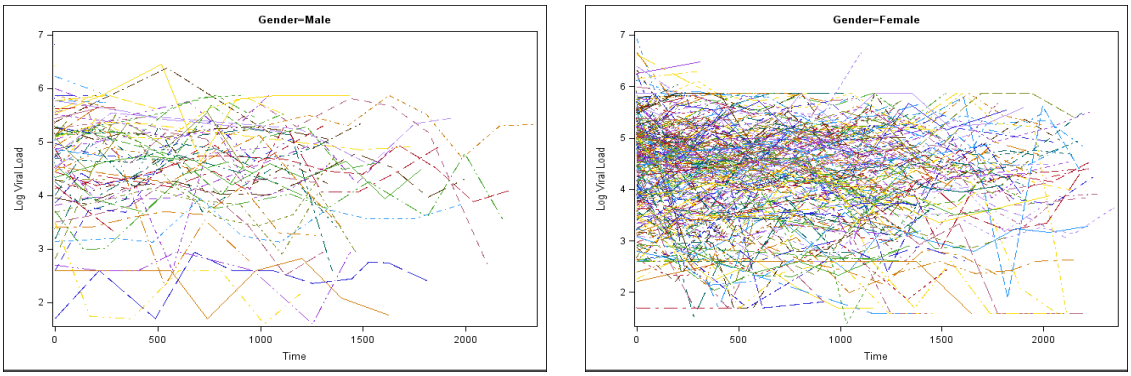


Figure 11: Log viral loads for males and females over time in days

It is evident from Figure 12 that the mean viral loads are different between males and females. It appears that females have a lower viral load at the majority of the time points. This implies that females are at an advantage over males in the sense that they have lower levels of the virus than males.

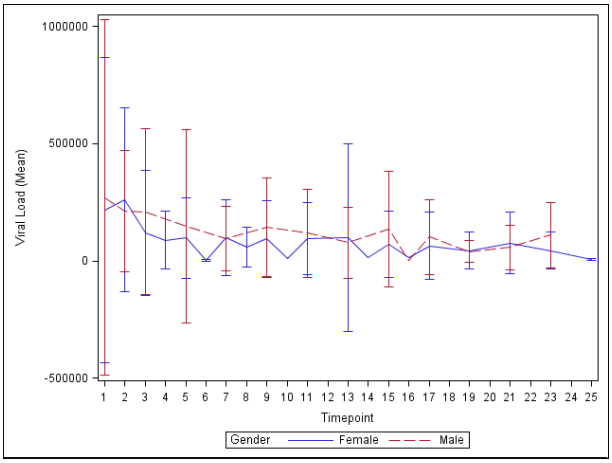


Figure 12: Mean viral loads for males and females at each time point with standard deviations

A summary of the repeated measurement for viral loads can be found in Table 4. It can be seen that the highest percentage of participants have only one observation for viral load as expected because initially all the participants were present. The lowest number of participants has fifteen viral load observations. This can be accounted for by participants being excluded after going onto treatment, withdrawing from the study or death. This can be seen graphically in Figure 13.

Table 4: Summary of repeated measurements for viral load

Number of viral load observations per subject	Number of subjects	% of subjects (cumulative %)	Total number of viral load observations (%)
0	2	0.47 (0.47)	0 (0.00)
1	80	18.78 (19.25)	80 (3.68)
2	65	15.26 (34.51)	13 (5.99)
3	33	7.75 (42.25)	99 (4.56)
4	29	6.81 (49.06)	116 (5.34)
5	20	4.69 (53.76)	100 (4.60)
6	30	7.04 (60.8)	180 (8.29)
7	43	10.08 (70.89)	301 (13.86)
8	58	13.62 (84.51)	464 (21.36)
9	18	4.23 (88.73)	162 (7.46)
10	16	3.76 (92.49)	16 (7.37)
11	12	2.82 (95.31)	132 (6.08)
12	13	3.05 (98.36)	156 (7.18)
13	6	1.14 (99.77)	78 (3.59)
14	1	0.23 (100.00)	14 (0.64)

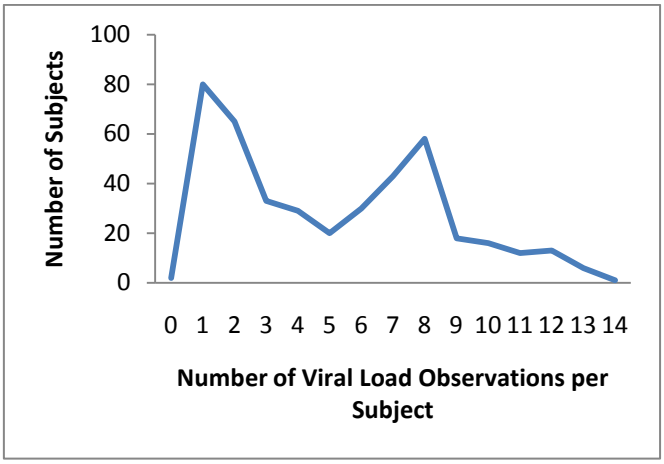


Figure 13: Number of viral load observations

Tests for normality indicate that the viral load data violates the normality conditions as shown by the results with p-values < 0.05 (results not shown). This can also be seen clearly by the histogram in Figure 14. Therefore transformations will need to be made.

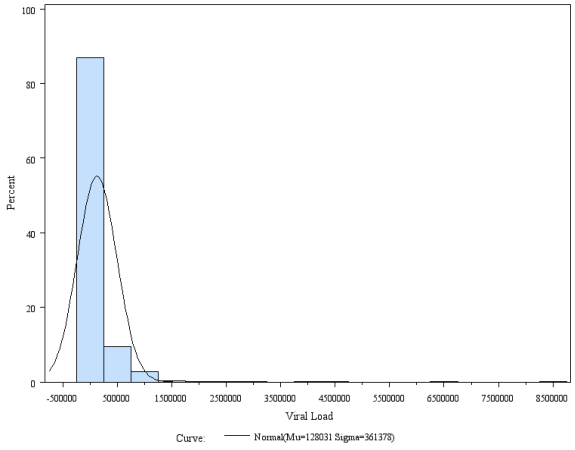


Figure 14: Histogram for viral load

By taking the log of the viral load, the normality conditions showed an improvement over viral load. Although the normality conditions are still violated as seen in Table 5, this may be due to the test detecting the longer than normal tails. Otherwise log transformation greatly improves adherence to normality. Descriptive statistics for the log viral load can be seen in Table 6. It can be seen that the log transformation introduces a slight negative skewness. The kurtosis in Table 6 shows that this data is approximately normally distributed. The histogram in Figure 15 shows that this is approximately normally distributed. Therefore the log viral load is acceptable and will be used in further analysis.

Table 5: Test for normality for log viral load

Test	Statistic	P-Value
Kolmogorov-Smirnov	0.0603	<0.0100
Cramer-von Mises	2.4901	<0.0050
Anderson-Darling	16.0202	<0.0050

Table 6: Location for log viral load

Mean	4.3431
Median	4.4726
Mode	2.6010
Standard Deviation	0.9872
Variance	0.9746
Range	5.4114
Interquartile Range	1.3194
Skewness	-0.5522
Kurtosis	-0.0872

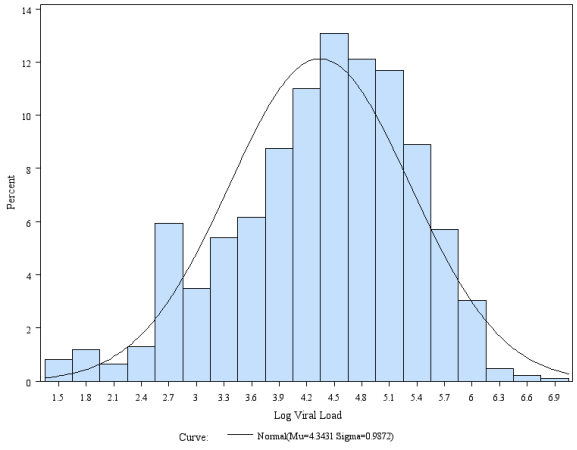


Figure 15: Histogram for log viral load

2.3.3 HLA-B Types

The most common HLA-B types found in this data set include HLA-B*0801, B*1503, B*1510, B*4201, B*4403 and B*5802. The distribution of these HLA-B types can be seen in Figure 16.

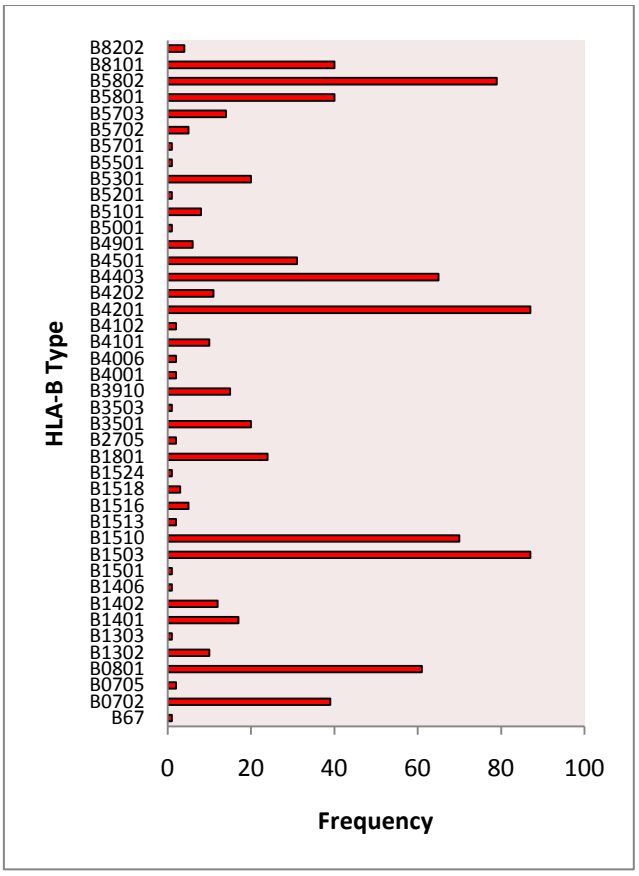


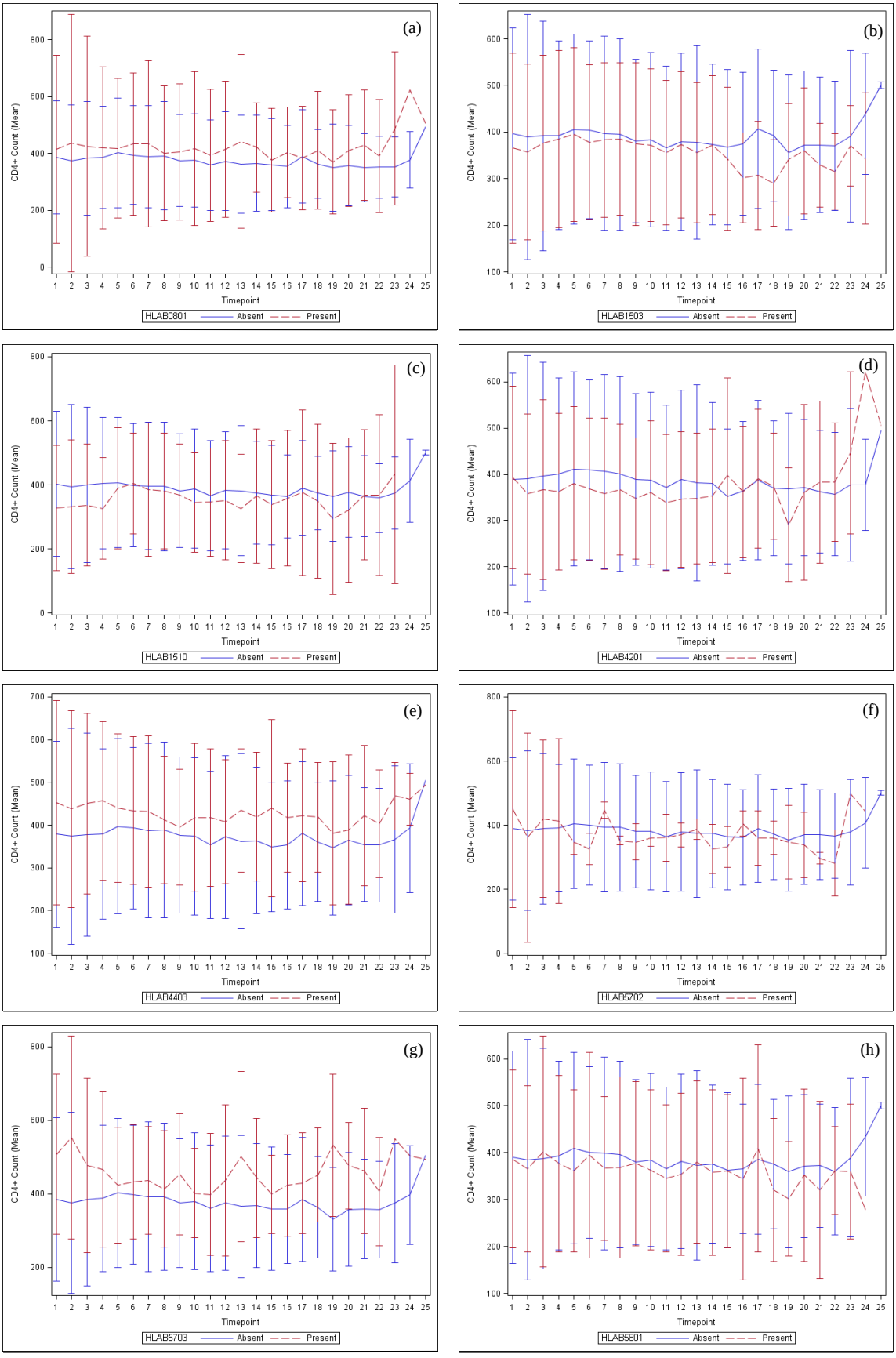
Figure 16: Frequency of each HLA-B type

Previous research on HLA-B types by Kiepiela et al. (2004) was based on cross-sectional data. It has been shown that HLA-B*57, B*5801 and B*4403 are associated with low viral load, and therefore are categorised as controllers. HLA-B*1510 and B*5802 were shown to be associated with high viral load, and hence viewed as facilitators for faster disease progression (Kiepiela, et al., 2004).

These HLA-B types will be investigated in the current longitudinal data analysis, as well as the most common HLA-B types found in Figure 16. Therefore the HLA-B types that will be included in the analysis are HLA-B*0801, B*1503, B*1510, B*4201, B*4403, B*5702, B*5703, B*5801, B*5802 and B*8100.

Plots seen in Figure 17 are used to graphically show the associations between CD4+ count and the HLA-B types. Since the mean CD4+ count for individuals that have these HLA-B types are above that of the overall mean CD4+ count for those that do not have these HLA-B types, it appears that B*0801, B*4403, B*5703 and B*8100 are associated with high CD4+ count, and therefore may be controllers. HLA-B*1503, B*1510, and B*5802 seem to be associated with low CD4+ count since the mean CD4+ counts for individuals with these HLA-B types all fall below that of the mean CD4+ count for those individuals who do not have these HLA-B types. They may therefore be considered as facilitators for faster disease progression. The mean CD4+ count for individuals with HLA-B*4201, B*5702 and B*5801 are very similar to the mean CD4+ count of those individuals without these HLA-B types and do not seem to be associated with either high or low CD4+ count. The effect of the presence and absence of these immune characteristics can be seen in the graphs shown in Figure 17 (a) - (j).

In Figure 18 (a) – (j), it is shown that individuals with HLA-B*4403 and B*5801 have lower mean viral load than the individuals without these HLA-B types. This indicates that these HLA-B types may be associated with a slower disease progression. From the mean viral loads of the individuals that contain the HLA-B types, HLA- B*1503, B*1510, B*5702 and B*5802 they appear to be associated with high viral load. This implies that these HLA-B types are associated with a faster replication of the virus. HLA-B*0801, B*4201, B*5703 and B*8100 seem to have similar viral loads for the individuals with or without these HLA-B types.


(a)
CD4+ Count (Mean)
Timepoint
HLAB0801 Absent Present
(b)
CD4+ Count (Mean)
Timepoint
HLAB1503 Absent Present
(c)
CD4+ Count (Mean)
Timepoint
HLAB1510 Absent Present
(d)
CD4+ Count (Mean)
Timepoint
HLAB4201 Absent Present
(e)
CD4+ Count (Mean)
Timepoint
HLAB4403 Absent Present
(f)
CD4+ Count (Mean)
Timepoint
HLAB5702 Absent Present
(g)
CD4+ Count (Mean)
Timepoint
HLAB5703 Absent Present
(h)
CD4+ Count (Mean)
Timepoint
HLAB5801 Absent Present

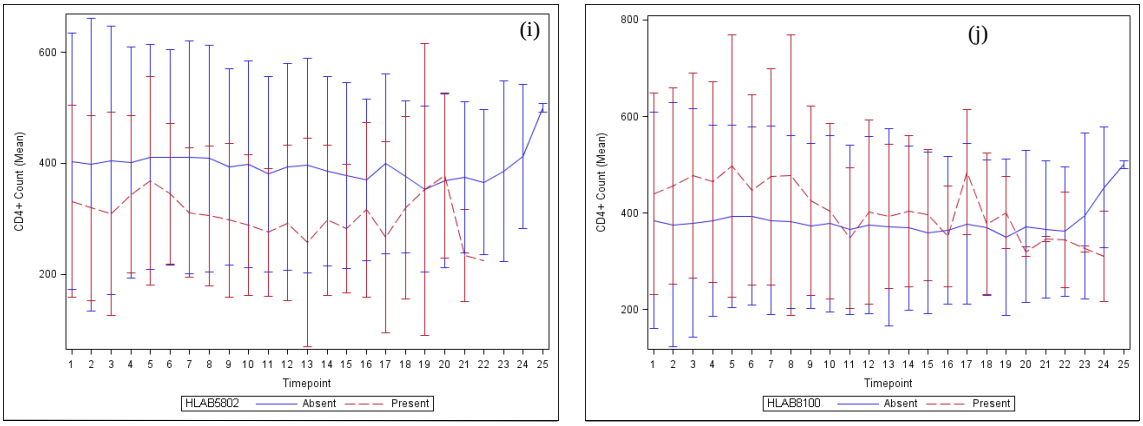
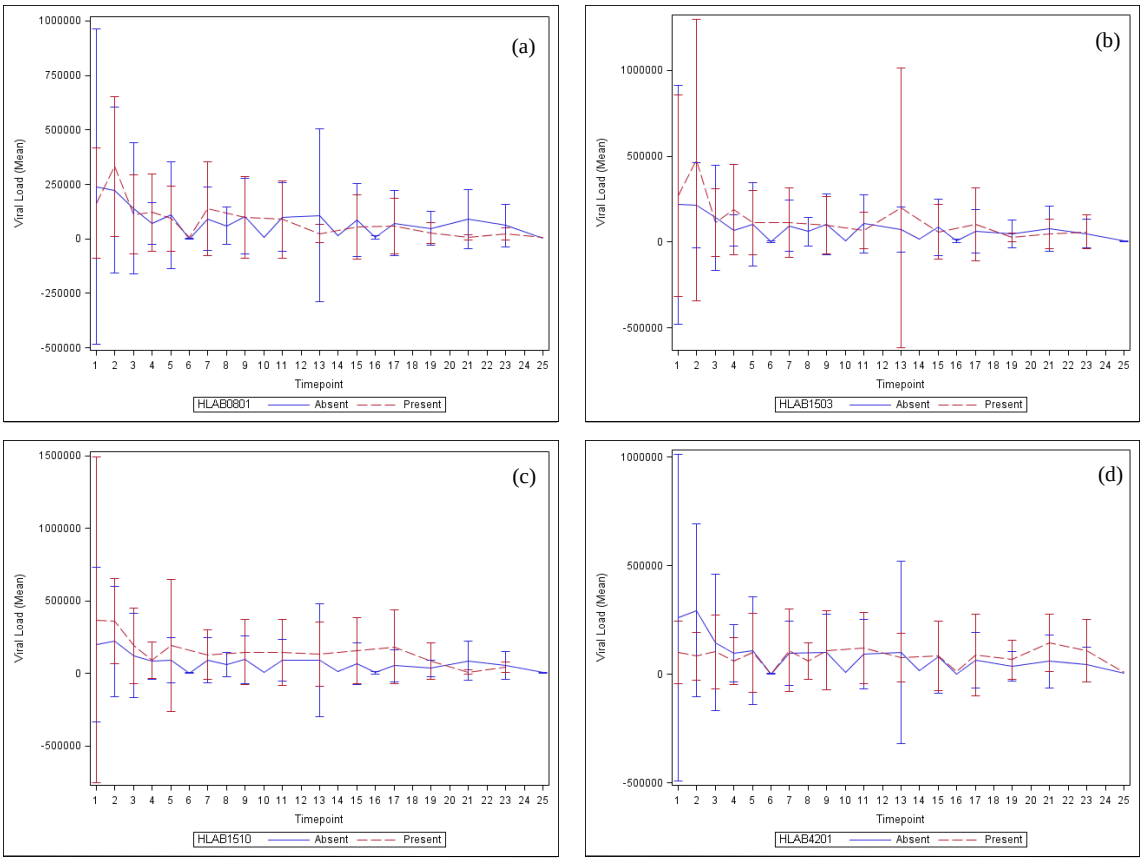


Figure 17: CD4+ count means by HLA-B types with standard deviations



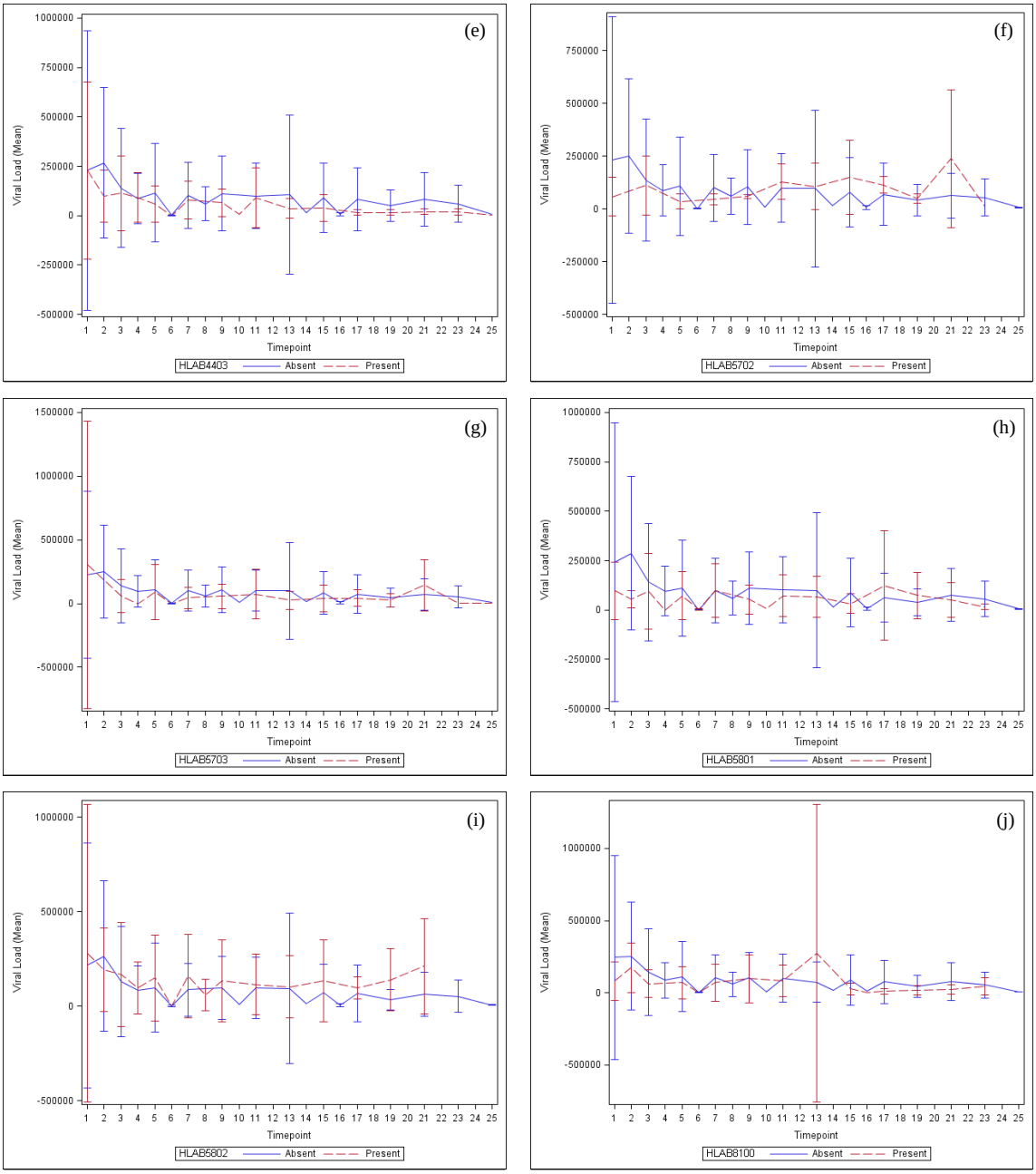


Figure 18: Viral load means by HLA-B types with standard deviations

2.3.4 IL-10 Genotypes

Out of the 426 participants, it is shown in Figure 19 that 45 had the "AA" -592 genotype, 198 had the "AC" -592 genotype and 183 had the "CC" -592 genotype. It can be seen in Figure 20 that 205 had the "AA" -1082 genotype, 173 had the "AG" -1082 genotype, and 48 had the "GG" -1082 genotype.

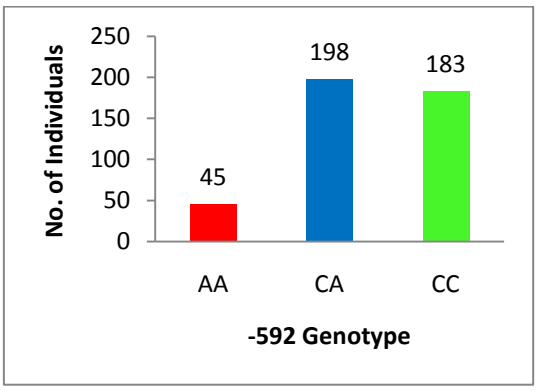


Figure 19: Number of individuals with each of the -592 genotypes

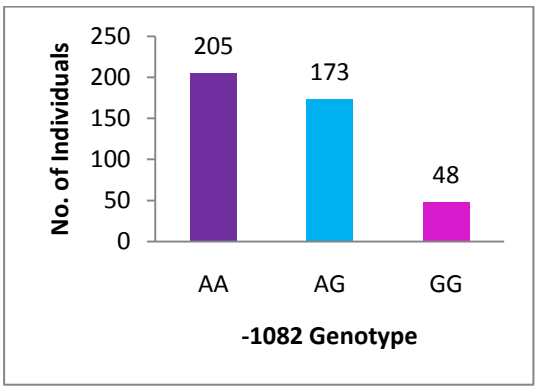


Figure 20: Number of individuals with each of the -1082 genotypes

Figure 21 illustrates the association between the median CD4+ counts and the -592 genotypes. The p-values for the differences are shown over horizontal bars. Using the Kruskal Wallis Test, it is seen that there is at least one significant difference in the median CD4+ counts between the -592 genotypes (p-value=0.0428). Conducting a pairwise comparison, it is evident that the CA and CC genotypes have significantly different median CD4+ counts (p-value=0.0179). Figure 22 highlights the association between median CD4+ count and the -1082 genotypes. There is at least one significant difference in the median CD4+ counts among the -1082 genotypes. The pairwise comparison reveals that the AA and AG genotypes (p-value=0.0093), the AA and GG genotypes (p-value<0.0001) and the AG and GG genotypes (p-value=0.0008) have significantly different median CD4+ counts. These results are before correcting for multiple comparisons. To solve the problem of multiple testing, the Bonferroni correction can be used. This would imply that a significance level of $\alpha = 0.05/3$ is used to test the differences between groups.

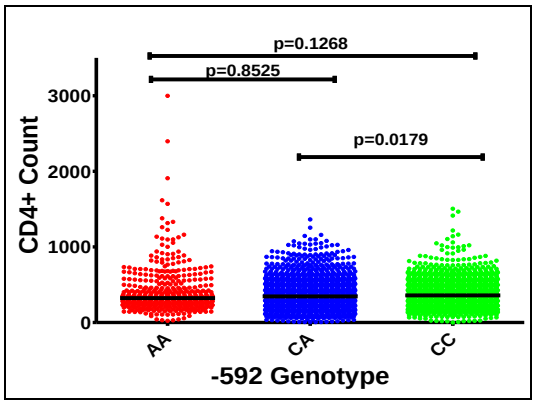


Figure 21: CD4+ counts for -592 genotypes

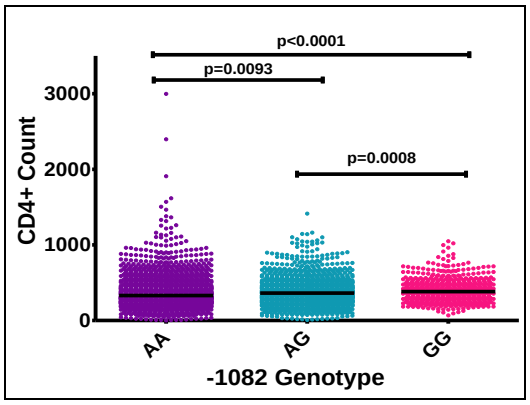


Figure 22: CD4+ counts for -1082 genotypes

Figure 23 reveals the association between log viral load and the -592 genotypes. Using the Kruskal Wallis Test, it is evident that there is at least one significant difference between log viral loads for the -592 genotypes (p-value=0.0038). The AA and CC genotypes (p-value=0.0094) and the CA and CC genotypes (p-value=0.0048) have significantly different median log viral loads. Figure 24 shows the association between log viral load and the -1082 genotypes. There is at least one significant difference in median log viral loads among the -1082 genotypes (p-value=0.0002). After pairwise comparisons, it is evident that the AA and GG genotypes (p-value<0.0001) and the AG and GG genotypes (p-value=0.0002) have significantly different median log viral loads. Again, correcting for multiple comparisons is done by using the Bonferroni correction with a significance level of $\alpha = 0.05/3$.

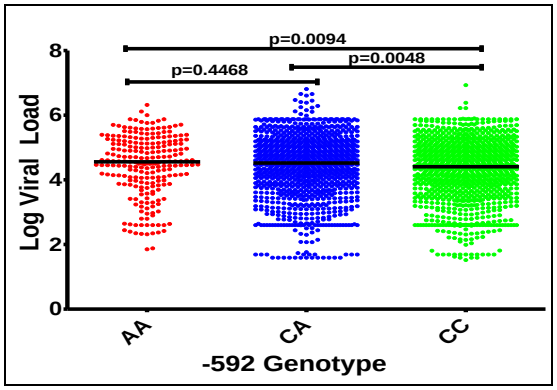


Figure 23: Log viral load for -592 genotypes

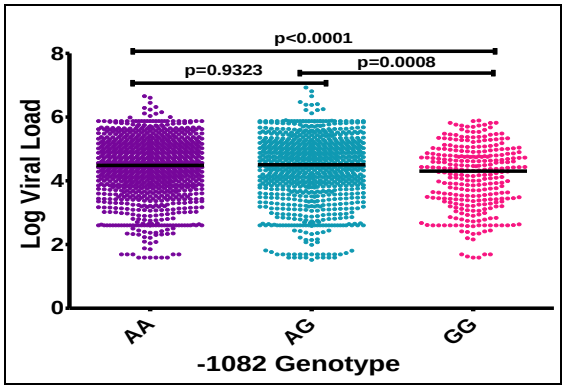


Figure 24: Log viral load for -1082 genotypes

The exploratory data analysis shows that there are complex relationships between the key disease markers, namely CD4+ counts and viral loads and the individual's specific genetic factors, namely the HLA-B types and IL-10 genotypes.

Chapter 3

Models for Longitudinal Data

In this chapter we review briefly the theory of longitudinal data analysis which is the most appropriate technique to use for the SK study data. We first introduce the model then we discuss how to estimate fixed and random effects in the model. Statistical inference about the model parameters is also discussed.

3.1 Linear Mixed Model

The linear mixed model (LMM) deals with continuous longitudinal data assumed to be normally distributed and can be written as

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

for i=1,2,…,N (Laird & Ware, 1982; Verbeke et al., 1998). Where Y_i is the response variable, X_i is matrix of fixed effects (design matrix and covariates), β is the fixed effects parameter, Z_i is the design matrix for random effects and U_i is the random effects parameter (Laird & Ware, 1982). According to Verbeke et al. (1998), we assume $U_i \sim N(0\ , G)$ and $\varepsilon_i \sim N(0\ , R_i)$ and we assume U_i and ε_i are independent, thus we can write

$$\mathrm{Y_i} = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{in_i} \end{bmatrix} \tag{3.2}$$

and

$$\begin{bmatrix} U_i \\ \varepsilon_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} G & 0 \\ 0 & R_i \end{bmatrix}\right). \tag{3.3}$$

Thus,

$$\mathrm{E}(\mathrm{Y_i}) = X_i\beta \tag{3.4}$$

and

$$Var(Y_i) = Z_i G Z_i' + R_i = V_i \tag{3.5}$$

Let the parameters contained in V_i be included in the vector α. The following demonstrates the estimation of the parameters discussed in the previous section.

Incomplete and unbalanced data are complications associated with longitudinal data. One method of dealing with this problem is to use Gaussian theory estimation procedures for the mixed model. This includes maximum likelihood (ML) and restricted maximum likelihood (REML) estimators.

The estimate for β and U can be shown to be

$$\widehat{\beta}\,(\alpha) = \left(\sum_{i=1}^{N}(X_i{}'W_iX_i)\right)^{-1}\left(\sum_{i=1}^{N}(X_i'W_iy_i)\right) \tag{3.6}$$

$$\widetilde{U} = E(U|Y) = GZ_i{}'PY_i \tag{3.7}$$

where $W_i = V_i^{-1}(\alpha)$ and $P = W_i - W_iX_i(X_i{}'W_iX_i)^{-1}X_i{}'W_i$ and α is the variance component (Verbeke & Molenberghs, 2000, p. 42).

The expression $\widehat{\beta}\,(\alpha)$ assumes that α is known otherwise an estimate for α may be used. The expression $\widetilde{U}$ in equation (3.7) is based on the conditional mean of U given data seen as a posterior mean of U given Y.

3.2 Maximum Likelihood Estimation (ML)

Let Y be a random variable such that $Y \sim N(\mu, \sigma^2)$. The maximum likelihood method uses the probability density function of the Normal distribution given by

$$f(Y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y-\mu)^2}{2\sigma^2}\right) \tag{3.8}$$

Suppose there are N normally, and independently distributed variables $Y_1; Y_2; \ldots; Y_N$, where each have a mean μ and variance σ^2. The basis of the inferential procedures is the likelihood function given by

$$L(\theta; Y) = \prod_{i=1}^{N} f(Y_i; \theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i-\mu)^2}{2\sigma^2}\right) \tag{3.9}$$

where $Y = (Y_1, \ldots, Y_N)'$ and $\epsilon\, \Omega\ \ \Theta = (\mu, \sigma^2)$ contains the parameters to be estimated and this is done by maximizing the likelihood. Maximizing the likelihood is equivalent to maximizing the log likelihood where the log likelihood can be written as

$$\log L(\theta) = l(\theta) = \sum_{i=1}^{N} \log f(Y; \theta) \quad = \sum_{i=1}^{N} \left(-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(Y_i-\mu)^2}{2\sigma^2}\right) \tag{3.10}$$

(Pawitan, 2001).

3.2.1 Extension to Multivariate Data

Let $Y_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})$, where y_{ij} is the j^{th} observation from the i^{th} individual or experimental unit in a longitudinal study. In this particular case the experimental unit is the individual, which constitutes a cluster of n_i observations. Assume a model for Y_i given by $Y_i \sim MVN(X_i'\beta, V_i)$. Then the likelihood formulation can be extended to this type of data as follows,

$$L_{ML}(\theta) = \prod_{i=1}^{N} (2\pi)^{-\frac{n_i}{2}} |V_i(\alpha)|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(Y_i - X_i\beta)' V_i^{-1}(\alpha)(Y_i - X_i\beta)\right) \tag{3.11}$$

(Verbeke & Molenberghs, 2000, p. 42).

Here there are two types of fixed parameters namely the fixed effect regression parameters β and the variance covariance parameters α. The model also contains random effects parameters U_i as shown in model equation (3.1).

3.2.2 Estimation of Fixed Effects Regression Parameters

The estimation of the fixed parameters requires maximization of the log likelihood function with respect for those parameters, β. This is done by differentiating the log-likelihood with respect to β and solving $\frac{\partial l}{\partial \beta} = 0$. Assuming α is known, we then have the following

$$\begin{aligned}\frac{\partial l}{\partial \beta} &= \frac{\partial}{\partial \beta}\left(-\frac{1}{2}(Y_i - X_i\beta)' V_i^{-1}(\alpha)(Y_i - X_i\beta)\right) \\ &= \frac{\partial}{\partial \beta}\left(-\frac{1}{2}(Y_i' V_i^{-1}(\alpha) Y_i - Y_i' V_i^{-1}(\alpha) X_i\beta - \beta' X_i' V_i^{-1}(\alpha) Y_i + \beta' X_i' V_i^{-1}(\alpha) X_i\beta)\right) \\ &= -(X_i' V_i^{-1}(\alpha) X_i\beta - X_i' V_i^{-1}(\alpha) Y_i) \end{aligned} \quad (3.12)$$

(Werner, 2009, p. 44).

We now equate $\frac{\partial l}{\partial \beta}$ to zero and solve for β as shown below.

$$\begin{aligned}&\frac{\partial l}{\partial \beta} = 0 \\ &X_i' V_i^{-1}(\alpha) X_i\beta - X_i' V_i^{-1}(\alpha) Y_i = 0 \\ &X_i' V_i^{-1}(\alpha) X_i\beta = X_i' V_i^{-1}(\alpha) Y_i \\ &\hat{\beta} = (X_i' V_i^{-1}(\alpha) X_i)^{-1}(X_i' V_i^{-1}(\alpha) Y_i) \end{aligned} \quad (3.13)$$

Combining the above information from N individuals will give the following estimate for β

$$\widehat{\beta}\,(\alpha) = \left(\sum_{i=1}^{N}(X_i{}'W_iX_i)\right)^{-1}\left(\sum_{i=1}^{N}(X_i{}'W_iY_i)\right) \tag{3.14}$$

(Laird & Ware, 1982, p. 966; Verbeke & Molenberghs, 2000, p. 42), where $W_i = V_i^{-1}$ which depends on α. If α is unknown then an estimate of α can be used.

3.2.3 Estimation of Random Effects Parameters

Consider the model for each individual i

$$Y_i = X_i\beta + Z_iU_i + \varepsilon_i \tag{3.18}$$

The estimation of random effects follows a similar process to that of the estimation of the fixed effects. If the $cov(U_i, Y_i) = GZ'$, where Y_i is the response vector and U_i is the vector of individual specific parameters, we then have

$$\begin{pmatrix} Y \\ U \end{pmatrix} \sim N\left(\begin{pmatrix} X\beta \\ 0 \end{pmatrix}, \begin{pmatrix} V & ZG \\ GZ' & G \end{pmatrix}\right) \tag{3.19}$$

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

$$\begin{aligned} \widehat{U_\iota} &= E(U_i|Y_i) \\ &= E(U_i) + cov(U_i|Y_i)\big(var(Y_i)\big)^{-1}\big(Y_i - E(Y_i)\big) \\ &= GZ_i'V_i^{-1}(\alpha)(Y_i - X_i\beta) \end{aligned} \tag{3.20}$$

It can easily be shown that

$$E(\widehat{U}_i) = 0 \tag{3.21}$$

and

$$var(\widehat{U}_\iota) = GZ_i{}' \left\{ W_i - W_i X_i \left(\sum_{i=1}^{n} X_i' W_i X_i \right)^{-1} X_i' W_i \right\} Z_i G' \tag{3.22}$$

.

If we assess the estimation error using equation (3.22), the variation in $\widehat{U}_\iota - U_i$ will be underestimated since this expression does not take into account the variation of U_i. The following may therefore be preferred

$$var(\widehat{U}_\iota - U_i) = G - GZ_i' W_i Z_i G + GZ_i' W_i X_i \left(\sum_{i=1}^{n} X_i' W_i X_i \right)^{-1} X_i' W_i Z_i G \tag{3.23}$$

(Laird & Ware, 1982, p. 966)**.**

3.2.4 Restricted Maximum Likelihood Estimation (REML)

Applying maximum likelihood to the linearly transformed response data vector is known as restricted maximum likelihood. The transformation is done in such a way that the linearly transformed data vector contains none of the fixed effects. The method of REML was introduced by Patterson and Thompson (1971). It was developed to avoid the biased variance component estimates that are produced by ordinary ML estimation. This is due to the fact that the maximum likelihood method of the variance components takes no account of the loss of degrees of freedom resulting from the estimation of the fixed effects (β). REML takes into account the loss of degrees of freedom from estimating the fixed effects and is therefore unbiased, while ML is biased (Verbeke & Molenberghs, 2000, p. 44). For this reason ML estimates of variance are biased downwards compared to REML estimates.

To demonstrate how REML works, consider a simple univariate random sample $Y_1, Y_2, \dots, Y_N \sim iid\ N(\mu, \sigma^2)$. Then all the data can be combined into one distributional model

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \vdots \\ \mu \end{pmatrix};\ \sigma^2 I_N \right) \tag{3.24}$$

where I_N is a N-dimensional identity matrix. To avoid estimating μ, the vector Y can be transformed such that μ is removed from the likelihood. Let Y^* be the transformation where $Y_i^* = Y_i - Y_{i+1}$, then

$$Y^* = \begin{pmatrix} Y_1 - Y_2 \\ Y_2 - Y_3 \\ \vdots \\ Y_{N-1} - Y_N \end{pmatrix} = A'Y \sim N(0, \sigma^2 A'A) \tag{3.25}$$

where A is a (N-1) × N matrix with elements $A_{i,i} = 1,\ A_{i,i+1} = -1$ and zero elsewhere (Verbeke & Molenberghs, 2000, p. 43). Using this transformation the REML of σ^2 is given by

$$s^2 = \sum_{i=1}^{N} \left(Y_i - \overline{Y}\right)^2 / (N-1) \tag{3.26}$$

which is unbiased for σ^2, while the ML estimate is $\left(1 - \frac{1}{N}\right) s^2 < s^2$, but for large N the difference will be negligible.

When dealing with longitudinal data, let Y_i denote the individual n_i-dimensional vector of observations given by $Y_i = \left(y_{i1}, y_{i2}, \dots, y_{in_i}\right)'$. We assume that $Y_i \sim N(X_i\beta, V_i)$. Combining all subject specific information into one vector *Y*, such that $Y \sim N(X\beta, V)$ gives

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix}, \qquad X = \begin{pmatrix} X_1 \\ \vdots \\ X_N \end{pmatrix}, \qquad V(\alpha) = \begin{pmatrix} V_1 & 0 & \cdots & 0 \\ 0 & V_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & V_N \end{pmatrix} \tag{3.27}$$

The REML adjusted likelihood for $\theta = (\alpha', \beta')$, is given by

$$L_{REML}(\theta) = \left| \sum_{i=1}^{N} X_i' W_i(\alpha) X_i \right|^{-\frac{1}{2}} L_{ML}(\theta) \tag{3.28}$$

(Verbeke & Molenberghs, 2000, p. 46), where $W_i(\alpha) = V_i(\alpha)^{-1}$.

The REML estimators can also be found by maximizing the adjusted likelihood seen in equation (3.28) with respect to α and β, since $\left|\sum_{i=1}^{N} X_i' W_i(\alpha) X_i\right|$ does not depend on β (Verbeke & Molenberghs, 2000, p. 46). $L_{REML}(\theta)$ can be seen as a modification to $L_{ML}(\theta)$ in equation (3.11) to include a penalty to account for degrees of freedom lost in estimating β.

3.2.5 Estimation of Unknown Variance Components

If the covariance matrices are unknown, but an estimate of fixed θ is available, hence estimates of R_i and G are also available, we then let the variance of Y_i given in equation (3.5) be estimated by $\widehat{V}_i = \hat{R}_i + Z_i \hat{G} Z_i' = \widehat{W}_i^{-1}$. From this we can then estimate α and U_i. To do this, we use the weighted least squares from equation (3.14), replacing each W_i by an estimate $\widehat{W}_i$. We denote these estimates by $\hat{\alpha}(\hat{\theta})$ and $\widehat{U}_i(\hat{\theta})$ (Laird & Ware, 1982, pp. 966-967; Verbeke & Molenberghs, 2000, p. 42).

3.3 The Random Intercept Model

Suppose time is the only covariate in the model such that a model for an outcome at a given time point for an individual i is given by the linear model,

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \varepsilon_{ij} \tag{3.29}$$

where i=1,2,…,N and Y = $\begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix}$.

Marginally

$$E(y_{ij}) = \beta_0 + \beta_1 t_{ij} \tag{3.30}$$

$$y_{ij} \sim N(\beta_0 + \beta_1 t_{ij}, \sigma^2) \tag{3.31}$$

The coefficients β_0 and β_1 are a measure of marginal or population averaged effects. In order to explain individual to individual variability one can add individual specific effects to the model above so that the intercepts and slopes are now written as

$$\beta_{0i} = \beta_0 + b_{0i} \tag{3.32}$$

$$\beta_{1i} = \beta_1 + b_{1i} \tag{3.33}$$

where

$$U_i = \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}; \begin{pmatrix} g_0^2 & g_{01} \\ g_{10} & g_1^2 \end{pmatrix} \right) \tag{3.34}$$

and $g_{01} = g_{10}$. Equations (3.32) and (3.33) suggest that now an individual possesses individual specific intercept and slope as opposed to the case in equation (3.30).

Then the final model is modified to:

$$\begin{aligned} y_{ij} &= \beta_0 + b_{0i} + (\beta_1 + b_{1i})t_{ij} + \varepsilon_{ij} \\ &= \beta_0 + \beta_1 t_{ij} + b_{0i} + b_{1i}t_{ij} + \varepsilon_{ij} \end{aligned} \tag{3.35}$$

Model (3.35) is thus called the random or subject specific random effects model.

If we consider a model with only a random intercept effect (no random slope), such a model is written as

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{0i} + \varepsilon_{ij} \tag{3.36}$$

$$E(y_{ij}) = \beta_0 + \beta_1 t_{ij} \tag{3.37}$$

$$var(y_{ij}) = g_0^2 + \sigma^2 \tag{3.38}$$

Note that in the case of the LMM for a normal response it is easy to switch from the subject specific random effects model to the marginal model by using expectation but this is not generally true for other non-normal responses.

Note b_{0i} and ε_{ij} are uncorrelated such that if an observation at time occasion k is given by

$$y_{ik} = \beta_0 + \beta_1 t_{ik} + b_{0i} + \varepsilon_{ik} \tag{3.39}$$

where ε_{ik} , ε_{ij} are uncorrelated then

$$var(y_{ij}) = var(y_{ik}) = g_0^2 + \sigma^2 \tag{3.40}$$

$$\begin{aligned} cov(y_{ij}, y_{ik}) &= E(y_{ij}y_{ik}) - E(y_{ij})E(y_{ik}) \\ &= E(b_{0i}^2) = g_0^2 \end{aligned} \tag{3.41}$$

$$corr(y_{ij}, y_{ik}) = \frac{cov\,(y_{ij}, y_{ik})}{\sqrt{var(y_{ij})var(yik)}} = \frac{g_0^2}{g_0^2 + \sigma^2} = \rho \tag{3.42}$$

Thus by allowing for subject to subject variability through a random intercept effect, this automatically induces a correlation between any two observations from the same subject. Random effect models for longitudinally measured or observed data were first described by Laird and Ware (1982). This means a model with a random intercept only leads to the exchangeable or compound symmetry correlation structure given as:

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

3.4 Random Intercept and Slope Model

The random intercept and slope model is just but a special case of the linear mixed model for longitudinal data which we now briefly describe below. There exists numerous literature on the analysis of longitudinal data accounting of for both fixed and random effects and among the most referenced include Laird and Ware (1982), Diggle et al. (2002) and Verbeke and Molenberghs (2000).

Now consider the full random intercept and slope model which can be written as

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{0i} + b_{1j} t_{ij} + \varepsilon_{ij} \tag{3.43}$$

or equivalently,

$$y_{ij} = x'_{ij}\beta + z'_{ij} b_i + \varepsilon_{ij} \tag{3.44}$$

where $x_{ij}' = \begin{pmatrix} 1 & t_{ij} \end{pmatrix}$, $z_{ij}' = \begin{pmatrix} 1 & t_{ij} \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$ and $b_i = \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix}$,

then

$$E(y_{ij}) = \beta_0 + \beta_1 t_{ij}, \tag{3.45}$$

$$var(y_{ij}) = g_0^2 + g_1^2 t_{ij}^2 + 2t_{ij} g_{01} + \sigma^2 \tag{3.46}$$

and

$$cov(y_{ij}, y_{ik}) = cov(\beta_0 + \beta_1 t_{ij} + b_{0i} + b_{i1} t_{ij} + \varepsilon_{ij}, \beta_0 + \beta_1 t_{ik} + b_{0i} + b_{1i} t_{ik} + \varepsilon_{ik}) \tag{3.47}$$

which we can expand to

$$\begin{aligned} cov(y_{ij}, y_{ik}) &= E(y_{ij} y_{ik}) - E(y_{ij}) E(y_{ik}) \\ &= g_0^2 + t_{ij} t_{ik} g_1^2 + g_{01} t_{ik} + g_{10} t_{ij} \\ &= g_0^2 + t_{ij} t_{ik} g_1^2 + g_{01}(t_{ik} + t_{ij}) \end{aligned} \tag{3.48}$$

$$corr(y_{ij}, y_{ik}) = \frac{g_0^2 + t_{ij} t_{ik} g_1^2 + g_{01}(t_{ij} + t_{ik})}{\sqrt{var(y_{ij}) var(y_{ik})}} \tag{3.49}$$

While the covariance in equation (3.41) is constant that specified in equation (3.48) is a second order quadratic function of time. The correlation between any two observations is no longer constant, but is now a function of time.

3.5 Types of Correlation or Covariance Structures

Although the exact correlation between any two observations in a random intercept and slope model is as given in equation (3.49), other simplifying assumptions can be used. Since the repeated measurements are correlated, the covariance among the repeated measures needs to be accounted for. The following covariance structures can be used to account for both (1) correlation between subjects and (2) correlation between random effects depending on the focus of the analysis.

Commonly used covariance structures include Compound Symmetry (CS), Toeplitz (Toep), Unstructured (UN), Autoregressive (1) (AR(1)), Power Spatial (SP(POW)(c-list)), Exponential Spatial (SP(EXP)(c-list)) and Gaussian Spatial (SP(GAU)(c-list)) as listed in the table below. Some of the common covariance structures assumed in practice are listed below. In Table 7, m denotes the number of observations per individual.

Table 7: Commonly assumed covariance structures

Structure	Description	No. of Parameters	{i,j}th element
AR(1)	Autoregressive(1)	2	$\sigma_{ij} = \sigma^2\rho^{\lvert i-j\rvert}$
CS	Compound Symmetry	2	$\sigma_{ij} = \sigma_1 + \sigma^2 1(i=j)$ $where\ 1(i=j)=1$ $if\ i=j\ and\ 0\ if\ i \neq j$
UN	Unstructured	m(m+1)/2	$\sigma_{ij} = \sigma_{ij}$
TOEP	Toeplitz	m	$\sigma_{ij} = \sigma_{\lvert i-j\rvert+1}$
VC	Simple	q	$\sigma_{ij} = \sigma_k^2 1\ (i=j)$ $where\ 1(i=j) = 1$ $if\ i=j\ and\ 0\ if\ i \neq 1$ and i corresponds to the kth effect
SP(POW)(c-list)	Power Spatial	2	$\sigma^2\rho^{dij}$
SP(EXP)(c-list)	Exponential Spatial	2	$\sigma^2\exp\{-d_{ij}/\theta\}$
SP(GAU)(c-list)	Gaussian Spatial	2	$\sigma^2\exp\{-d_{ij}^2/\rho^2\}$

The c-list under the spatial type structures is used to refer to the names of the numeric variables used as coordinates of the location of the observation in space, and d_{ij} is the Euclidean distance between the *ith* and *jth* vectors of these coordinates, which correspond to the *ith* an *jth* observations in the input data set.

The VC structure is the standard variance components and is the default used by SAS.

$$\begin{bmatrix} \sigma_A^2 & 0 & 0 \\ 0 & \sigma_B^2 & 0 \\ 0 & 0 & \sigma_C^2 \end{bmatrix}$$

The Compound Symmetry structure assumes the covariances are homogeneous. There is a correlation between two separate measurements, but it is assumed that the correlation is constant regardless of how far apart the measurements are. This is unrealistic in a longitudinal data problem in the sense that observations closer to each other are more correlated than observations which are further apart.

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

However a compound symmetry structure is ideal if observations are clustered arbitrarily say in a household location, class and so on. Since there is no time ordering as in longitudinal data one can realistically assume a constant correlation across observations.

This problem of unequal correlation can be solved by using several approaches such as the Autoregressive (1) covariance structure. The correlation between m time units apart is ρ^m, $0<\rho<1$. The greater the distance (m), the smaller the magnitude of the covariance will be. In this case the covariance matrix is given by (taking four observations per individual as an illustration)

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

The Toeplitz structure is similar to AR(1) in that all measurements next to each other have the same correlation, measurements two time units apart have the same correlation different from the measurements one time unit apart, measurements three time units have the same correlation different from the measurements one and two time units apart, etc. There is however no assumption of exponential decay. Technically, the AR(1) is a special case of the Toeplitz. The Toeplitz model has as many parameters as there is distance.

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

The Unstructured covariance structure is the most flexible since it assumes all the variances and covariances are different. The covariance matrix for such a structure is given by

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

When observations are not necessarily equally spaced within and between individuals the best covariance structures to consider are the spatial type of structures. The correlations are positive and decreasing functions of the Euclidean distances between observations. These structures are advantageous when dealing with repeated measurement since they take into account the distance between the observations within each subject. The three most common spatial covariance structures are Power, Exponential and Gaussian as listed in Table 7.

The power spatial covariance structure is given by

$$\sigma^2\begin{bmatrix} 1 & \rho^{di12} & \rho^{di13} & \rho^{di14} \\ \rho^{di21} & 1 & \rho^{di23} & \rho^{di24} \\ \rho^{di31} & \rho^{di32} & 1 & \rho^{di34} \\ \rho^{di41} & \rho^{di42} & \rho^{di43} & 1 \end{bmatrix}$$

where d_{ijk} is the distance between the jth and kth observation within subject i and $0 < \rho < 1$.

The exponential spatial covariance structure is given by

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

The Gaussian spatial covariance structure given by

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

In all the specifications of the spatial structures above, d_{ijk} is the distance between the jth and kth observations within the same individual i. The advantage of the spatial type structures over the standard AR structure (which assumes equal spaced observations) is that they make use of the actual distance between observations. This is important because now the modeller can deal with unequally spaced observations within and between observations.

3.5.1 Selection of a Covariance Structure

There are various model selection criteria that can be used to discriminate between model choice and these include the Akaike information criteria (AIC), the Bayesian information criteria (BIC), the marginal quasi-information criteria (MQIC) and Akaike conditional information criteria (AICC). Generally the model with the smallest AIC, BIC, MQIC or AICC should be selected. Choosing a covariance structure which is too simple will increase the fixed effects Type I error rate, while choosing a covariance structure which is too complex will sacrifice power and efficiency. Thus a choice between possible covariance structures is a tradeoff between these two extreme limits. The algebraic expressions showing how the values of the different criteria are calculated are shown in the Table 8. They are all likelihood based thus existence of a likelihood is a pre-requisite to their utility.

Table 8: Selection criteria

AIC	$-2l + 2d$
MQIC	$-2l + 2d\log(\log(n))$
BIC	$-2l + dlog(n)$
AICC	$-2l + d(\log(n) + 1)$

Where l is the log likelihood, d is the number of parameters in the model and n is the sample size. The most common used method of model selection is the AIC. All the model selection criteria rely on imposing a penalty on the log-likelihood based on the number of parameters in the model. For the MQIC, BIC and AICC the expressions also include the sample size.

3.5.2 Selection of Mean Structure

Consider the null hypothesis where parameter β is from the subspace $\theta_{\beta,0}$ of the parameter space Θ_β such that $H_o: \beta \in \theta_{\beta,0}$, then the LR tests follows a chi-squared distribution with degrees of freedom equal to the difference between the dimension d of Θ_β and the dimension $\Theta_{\beta,0}$. Note that the above result is only valid if the models are fitted using ML estimation (Verbeke & Molenberghs, 2000, p. 63).

In order 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, the likelihood ratio (LR) test is commonly used. The test statistic is defined as

$$-2ln\lambda_N = -2ln\left[\frac{L_{ML}(\widehat{\Theta}_{\beta,0})}{L_{ML}(\widehat{\Theta}_\beta)}\right] \tag{3.50}$$

where L_{ML} denotes the maximum likelihood function and $\widehat{\Theta}_{\beta,0}$ and $\widehat{\Theta}_\beta$ are the maximum likelihood estimates obtained from maximizing L_{ML} (Verbeke & Molenberghs, 2000, pp. 62-63).

3.5.3 Selection of Random Effects

For valid conditional and marginal distributions of the data the positivity constraint on Ω is required. Consider the case where $H_o: \Theta_i = 0$ and $\Omega = (0, \infty)$. The value of Θ_i under the null hypothesis is on the boundary of the parameter space, and the distribution of the likelihood ratio test statistic λ is therefore nonstandard. This is a mixture of chi-square distributions. Testing hypotheses such as the need for random effects uses the likelihood ratio test statistic which has an asymptotic null distribution that is often a mixture of chi-squared distributions rather than the classical single chi-squared distribution (Verbeke & Molenberghs, 2000).

3.5.4 Sandwich Estimator

By using the sandwich estimator, we can correct the standard errors of the estimated fixed effects which are due to possible misspecification of the variance-covariance structure. This reflects the cautious nature of repeated measurements analysis assuming that we can never fit a perfect model, but at best we can fit a reasonable model to the data.

3.6 Model Diagnostics

After fitting a given model it is always appropriate to carry out a model diagnostics to assess how well the model fits. It is common to use model residuals as one of the byproducts of analysis to achieve this goal (Fitzmaurice et al., 2004, p. 237).

For analysis of longitudinal data we can extract a vector of residuals for each individual. The vector of residuals can be defined as

$$r_i = Y_i - X_i \hat{\beta} \tag{3.51}$$

with a mean equal to zero. Note that this is the marginal case, where the mean of Y_i is given by $E(Y_i) = X_i\beta$.

This gives an estimate of the vector of errors given as

$$e_i = Y_i - X_i\beta \tag{3.52}$$

(Fitzmaurice et al., 2004, p. 237).

The residuals can be used to assess the adequacy of the model for the covariance. This can be done by comparing the scatter-plot of the residuals against the predicted mean response to determine if there is any systematic trend where the residual for a given time occasion within an individual is given by

$$r_{ij} = Y_{ij} - X'_{ij}\hat{\beta} \tag{3.53}$$

and the predicted mean response can be written as

$$\hat{\mu}_{ij} = X'_{ij}\hat{\beta}. \tag{3.54}$$

For the model to be adequate there should be no systematic trend. The presence of this trend could imply that a quadratic term may need to be included, or that the covariate may need to be transformed (Fitzmaurice et al., 2004, p. 238).

Note that the components of the vector of residuals in equation (3.51) are correlated and may not have constant variance. The vectors of the errors is given in equation (3.52) with mean equal to zero, which is the same as the mean of the vector of residuals. The covariance of the residuals is however not the same as the covariance of the errors. The covariance of the residuals can be approximated by

$$Cov(r_i) \approx Cov(e_i) = \Sigma_i. \tag{3.56}$$

Since the covariance of the residuals have approximate covariance matrix, Σ_i, this has important implications for the examination of the residuals. Since the variance may not be constant, the range may not be constant when comparing the scatter-plot of the residuals against the predicted values or against time. For this reason, standard residual diagnostics should be avoided for examining the homogeneity of the residual variance or autocorrelation among the residuals. Another implication is that since the residuals from a regression analysis of longitudinal data may be correlated with the covariates, the scatter-plot may show a systematic trend when plotting the residuals against a selected covariate (Fitzmaurice et al., 2004, p. 238).

An outlier is defined as an observation that lies an abnormal distance from other observations in a random sample. Influential observations are those that appear to have a large influence on the parameter estimate. These should be investigated and removed if necessary.

3.7 Summary

A major problem with longitudinal data is that of missing values, inconsistent timed observations and incomplete data. The mixed model approach has assumptions and procedures that work well when dealing with these problems (Muller & Stewart, 2006). This approach involves building a model for the expected values and for the covariances of the data (Muller & Stewart, 2006). Choosing the best covariance structure is important for the interpretation of the random variation found in the data. It is also essential in order to obtain valid inferences for the parameters in the mean structure of the model (Verbeke, et al., 1998). A structure that is too restrictive will invalidate inferences, while a structure that is too simple will lead to inefficient estimation and poor estimation of standard errors (Verbeke, et al., 1998).

Missing data can be missing completely at random (MCAR) where the probability that an observation is missing does not depend either on the observed (y_{obs}) or the missed (y_{mis}) observation. Under this assumption missing observations cannot in any way bias the analysis if the assumption is true. Secondly data can be missing at random (MAR) where the probability that an observation outcome is missing may depend on the observed outcome (y_{obs}) but not on the missing values (y_{mis}). Lastly we can have missing not at random (MNAR) where the probability that an observation is missing depends on the missing outcome (y_{mis}) and possibly also on the observed (y_{obs}) (Diggle, et al., 2002, p. 283). Since the methods we adopt in the thesis are likelihood based we assume any missing data is MAR which is comfortably accounted for under the likelihood based method of analysis.

Chapter 4

Application of Linear Mixed Models to Sinikithemba Data

In most linear model applications involving several predictor variables a common step of analysis is to choose the set of variables that best predict the mean response or its function. Model selection is one of the most frequent problems encountered in data analysis. Model building can be done in numerous ways, such as backward, forward or stepwise procedures (Ngo & Brand, 2002). In this chapter we will first use a backward procedure using square root CD4+ count and log viral load separately as the response variables using linear mixed models. This will be done using the statistical software SAS version 9.2 (SAS Institute Inc., Cary, NC, USA) and specifically, we will use the PROC MIXED procedure in SAS.

4.1 Square Root CD4+ Count as the Response

Since CD4+ count was initially not normally distributed, a square root transformation was applied on CD4+ count and the transformed variable used in the analysis. The explanatory variables that will be included are gender (male or female), HLA-B type (B*0801, B*1503, B*1510, B*4201, B*4403, B*5702, B*5703, B*5801, B*5802 and B*8100) and IL-10 genotypes (-592 genotype and -1082 genotype). The -592 genotype includes AA, CA and CC. The -1082 genotype includes AA, AG and GG.

The aim is to find the best population averaged or marginal model that best describes the mean responses. For each model the best covariance structure and estimation method are chosen on the basis of comparing the AIC's for different models. Covariance structures that will be used and compared include Compound Symmetry

(CS), Spatial Power (sp(pow)(c-list)), Spatial Exponential (sp(exp)(c-list)) and Spatial Gaussian (sp(gau)(c-list)) where c-list is defined as the list containing names of the numeric variables that are used as coordinates of the observations location in space or time. The methods of estimation that will be used are the maximum likelihood method and the restricted maximum likelihood method. The differences between the AIC, AICC and BIC for the ML and REML procedures is due to the fact that REML takes into account the degrees of freedom when estimating the fixed effects mean parameters whilst ML does not. This model building strategy will be done in two steps. Firstly, the covariance structure is sought and chosen then secondly, the mean structure must be chosen for the covariance structure that is adopted.

4.1.1 Covariance Structure

If the mean structure is not correctly chosen, it is likely that a more complicated covariance structure than necessary will be used. Therefore to choose the best covariance structure, we will first start with a full (Verbeke & Molenberghs, 2000, p. 123). This model includes gender, all the HLA-B types, both -592 and -1082 IL-10 genotypes, time, the two-way interactions between each of the HLA-B types and the IL-10 genotypes, the two-way interactions between the HLA-B types and time, the two-way interactions between the IL-10 genotypes and time and the three-way interactions between each of the HLA-B types, the IL-10 genotypes and time. Maximum likelihood estimation is used. The covariance structures are then compared using a likelihood ratio test (for the nested covariance structures) or AIC (for the unnested covariance structures).

The unstructured (UN), autoregressive (AR(1)) and toeplitz (TOEP) covariance structures were found not to be appropriate for this analysis. Due to a large number of repeated measurements per subject, the unstructured covariance would not allow the model to converge. The autoregressive structure was not suitable as these measurements were not equally spaced. The model with the toeplitz covariance structure was unable to converge to a positive definite hessian matrix. Therefore the compound symmetry and spatial covariance structures will be compared. Since the compound symmetry and

spatial covariance structures are not nested within each other, the AICs were used to compare and choose the best covariance structure. From Table 9 and Table 10, we can see that the compound symmetry covariance structure is the best covariance structure using the maximum likelihood method (AIC=19132.7). Therefore this will be the model we will use to determine the mean structure of the model.

Table 9: Fit statistics for full model using ML method

	CS	SP(POW)	SP(EXP)	SP(GAU)
-2 Log Likelihood	18920.7	19159.6	19159.6	22141.9
AIC	19132.7	19371.6	19371.6	22353.9
AICC	19138.4	19377.3	19377.3	22359.6
BIC	19562.5	19801.4	19801.4	22783.6

Table 10: Fit statistics for full model using REML method

	CS	SP(POW)	SP(EXP)	SP(GAU)
-2 Log Likelihood	19477.5	19606.7	19606.7	22728.8
AIC	19481.5	19610.7	19610.7	22732.8
AICC	19481.5	19610.7	19610.7	22732.8
BIC	19489.6	19618.8	19618.8	22740.9

Fitting the full model, together with the compound symmetry covariance structure using the maximum likelihood procedure, the Type III tests of fixed effects are found. These can be found in Table 1B in Appendix B. From this, the mean structure can now be chosen.

4.1.2 Mean Structure

Choosing the mean structure is done using the full model and the Type III tests of fixed effects shown in Table 1B in Appendix B. For this step, the maximum likelihood method is always used. From the Type III effects, terms are systematically dropped from the model, starting with the term that is the least significant. This newer, simpler model is then compared to the original model, using the likelihood ratio test. If the reduced model is not found to be significantly different from the original model (p-value>0.05), it can then be used over the original model. This step is repeated, each time taking out more terms, until only the significant terms are left in the model. Terms that are found to be insignificant, but their interaction with another term is found to be significant, will need to be included in the final model. The final analysis is then repeated, using the correct covariance structure, and final mean structure. Table 11 gives the final model with square root CD4+ count as the response, including gender, HLA-B types, IL-10 genotypes and their interactions with each other and time. We find that time is significantly associated with mean square root CD4+ count at the 5% level of significance (p-value<0.0001). HLA-B*5703 (p-value=0.0064) and B*5802 (p-value=0.0301) are both significantly associated with square root CD4+ count. The two-way interactions between HLA-B*1503, B*1510, B*5702, B*5703, B*5801, B*5802, B*8100 and time are significantly associated with mean square root CD4+ count (p-value<0.05). The interaction of IL-10 -592 genotype and time is also significantly associated with square root CD4+ count (p-value<0.05). The interaction between HLA-B*5802 and -592 genotype is significantly associated with square root CD4+ count (p-value=0.0307). The three-way interaction between HLA-B*1503, -592 genotype and time (p-value=0.0134), B*1510, -592 genotype and time (p-value<0.0001), B*4201, -592 genotype and time (p-value<0.0001), B*4403, -592 genotype and time (p-value<0.0001), B*5702, -592 genotype and time (p-value<0.0001), B*5801, -592 genotype and time (p-value=0.0005), B*5802, -592 genotype and time (p-value=0.0163), B*8100, -592 genotype and time (p-value<0.0001), B*0801, -1082 genotype and time (p-value=0.0012), B*1510, -1082 genotype and time (p-value=0.0440), B*4201, -1082 genotype and time (p-value=0.0451), B*4403, -1082 genotype and time (p-value<0.0001), B*5801, -1082 genotype and time (p-

value=0.0009), and B*8100, -1082 genotype and time (p-value<0.0001) are significantly associated with mean square root CD4+ count. The significant two-way interactions involving time imply that the slope or increase in mean square root CD4+ count differs for different levels of the categorical factor. The significant three-way interaction involving time imply that the slope also depends on the level of combination of the other remaining two categorical variables.

Table 11: Type III tests of fixed effects for final model with square root CD4+ count as the response

Effect	Num DF	Den DF	F Value	Pr>F
Time	1	3590	21.31	<.0001
B0801	1	403	0.14	0.7117
B1503	1	403	2.84	0.0925
B1510	1	403	0.57	0.4508
B4201	1	403	0.00	0.9494
B4403	1	403	1.40	0.2368
B5702	1	403	0.00	0.9543
B5703	1	403	4.67	0.0313
B5801	1	403	1.66	0.1990
B5802	1	403	7.40	0.0068
B8100	1	403	2.87	0.0909
-592 Genotype	2	403	0.39	0.6788
-1082 Genotype	2	403	0.28	0.7591
B1503*Time	1	3590	6.79	0.0092
B1510*Time	1	3590	15.45	<.0001
B5702*Time	1	3590	18.01	<.0001
B5703*Time	1	3590	22.72	<.0001
B5801*Time	1	3590	19.03	<.0001
B5802*Time	1	3590	13.03	0.0003
B8100*Time	1	3590	4.02	0.0451
-592 Genotype*Time	2	3590	5.53	0.0040
B1510*-592 Genotypes	2	403	1.41	0.2451
B5801*-592 Genotypes	2	403	2.44	0.0888
B5802*-592 Genotypes	2	403	3.51	0.0307
B5801*-1082 Genotypes	2	403	1.77	0.1715
B0801*-592 Genotypes*Time	2	3590	2.64	0.0713
B1503*-592 Genotypes*Time	2	3590	4.32	0.0134
B1510*-592 Genotypes*Time	2	3590	11.91	<.0001
B4201*-592 Genotypes*Time	2	3590	12.37	<.0001
B4403*-592 Genotypes*Time	2	3590	9.59	<.0001
B5702*-592 Genotypes*Time	2	3590	18.52	<.0001
B5801*-592 Genotypes*Time	2	3590	7.63	0.0005
B5802*-592 Genotypes*Time	2	3590	4.12	0.0163
B8100*-592 Genotypes*Time	2	3590	22.07	<.0001
B0801*-1082 Genotypes*Time	2	3590	6.75	0.0012
B1510*-1082 Genotypes*Time	2	3590	3.13	0.0440
B4201*-1082 Genotypes*Time	2	3590	3.10	0.0451
B4403*-1082 Genotypes*Time	2	3590	10.32	<.0001
B5801*-1082 Genotypes*Time	2	3590	6.98	0.0009
B8100*-1082 Genotypes*Time	2	3590	14.26	<.0001

In Table 12 the reference category for gender is female, ‘absent’ for the HLA-B types, CC for -592 genotypes and GG for -1082 genotypes. The gender effect is found not to be significant in this model. The estimate for time implies that the overall mean square root CD4+ count decreases at a rate of 0.00310 units per day (p-value<0.0001). Individuals who have HLA-B*1510 have a mean square root CD4+ count 2.6307 units lower (p-value=0.0219) and individuals with HLA-B*5802 have a mean square root

CD4+ count 2.9861 lower (p-value=0.0045) than for those without the respective HLA-B types. This implies that HLA-B*1510 and B*5802 are associated with a faster progression of HIV. Individual with HLA-B*5703 have a mean square root CD4+ count 2.9525 higher (p-value=0.0313) than individuals without B*5703. This implies that HLA-B*5703 is a good controller of HIV. On the -1082 loci, when the AA genotype is present the mean square root CD4+ count for such an individual is 2.1249 units lower than when the CC genotype is present (p-value=0.0374). The estimate for the interaction between HLA-B*1503 and time shows that individuals with B*1503 have a mean square root CD4+ count that increases at an extra rate of 0.001349 units per day than that of an individual without this HLA-B type (p-value<0.0001). The interaction between HLA-B*1510 and time indicates that when HLA-B*1510 is present, the mean square root CD4+ count increases at an extra rate of 0.001919 units per day than when HLA-B*1510 is absent (p-value<0.0001). When HLA-B*5703 is present, the estimate for the interaction between B*5703 and time shows that the mean square root CD4+ count increases at an extra rate of 0.001467 units per day than when this HLA-B type is not present (p-value<0.0001). The interaction between HLA-B*5801 and time has a coefficient of 0.001696. This is interpreted to mean that compared to the reference HLA-B level, individuals with B*5801 have a mean square root CD4+ count that increases at an extra rate of 0.001696 units per day (p-value=0.0147). This implies that HLA-B*1503, B*1510, B*5703 and B*5801 are all associated with a slower progression of HIV over time. The interaction between HLA-B*5702 and time shows that the mean square root CD4+ count decreases at an extra rate of 0.02243 units per day for an individual with B*5702 than for an individual without this HLA-B type (p-value<0.0001), thus HLA-B*5702 is associated with a faster progression of HIV over time. If we consider the -592 loci, the interaction between the CA genotype and time shows that the square root CD4+ count increases at an extra rate of 0.001499 units per day than when the CC genotype is present (p-value<0.0001). This implies that the CA genotype acts as a good controller of HIV. The significant interaction between HLA-B*5801 and CA genotype shows that the mean square root CD4+ count is 4.7099 units more if HLA-B*5801 is present (p-value=0.0461) than when it is absent. The coefficient for the interaction between HLA-B*5802 and the CA genotype is 3.1686. This implies that the mean square root CD4+ count is 3.1686 units more when this

HLA-B type is present (p-value=0.0297) compared to when this is absent. This indicates that the CA -592 genotype is associated with a slower progression of HIV when HLA-B*5801 is present and associated with a faster progression of HIV when B*5802 is present. When HLA-B*0801 and the CA genotype is present, the interaction between B*0801, the CA genotype and time shows that an individual's mean square root CD4+ count increases at an extra rate of 0.001546 units per day (p-value=0.0300), and when the CC genotype is present this shows that an individual's mean square root CD4+ count increases at an extra rate of 0.002450 units per day (p-value<0.0001), compared to the absence of this HLA-B type. An individual who has B*1503 and CA genotype has a mean square root CD4+ count that decreases at an extra rate of 0.00138 units per day than an individual without B*1503 (p-value=0.0042). Still considering the -592 loci, we see that the interaction between B*1510, the AA genotype and time is significant which shows that the predicted square root CD4+ count increases at an extra rate of 0.003922 units per day, compared to the absence of this HLA-B type (p-value<0.0001). The coefficient of the interaction between HLA-B*4201, the CA genotype and time is -0.00121. This demonstrates that the mean square root CD4+ count decreases at an extra rate of 0.00121 units per day (p-value=0.0377), while it increases at an extra rate of 0.000857 units per day when the CC genotype is present (p-value=0.0349) compared to when B*4201 is absent. The estimate for an individual with HLA-B*4403 and the CA genotype over time is -0.00314. This suggests that the mean square root CD4+ count decreases at an extra rate of 0.00314 units per day (p-value<0.0001) than an individual without this HLA-B type, and when HLA-B*4403 and the CC genotype is present the an individual's mean square root CD4+ count decreases at an extra rate of 0.00118 units per day (p-value=0.0386), compared to that for an individual without B*4403. The coefficient for HLA-B*5702, the AA genotype and time is 0.02516 and the coefficient for HLA-B*5702, the CA genotype and time is 0.02109. This indicates that the predicted square root CD4+ count increases at an extra rate of 0.02516 units per day for an individual with B*5702 and the AA genotype (p-value<0.0001), and increases at an extra rate of 0.02109 units per day for an individual with B*5702 and the CA genotype (p-value<0.0001), compared to that of an individual without this HLA-B type and the given genotype. The estimate for the interaction HLA-B*5801, the CA genotype and time is 0.002496. This illustrates that the mean square

root CD4+ count increases at an extra rate of 0.002496 units per day (p-value<0.0001), compared to the mean square root CD4+ count when B*5801 is absent. The estimate for an individual with HLA-B*5802 and the AA genotype shows that the mean square root CD4+ count decreases at an extra rate of 0.00332 units per day (p-value=0.0159) than that for an individual without this HLA-B type. When HLA-B*8100 and the AA genotype is present, the interaction between B*8100, -592 genotype and time indicates that the mean square root CD4+ count decreases at an extra rate of 0.00527 units more per day (p-value<0.0001), and decreases by 0.00413 units per day (p-value<0.0001) when B*8100 is present compared to when B*8100 is absent. Looking now at the -1082 loci, the coefficient of the interaction between HLA-B*0801, AA genotype and time shows that the mean square root CD4+ count decreases at an extra rate of 0.00314 units per day (p-value<0.0001) when HLA-B*0801 is present compared to when this HLA-B type is absent. The coefficient of the interaction between the same HLA-B type, the AG genotype and time shows that the mean square root CD4+ count decreases at a rate of 0.00149 units more per day (p-value=0.0120) than for the same interaction without this HLA-B type. Still considering the -1082 loci, the interaction between HLA-B*1510, the AA genotype and time shows that the mean square root CD4+ count decreases at a rate of 0.00255 units more per day (p-value=0.0132) and the interaction between B*1510, the AG genotype and time shows that the mean square root CD4+ count decreases at an extra rate of 0.00198 units per day (p-value=0.0411) compared to that when B*1510 is absent. When HLA-B*4403 and the AA genotype is present, the interaction between B*4403, -1082 genotype and time shows that the mean square root CD4+ count increases at an extra rate of 0.003114 units per day (p-value<0.0001), and shows that when the same HLA-B type and the AG genotype is present the mean square root CD4+ count increases at an extra rate of 0.001931 units per day (p-value=0.0039), compared to that when B*4403 is absent. The coefficient of the interaction between B*5801, the AG genotype and time implies that the mean square root CD4+ count decreases at an extra rate of 0.00318 units per day, compared to when the HLA-B type is absent (p-value=0.0007). The interaction between B*8100, the AA genotype and time illustrates that the mean square root CD4+ count increases at an extra rate of 0.004033 units per day (p-value<0.0001) and increases by 0.002314 units per day (p-value=0.0022) when B*8100 and the AG -1082 genotype is present, compared to when B*8100 is absent.

Table 12: Solution for fixed effects with square root CD4+ count as the response

Effect	Genotype	Estimate	Std Error	DF	t-value	p-value
Intercept		21.1066	1.0146	403	20.80	<.0001
Time		-0.00310	0.000366	3590	-8.48	<.0001
B0801		-0.2991	0.8088	403	-0.37	0.7117
B1503		-1.2200	0.7234	403	-1.69	0.0925
B1510		-2.6307	1.1430	403	-2.30	0.0219
B4201		-0.04540	0.7153	403	-0.06	0.9494
B4403		0.9337	0.7880	403	1.18	0.2368
B5702		0.1293	2.2564	403	0.06	0.9543
B5703		2.9525	1.3667	403	2.16	0.0313
B5801		-1.9469	2.5696	403	-0.76	0.4491
B5802		-2.9861	1.0463	403	-2.85	0.0045
B8100		1.6411	0.9683	403	1.69	0.0909
-592 Genotype	AA	0.3009	1.2264	403	0.25	0.8063
-592 Genotype	CA	-0.7652	0.7648	403	-1.00	0.3177
-1082 Genotype	AA	-2.1249	1.0177	403	-2.09	0.0374
-1082 Genotype	AG	-1.4756	0.9888	403	-1.49	0.1364
B1503*Time		0.001349	0.000282	3590	4.78	<.0001
B1510*Time		0.001919	0.000889	3590	2.16	0.0310
B5702*Time		-0.02243	0.004839	3590	-4.64	<.0001
B5703*Time		0.001467	0.000308	3590	4.77	<.0001
B5801*Time		0.001696	0.000695	3590	2.44	0.0147
B5802*Time		-0.00030	0.000303	3590	-1.00	0.3172
B8100*Time		0.000243	0.000446	3590	0.55	0.5857
-592 Genotype*Time	AA	0.000386	0.000706	3590	0.55	0.5847
-592 Genotype*Time	CA	0.001499	0.000352	3590	4.26	<.0001
B1510*-592 Genotype	AA	4.2517	2.6511	3590	1.60	0.1095
B1510*-592 Genotype	CA	1.4279	1.4821	403	0.96	0.3359
B5801*-592 Genotype	AA	-6.5368	3.4741	403	-1.88	0.0606
B5801*-592 Genotype	CA	-4.7099	2.3544	403	-2.00	0.0461
B5802*-592 Genotype	AA	-2.2143	2.6208	403	-0.84	0.3987
B5802*-592 Genotype	CA	3.1686	1.4519	403	2.18	0.0297
B5801*-1082 Genotype	AA	6.6686	3.5636	403	1.87	0.0620
B5801*-1082 Genotype	AG	5.3379	3.2681	403	1.63	0.1032
B0801*-592 Genotype*Time	AA	0.000513	0.001141	3590	0.45	0.6531
B0801*-592 Genotype*Time	CA	0.001546	0.000712	3590	2.17	0.0300
B0801*-592 Genotype*Time	CC	0.002450	0.000514	3590	4.76	<.0001
B1503*-592 Genotype*Time	AA	-0.00085	0.000581	3590	-1.46	0.1437
B1503*-592 Genotype*Time	CA	-0.00138	0.000481	3590	-2.86	0.0042
B1510*-592 Genotype*Time	AA	0.003922	0.000934	3590	4.20	<.0001
B1510*-592 Genotype*Time	CA	-0.00029	0.000499	3590	-0.58	0.5636
B4201*-592 Genotype*Time	AA	0.000352	0.000770	3590	0.46	0.6478
B4201*-592 Genotype*Time	CA	-0.00121	0.000582	3590	-2.08	0.0377
B4201*-592 Genotype*Time	CC	0.000857	0.000406	3590	2.11	0.0349
B4403*-592 Genotype*Time	AA	-0.00173	0.000969	3590	-1.78	0.0747
B4403*-592 Genotype*Time	CA	-0.00314	0.000713	3590	-4.40	<.0001
B4403*-592 Genotype*Time	CC	-0.00118	0.000572	3590	-2.07	0.0386
B5702*-592 Genotype*Time	AA	0.02516	0.004895	3590	5.14	<.0001
B5702*-592 Genotype*Time	CA	0.02109	0.004898	3590	4.31	<.0001
B5801*-592 Genotype*Time	AA	0.001530	0.000792	3590	1.93	0.0535
B5801*-592 Genotype*Time	CA	0.002496	0.000639	3590	3.91	<.0001
B5802*-592 Genotype*Time	AA	-0.00332	0.001376	3590	-2.41	0.0159
B5802*-592 Genotype*Time	CA	-0.00089	0.000472	3590	-1.89	0.0585
B8100*-592 Genotype*Time	AA	-0.00527	0.001113	3590	-4.73	<.0001
B8100*-592 Genotype*Time	CA	-0.00413	0.000665	3590	-6.21	<.0001
B0801*-1082 Genotype*Time	AA	-0.00314	0.000706	3590	-4.45	<.0001

B0801*-1082 Genotype*Time	AG	-0.00149	0.000592	3590	-2.51	0.0120
B0801*-1082 Genotype*Time	AA	-0.00046	0.000418	3590	-1.09	0.2745
B0801*-1082 Genotype*Time	AG	0.000484	0.000393	3590	1.23	0.2184
B1510*-1082 Genotype*Time	AA	-0.00255	0.001027	3590	-2.48	0.0132
B1510*-1082 Genotype*Time	AG	-0.00198	0.000967	3590	-2.04	0.0411
B4201*-1082 Genotype*Time	AA	0.000994	0.000563	3590	1.77	0.0775
B4201*-1082 Genotype*Time	AG	-0.00002	0.000500	3590	-0.05	0.9629
B4403*-1082 Genotype*Time	AA	0.003114	0.000702	3590	4.44	<.0001
B4403*-1082 Genotype*Time	AG	0.001931	0.000669	3590	2.89	0.0039
B5801*-1082 Genotype*Time	AA	-0.00163	0.000874	3590	-1.87	0.0620
B5801*-1082 Genotype*Time	AG	-0.00318	0.000936	3590	-3.40	0.0007
B8100*-1082 Genotype*Time	AA	0.004033	0.000760	3590	5.31	<.0001
B8100*-1082 Genotype*Time	AG	0.002314	0.000756	3590	3.06	0.0022

4.1.3 Diagnostic Analysis for the Model Excluding Random Effects

The histogram of the residuals shown in Figure 25 indicates approximate normality for this model. The normal quantile plot of the residuals displays slight systematic departures from a straight line in the tails. This reveals that this model may be a good representation of the data, but models including random effects need to be examined.

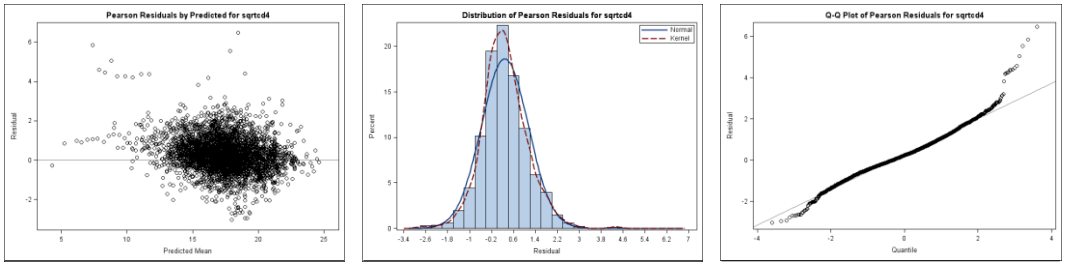


Figure 25: Model diagnostics for square root CD4+ count as the response

4.1.4 Random Effects

The variability in the subject-specific intercepts and slopes may not be completely explained by the covariates in the model. This can be solved by using random effects, which represent this variability (Verbeke & Molenberghs, 2000, p. 69). The model helps to account for the extra variability due to individual to individual heterogeneity.

For the purpose of this analysis, we will compare models that include a random intercept and slope, a random intercept, and no random effects.

Testing for the need of random effects cannot be done using classical likelihood ratio tests. This is due to the fact that the likelihood ratio statistic for the null hypothesis does not have the classical asymptotic chi-squared distribution (Verbeke & Molenberghs, 2000, p. 133). Instead of using the classical single chi-squared distribution, a mixture of chi-squared distributions should be used (Verbeke & Molenberghs, 2000, p. 69).

Comparing the model with a random intercept and slope (time) to a model with only a random intercept, we get a p-value > 0.05 (using degrees of freedom equal to 2 and 3 respectively). This implies that we do not need to keep the random intercept and slope for the final model. Comparing the model with only a random intercept and the model with no random effects gives a p-value > 0.05 (using degrees of freedom equal to 1 and 0 respectively). This implies that we do not need to keep the random intercept in the final model. From this, we can conclude that the best model is in fact the model that does not contain any random effects. This is confirmed by fit criteria shown in Table 13. It can be seen that the model including no random effects it the best model (AIC=19133.5).

Table 13: Fit criteria for CD4+ count comparing random effects

	Random Intercept and Slope	Random Intercept	No random Effects
-2 Log Likelihood	18919.5	18919.5	18919.5
AIC	19135.5	19135.5	19133.5
AICC	19141.4	19141.4	19139.3
BIC	19573.3	19573.3	19567.3

4.2 Log Viral Load as the Response

The same procedure used to analyze the square root CD4+ count will now be used in the analysis of log viral load.

4.2.1 Covariance Structure

Using the full model, the spatial power covariance structure with the maximum likelihood procedure is chosen (AIC=4220.0). This can be seen in Table 14 and Table 15. The type III tests of fixed effects can be found in Table 2B in Appendix B. From these results the mean structure can be chosen.

Table 14: Fit statistics for full model using ML method

	CS	SP(POW)	SP(EXP)	SP(GAU)
-2 Log Likelihood	4301.9	4220.0	5547.0	5547.0
AIC	4515.9	4434.0	5761.0	5761.0
AICC	4527.1	4445.2	5772.2	5772.2
BIC	4949.7	4867.9	6194.9	6194.9

Table 15: Fit statistics for full model using REML method

	CS	SP(POW)	SP(EXP)	SP(GAU)
-2 Log Likelihood	5164.3	5028.3	6453.7	6453.7
AIC	5168.3	5032.3	6457.7	6457.7
AICC	5168.3	5032.3	6457.7	6457.7
BIC	5176.4	5040.4	6465.8	6465.8

4.2.2 Mean Structure

The same procedure for choosing the mean structure as described under the square root CD4+ count model is used. The terms in the final means structure model are shown in Table 16. It can be seen from this table using the log viral load as the response, the final model includes gender, HLA-B types, IL-10 genotypes and their interactions with each other and time. We find that gender is significantly associated with log viral load at the 5% level of significance (p-value=0.0308). HLA-B*5703 (p-value=0.0006), B*5802 (p-value=0.0322) and the -592 genotype (p-value=0.0041) are also significantly associated with log viral load. The interactions between HLA-B*1503 and the -592 genotype (p-value=0.0049), B*8100 and the -592 genotype (p-value=0.0222) and the interaction between B*5702, the -592 genotype and time (p-value=0.0135), are significantly associated with log viral load.

Table 16: Type III tests of fixed effects for log viral load as response for model with no random effects

Effect	Num DF	Den DF	F Value	Pr>F
Gender	1	404	4.70	0.0308
Time	1	1739	3.07	0.0798
B1503	1	404	1.05	0.3051
B4201	1	404	1.75	0.1866
B5702	1	404	2.01	0.1571
B5703	1	404	11.94	0.0006
B5802	1	404	4.62	0.0322
B8100	1	404	3.49	0.0626
-592 Genotype	2	404	5.56	0.0041
-1082 Genotype	2	404	0.08	0.9192
B4201*Time	1	1739	3.36	0.0668
B1503*-592 Genotype	2	404	5.38	0.0049
B8100*-592 Genotype	2	404	3.84	0.0222
B4201*-1082 Genotype	2	404	2.02	0.1337
B5802*-1082 Genotype	2	404	2.64	0.0726
B5702*-1082 Genotype*Time	3	1739	3.57	0.0135
B5703*-1082 Genotype*Time	2	1739	2.93	0.0536

From the results in Table 17 it can be seen that at the 5 % level of significance, gender is found to be significantly associated with log viral load (p-value=0.0308). The estimate shows the predicted log viral load for a female is 0.2017 units lower than that for a male. The estimate for time shows that predicted log viral load decreases by 0.00025 units per day (p-value=0.0013). An individual with HLA-B*8100 has a predicted log viral load that is 0.8154 units lower than an individual without this HLA-B type (p-value<0.0001). This implies that B*8100 is a good controller of HIV replication. Looking at the -592 loci, an individual who possesses HLA-B*1503 and the CA genotype has a predicted log viral load that is 0.6502 units higher than an individual who possesses B*1503 and the CC genotype (p-value=0.0016). This implies that an individual with B*1503 and the CC genotype is protected against HIV better than an individual with B*1503 and the CA genotype. An individual who possesses HLA-B*8100 and the CA genotype has a predicted log viral load that is 0.6744 units higher than an individual who possesses B*8100 and the CC genotype (p-value=0.0092). This implies that an individual with HLA-B*8100 and the CA genotype is less protected than an individual without HLA-B*8100 and the CC genotype. Over time, the predicted log viral load for an individual who possesses B*5702 and the AG genotype will have an additional rate of increase of 0.001496 units per day than an individual who does not

possess this HLA-B*5702 (p-value=0.0020) and the predicted log viral load for an individual who possesses B*5703 and the AG genotype will have an additional rate of increase of 0.000552 units per day than for an individual who does not possess this HLA-B*5703 (p-value=0.0212). Thus these HLA-B and genotype interactions are disadvantageous for an individual's ability to control HIV replication over time.

Table 17: Solution for fixed effects for log viral load for model with no random effects

Effect	Genotype	Estimate	STD Error	DF	t Value	Pr>|t|
Intercept		4.9106	0.2258	404	21.75	<.0001
Gender		-0.2017	0.09308	404	-2.17	0.0308
Time		-0.00025	0.000076	1739	-3.23	0.0013
B1503		-0.1326	0.1371	404	-0.97	0.3342
B4201		-0.1934	0.2640	404	-0.73	0.4642
B5702		-0.5384	0.3799	404	-1.42	0.1571
B5703		-0.8043	0.2328	404	-3.45	0.0006
B5802		0.3715	0.2808	404	1.32	0.1866
B8100		-0.8154	0.1888	404	-4.32	<.0001
-592 Genotype	AA	0.04594	0.1860	404	0.25	0.8050
-592 Genotype	CA	-0.1370	0.1108	404	-1.24	0.2171
-1082 Genotype	AA	0.1223	0.1569	404	0.78	0.4359
-1082 Genotype	AG	0.1573	0.1516	404	1.04	0.3000
B4201*Time		0.000203	0.000111	1739	1.83	0.0668
B1503*-592 Genotype	AA	0.07553	0.2753	404	0.27	0.7839
B1503*-592 Genotype	CA	0.6502	0.2042	404	3.18	0.0016
B8100*-592 Genotype	AA	0.7949	0.4885	404	1.63	0.1045
B8100*-592 Genotype	CA	0.6744	0.2578	404	2.62	0.0092
B4201*-1082 Genotype	AA	-0.1581	0.2841	404	-0.56	0.5781
B4201*-1082 Genotype	AG	0.2275	0.2843	404	0.80	0.4239
B5802*-1082 Genotype	AA	0.06013	0.3139	404	0.19	0.8482
B5802*-1082 Genotype	AG	-0.4115	0.3203	404	-1.28	0.1996
B5702*-1082 Genotype*Time	AA	0.000230	0.000449	1739	0.51	0.6087
B5702*-1082 Genotype*Time	AG	0.001496	0.000484	1739	3.09	0.0020
B5702*-1082 Genotype*Time	GG	0.002660	0.002156	1739	1.23	0.2176
B5702*-1082 Genotype*Time	AA	0.000094	0.000155	1739	0.60	0.5455
B5702*-1082 Genotype*Time	AG	0.00000557	0.000102	1739	-0.05	0.9563
B5703*-1082 Genotype*Time	AG	0.000552	0.000239	1739	2.31	0.0212
B5703*-1082 Genotype*Time	GG	-0.00003	0.000285	1739	-0.11	0.9115

4.2.3 Diagnostics Analysis for the Model Excluding Random Effects

The histogram of the residuals after fitting the model shown in Figure 26 indicates slight skewness. The normal quantile plot of the residuals shows that there is a slight systematic departure from a straight line.

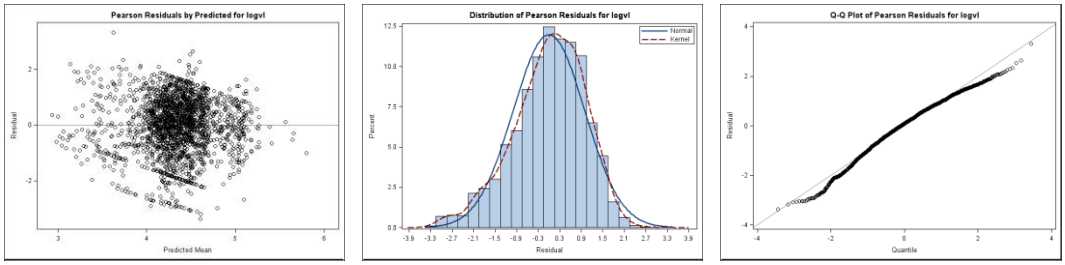


Figure 26: Model diagnostics for log viral load

4.2.4 Random Effects

Comparing the model with a random intercept and slope (time) to a model with only a random intercept gives a p-value > 0.05, using a chi-squared distribution with 2 and 1 degrees of freedom respectively. This implies that the simpler model having a random intercept only is sufficient. Comparing the model with a random intercept and no random effects gives a p-value ≈ 0 using a chi-squared distribution with 1 and 0 degrees of freedom. From this we can conclude that the model including a random intercept is the best model and will therefore be used for the final analysis. This is confirmed by the AIC statistics given in Table 18, with the model including the random intercept having the smallest AIC (AIC=4324.3). It should be noted that since the best model includes a random intercept, this implies that there is a strong individual to individual variability in log viral load at baseline.

Table 18: Information criteria for random effects with log viral load as the response

	Random Intercept and Slope	Random Intercept	No random Effects
-2 Log Likelihood	4107.1	4108.3	4220.0
AIC	4327.1	4324.3	4434.0
AICC	4338.9	4335.7	4445.2
BIC	4773.1	4762.1	4867.9

4.2.5 Model Including Random Intercept

The mean structure is chosen using the Type III Tests for fixed effects shown in Table 3B in Appendix B. After removing the non-significant terms and using the likelihood ratio test to compare models, we find the simplest mean structure. The Type III tests for fixed effects for the final model are shown in Table 19. At a 5% level of significance, gender (p-value=0.0296) and time (p-value=0.0154) are significantly associated with log viral load. HLA-B*5703 (p-value<0.0001), B*5801 (p-value=0.0375), B*8100 (p-value=0.0037) and the -1082 genotype (p-value=0.0198) are shown to be significantly associated with log viral load. The interactions between HLA-B*4201 and time (p-value=0.0039), B*4403, the -592 genotype and time (p-value=0.0443), B*5702, the -592 genotype and time (p-value=0.0002), B*8100, the -592 genotype and time (p-value=0.0400) and B*0801, the -1082 genotype and time (p-value=0.0388) are significantly associated with log viral load.

Table 19: Type III tests for fixed effects for final model with log viral load as the response

Effect	Num DF	Den DF	F Value	Pr>F
Gender	1	1723	4.74	0.0296
Time	1	1723	5.88	0.0154
B0801	1	1723	0.23	0.6315
B4201	1	1723	3.13	0.0768
B4403	1	1723	1.11	0.2915
B5702	1	1723	2.47	0.1162
B5703	1	1723	15.62	<.0001
B5801	1	1723	4.33	0.0375
B5802	1	1723	0.10	0.7528
B8100	1	1723	8.47	0.0037
-592 Genotype	2	1723	0.70	0.4957
-1082 Genotype	2	1723	3.93	0.0198
B4201*Time	1	1723	8.34	0.0039
B5703*Time	1	1723	3.61	0.0574
B0801*-1082 Genotype	2	1723	2.89	0.0557
B4201*-1082 Genotype	2	1723	2.83	0.0593
B4403*-592 Genotype*Time	3	1723	2.70	0.0443
B5702*-592 Genotype*Time	3	1723	6.75	0.0002
B5802*-592 Genotype*Time	3	1723	2.48	0.0599
B8100*-592 Genotype*Time	2	1723	3.23	0.0400
B0801*-1082 Genotype*Time	3	1723	2.80	0.0388
B8100*-1082 Genotype*Time	2	1723	2.75	0.0641

At a 5% level of significance, it is shown in Table 20 that females have a predicted log viral load that is 0.2260 units lower than that for males. Individuals who have HLA*B5703 have a predicted log viral load that is 0.8926 units lower (p-value<0.0001) than those without this HLA-B type, those with B*5801 have a predicted log viral load 0.2944 units lower (p-value=0.0375) than individuals who do not have this HLA-B type and individuals with B*8100 have a predicted log viral load 0.4491 units lower (p-value=0.0037) than individuals who do not have this HLA-B type. The interaction between HLA-B*4201 and time shows that when HLA-B*4201 is present, the predicted log viral load increases at an extra rate of 0.000224 units more per day than when HLA-B*4201 is absent (p-value=0.0039). On the -592 loci, the estimate of the interaction between HLA-B*5702, the CA genotype and time shows that the predicted log viral load increases at an extra rate of 0.001509 units more per day than that when B*5702 is absent (p-value<0.0001). The estimate of the interaction between HLA-B*5802, the CA genotype and time indicates that the predicted log viral load increases at an extra rate of 0.000364 units more per day compared to that of an individual without B*5802 (p-

value=0.0082). The coefficient of the interaction between HLA-B*8100, the AA genotype and time is 0.001009. This is interpreted to mean that compared to the absence of HLA-B*8100, an individual with B*8100 and the AA genotype has a predicted log viral load which increases at a rate of 0.001009 units more per day (p-value=0.0401). Now looking at the -1082 loci, the estimate of the interaction between HLA-B*0801, the GG genotype and time illustrates that the predicted log viral load decreases at a rate of 0.00055 units more per day than if HLA-B*0801 is absent (p-value=0.0200). This shows that being female, having the HLA-B*5703, B*5801 and B*8100 is protective against worse conditions of the disease. In addition, although the HLA-B*4201 effect was not significant, its interaction with time was significant causing an increased rate of 0.00224 per unit. Significant three-way interactions that were found are B5702*-592 CA*time, B5802*-592 CA*time, B8100*-592 AA*time, B0801*-1082 GG*time.

Table 20: Solutions for fixed effects for final model log viral load as the response

Effect	Category	Estimate	Std Error	DF	t-value	p-value
Intercept		4.9375	0.2536	407	19.47	<.0001
Gender		-0.2260	0.1038	1723	-2.18	0.0296
Time		-0.00014	0.000110	1723	-1.27	0.2026
B0801		-0.05149	0.3856	1723	-0.13	0.8938
B4201		-0.08356	0.2948	1723	-0.28	0.7769
B4403		-0.1306	0.1237	1723	-1.06	0.2915
B5702		-0.5919	0.3766	1723	-1.57	0.1162
B5703		-0.8926	0.2258	1723	-3.95	<.0001
B5801		-0.2944	0.1414	1723	-2.08	0.0375
B5802		0.03731	0.1184	1723	0.31	0.7528
B8100		-0.4491	0.1543	1723	-2.91	0.0037
-592 Genotype	AA	0.1940	0.1683	1723	1.15	0.2493
-592 Genotype	CA	0.07266	0.1021	1723	0.71	0.4770
-1082 Genotype	AA	0.2355	0.1846	1723	1.28	0.2021
-1082 Genotype	AG	0.1110	0.1829	1723	0.61	0.5442
Time*B4201		0.000224	0.000077	1723	2.89	0.0039
Time*B5703		0.000263	0.000138	1723	1.90	0.0574
B0801*-1082 Genotype	AA	-0.3589	0.4341	1723	-0.83	0.4085
B0801*-1082 Genotype	AG	0.2832	0.4216	1723	0.67	0.5018
B4201*-1082 Genotype	AA	-0.4685	0.3292	1723	-1.42	0.1549
B4201*-1082 Genotype	AG	0.02753	0.3306	1723	0.08	0.9336
B4403*-592 Genotype*Time	AA	-0.00019	0.000251	1723	-0.75	0.4545
B4403*-592 Genotype*Time	CA	0.000147	0.000128	1723	1.14	0.2527
B4403*-592 Genotype*Time	CC	0.000194	0.000122	1723	1.59	0.1116
B4403*-592 Genotype*Time	AA	0.000072	0.000138	1723	0.52	0.6019
B4403*-592 Genotype*Time	CA	-0.00012	0.000092	1723	-1.30	0.1923
B5702*-592 Genotype*Time	AA	-0.00006	0.000319	1723	-0.18	0.8539
B5702*-592 Genotype*Time	CA	0.001509	0.000346	1723	4.36	<.0001
B5702*-592 Genotype*Time	CC	0.002434	0.002025	1723	1.20	0.2296
B5802*-592 Genotype*Time	AA	0.000192	0.000430	1723	0.45	0.6558
B5802*-592 Genotype*Time	CA	0.000364	0.000138	1723	2.65	0.0082
B5802*-592 Genotype*Time	CC	0.000088	0.000125	1723	0.71	0.4808
B8100*-592 Genotype*Time	AA	0.001009	0.000491	1723	2.05	0.0401
B8100*-592 Genotype*Time	CA	0.000522	0.000323	1723	1.61	0.1068
B8100*-592 Genotype*Time	CC	-0.00002	0.000187	1723	-0.08	0.9330
B0801*-1082 Genotype*Time	AA	0.000210	0.000181	1723	1.16	0.2466
B0801*-1082 Genotype*Time	AG	-0.00014	0.000154	1723	-0.93	0.3517
B0801*-1082 Genotype*Time	GG	-0.00055	0.000234	1723	-2.33	0.0200
B0801*-1082 Genotype*Time	AA	0.000024	0.000127	1723	0.19	0.8532
B0801*-1082 Genotype*Time	AG	-4.09E-6	0.000121	1723	-0.03	0.9732
B8100*-1082 Genotype*Time	AA	-0.00060	0.000308	1723	-1.94	0.0529
B8100*-1082 Genotype*Time	AG	-0.00010	0.000314	1723	-0.32	0.7471

4.2.6 Diagnostic Analysis for the Model Including Random Intercept

The histogram of the residuals shown in Figure 27 does not indicate any skewness. The normal quantile plot of the residuals shows that there is only a minor systematic departure from a straight line at the tails. This indicates a fairly adequate model. Although these are fairly similar diagnostics to those from the model with no random effects, the current model is preferred since the fit statistics indicated that this model is a better fit to the data.

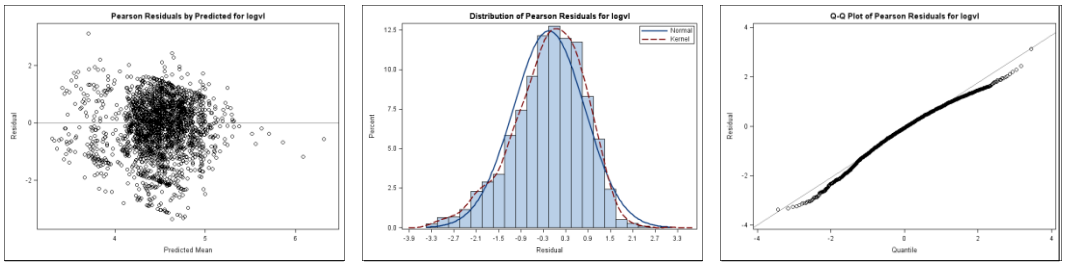


Figure 27: Model diagnostics for log viral load with random intercept

Chapter 5

Generalized Estimating Equations

Generalized estimating equations (GEEs), which were first introduced by Liang and Zeger (1986), are an extension to the theory of generalized linear models (GLMs) and are based on the concept of quasi-likelihood as opposed to direct likelihood. Hence we firstly consider the theory of the generalized linear models. Generalized estimating equations are also referred to as marginal or population averaged models. The focus in these models is on the population averaged effects rather than on the individual specific effects. The crucial characteristics about GEEs is that they have the capacity to account for correlation in clustered data such as repeated measurements from the same experimental unit.

5.1 Generalized Linear Models

The generalized linear model (GLM) extends the ordinary regression model to allow for non-normal response variables. There are three components that make up the GLM and are defined as the random component, the systematic component, and the link function (Agresti, 2002).

An important characteristic of generalized linear models is that they assume independent observations (McCullagh & Nelder, 1989, p. 21). These observations form the first component of the GLM, which are the random component. The response variable is Y with independent observations $(y_1, y_2, \dots , y_N)$ from a distribution that falls in the exponential family, which has probability density function or mass function given by

$$f(y_i;\theta_i,\Phi) = \exp\left\{\frac{[y_i\theta_i - b(\theta_i)]}{a(\Phi)} + c(y_i,\Phi)\right\} \tag{5.1}$$

The parameter Φ is known as the dispersion parameter and θ_i is known as the natural parameter.

The systematic component of a GLM relates a function η_i of explanatory variables by making use of the linear model given by

$$\eta_i = \sum_j \beta_j x_{ij} \tag{5.2}$$

where x_{ij} is the value of predictor where $j = 1,2 \ldots, p$ and $i = 1,2, \ldots, N$. The linear combination of explanatory variables to the right of equation (5.2) is known as the linear predictor. The coefficient of an intercept in the model can be accommodated by letting one $x_{ij} = 1$ for all i.

The random and systematic components are connected by the third component of the GLM, the link function. By letting $\mu_i = E(Y_i)$, where $i = 1,2, \ldots, N$, the model links μ_i to η_i by $\eta_i = g(\mu_i)$. The link function denoted by g is a monotonic differentiable function.

This implies that g links $E(Y_i)$ to the explanatory variables through the equation

$$g(\mu_i) = \sum_j \beta_j x_{ij.} \tag{5.3}$$

For the ordinary regression model where the response Y is normally distributed, we have the identity link $g(\mu) = \mu$, with $\eta_i = \mu_i$ (Agresti, 2002, pp. 116-117). The simplest assumption in the specification of the model stated in equation (5.3) is to consider the outcomes within a unit as independent, which is clearly not a realistic assumption.

5.2 Generalized Estimation Equations

5.2.1 Introduction

GEE's allow for repeated measurements, $y_i = (y_{i1}, y_{i2}, \dots, y_{in_i})$ where n_i can vary for each subject such as in the case of unbalanced data (Agresti, 2002, p. 467). The vector y_i is a realization of a random vector Y_i.

An alternate to maximum likelihood estimation is the quasi-likelihood estimation. This estimation assumes only a mean-variance relationship, instead of assuming a specific distribution for Y_i. It has a link function and linear predictor of the usual GLM form, but instead of assuming a distributional type for Y_i it assumes only $var(Y_i) = v(\mu_i)$ for some chosen variance function v (Agresti, 2002, p. 149). The quasi-likelihood method specifies a model for $\mu = E(Y)$. The variance function $v(\mu)$ describes how $var(Y)$ is dependent on the mean μ. Unlike the GLM approach, the GEE model applies to the marginal distribution for each Y_i (Agresti, 2002, p. 467).

When dealing with repeated measurements, we have a multivariate response for Y_i, such that $Y_i = (Y_{i1}, \dots, Y_{in_i})'$. Where it is assumed that the Y_{it}, $t = 1,2, \dots, n_i$, are independent, then a GLM approach can be used. Also, $\mu_i = (\mu_{i1}, \dots, \mu_{in_i})'$, where $\mu_{it} = E(Y_{it})$. Each outcome, Y_{it}, for subject i at time t may have a different number n_i of responses (Zeger, Liang, & Albert, 1988, p. 1049). Let the vector of explanatory variables values for Y_{it} be denoted by x_{it} with dimension $p \times 1$ (Liang & Zeger, 1986, p. 13). For the link function g, the linear predictor of the model can be written as

$$\eta_{it} = g(\mu_{it}) = x_{it}'\beta. \tag{5.4}$$

(Agresti, 2002, p. 472).

This is the model for the marginal distribution at each t rather than the joint distribution. The matrix of predictor values for subject i is then denoted by X_i, and has dimension $n_i \times p$ and x_{it}' is row t of X_i.

The probability mass function for y_{it} is given by

$$f(y_{it};\theta_{it},\Phi) = \exp\left\{\frac{[y_{it}\theta_{it} - b(\theta_{it})]}{\Phi} + c(y_{it},\Phi)\right\}. \quad (5.5)$$

(Agresti, 2002)

If Φ is known, the probability mass function is that of the natural exponential family with natural parameter θ_{it}.

Similar to the case of GLM's, we have

$$\mu_{it} = E(Y_{it}) = b'(\theta_{it}) \quad (5.6)$$

$$v(\mu_{it}) = var(Y_{it}) = \Phi b''(\theta_{it}) \quad (5.7)$$

5.2.2 Correlated Data

Consider the case where the observations in the vector y_i are not necessarily independent. Let the assumed or working correlation matrix of Y_i be $R(\alpha)$, which depends on a vector of unknown parameters α.

Let

$$b_i(\theta) = \left(b(\theta_{i1}), \dots, b(\theta_{iT_i})\right) \quad (5.8)$$

and B_i denote a diagonal matrix with diagonal elements $b_i''(\theta)$.

Then

$$V_i = B_i^{\frac{1}{2}} R(\alpha) B_i^{\frac{1}{2}} \Phi \quad (5.9)$$

is the working covariance matrix for Y_i (Agresti, 2002, p. 472; Liang & Zeger, 1986, p. 15).

In GLM's the score equation is given by

$$U(\beta) = \sum_i \frac{\partial \mu_i}{\partial \beta} \, v(\mu_i)^{-1}(y_i - \mu_i) = 0 \tag{5.10}$$

(Agresti, 2002, p. 470; Ramroop, 2008, p. 92).

For the case when the outcome Y_i is multivariate, equation (5.10) then becomes

$$\begin{aligned} U(\beta) &= \sum_i \sum_t \frac{\partial \mu_{it}}{\partial \beta} v(\mu_{it})^{-1}(y_i - \mu_i) \\ &= \sum_i \frac{\partial \mu_i'}{\partial \beta} V_i^{-1}(y_i - \mu_i) \\ &= \sum_i F_i' V_i^{-1}(y_i - \mu_i) = 0 \end{aligned} \tag{5.11}$$

where $F_i = \frac{\partial \mu_i'}{\partial \beta}$, $\mu_i = E(Y_i)$ (Liang & Zeger, 1986, p. 15; Agresti, 2002, p. 474; Ramroop, 2008, p. 93) and V_i is given in equation (5.9).

The GEE estimator $\hat{\beta}$ is the solution to equation (5.11), where

$$\mu_i = \mu_i(\beta) = g^{-1}(x_i'\beta) \tag{5.12}$$

(Agresti, 2002, pp. 472-473).

In the case where the identity link assumption holds with $\mu_i = X_i\beta$, we then have

$$U(\beta) = \sum_i X_i' v(\mu_{it})^{-1}(y_i - X_i\beta) = 0. \tag{5.13}$$

(Ramroop, 2008, p. 93). Thus data which requires the adoption of the identity link becomes a special case of the equation presented in (5.11).

5.3 Measures of Goodness of Fit

Since the GEE is strictly not likelihood based a challenge in its utility is the type of statistics to use for model selection. As an approximation the Quasi-likelihood under the Independence Model Criteria (QIC) can be used to choose the best subset of covariates. The QIC can be approximated by the QIC_u. The algebraic expressions showing how the QIC and the QIC_u are defined below in Table 21, with $\hat{\mu} = g^{-1}\left(x'\beta_R\right)$,

Table 21: Goodness of fit criteria for GEEs

QIC	$-2Q\left(g^{-1}(x'\beta_R)\right) + 2\,trace\left(A_I^{-1}V_{MS,R}\right)$
QIC_u	$-2Q\left(g^{-1}(x'\beta_R)\right) + 2d$

where Q stands for quasi-likelihood, $g^{-1}()$ is the inverse link function, A_I is the variance matrix which is obtained assuming the independence model, $V_{MS,R}$ is the modified sandwich estimate of variance from the model that uses R as the hypothesised correlation structure and *d* is the number of parameters in the model (Pan, 2001, p. 122; Hardin & Hilbe, 2003, pp. 139-142).

5.4 Application of Generalized Estimating Equations to Sinikithemba Data

In this chapter we will use a backward procedure using square root CD4+ count and log viral load separately as the response variables under the generalized estimating equations methodology. This will be done using the statistical software SAS version 9.2 (SAS Institute Inc., Cary, NC, USA). Using the 'proc genmod' procedure.

5.5 Square Root CD4+ Count as the Response

As with the linear mixed model analysis, square root CD4+ count will be used as the response variable in the analysis. The explanatory variables that will be included are the same variables that were included for the linear mixed model analysis, i.e. gender (male or female), HLA-B type (B*0801, B*1503, B*1510, B*4201, B*4403, B*5702, B*5703, B*5801, B*5802, B*8100) and IL-10 genotypes (-592 genotype and -1082 genotype). The -592 genotype includes AA, AC and CC. The -1082 genotype includes AA, AG and GG.

The model building strategy will be done in two steps. Firstly, the covariance structure of observations from the same individuals needs to be chosen. Unlike with linear mixed models, if the covariance structure is chosen incorrectly, the estimates of the model parameters will still be correct (Agresti, 2002). The covariance structures that will be utilized include Independence (IND), Compound Symmetry (CS) and Autoregressive (1) (AR(1)). Secondly, the mean structure must be chosen.

5.5.1 Covariance Structure

In order to choose the best covariance structure we make use of the full mean model. This includes all the HLA-B types, both -592 and -1082 genotypes, time, the interactions between each of the HLA-B types and the IL-10 genotypes, the interactions between the HLA-B types and time, the interactions between the IL-10 genotypes and time and the three-way interactions between the HLA-B types, the IL-10 genotypes and time. The covariance structures are then compared using the empirical standard error estimates and the model-based standard error estimates. The empirical standard error estimates are based on the actual variation of the data, whereas the model-based standard error estimates are based on the estimated correlation from the chosen model. The covariance structure with the least difference between the standard errors is the best covariance structure (Hanley, et al., 2003). For this data, the compound symmetry covariance structure was found to be the best and will therefore be used in this analysis. Using the compound symmetry covariance structure, the score statistics were found as

seen in Table 1C in Appendix C. This will be used to select the best model structure in the following section.

5.5.2 Mean Structure

Using the Score Statistics in Table 1C in Appendix, the covariate that is the least significant is dropped from the model. This new model is then compared to the original model using the QIC_u values, with the smaller QIC_u being better. This is then repeated until all the non-significant terms are dropped from the model (unless the QIC_u value is smaller, in which case the term will remain in the model). Note that any term that is included in an interaction term needs to be included as an individual or main effect term. If one of these terms has already been dropped, it will still need to be included in the final model. Following this procedure the final model can be seen in Table 22. From this table it can be seen that time is significantly associated with square root CD4+ count (p-value<0.0001) at the 5% significance level. HLA-B*5703 (p-value=0.0252) and B*5802 (p-value=0.0240) are significantly associated with square root CD4+ count. The interactions between HLA-B*1503 and time (p-value=0.0105), B*5703 and time (p-value=0.0283) and B*5801 and time (p-value=0.0050) are significantly associated with square root CD4+ count. The interactions between HLA-B*5802 and the -592 genotype (p-value=0.0208), and the interaction between HLA B*4403, the -1082 genotype and time (p-value=0.0274) are significantly associated with square root CD4+ count.

Table 22: Score statistics for type III GEE analysis with square root CD4+ count as response for final model

Source	DF	Chi-Square	Pr > Chi-Sq
Time	1	28.73	<0.0001
B0801	1	0.10	0.7558
B1503	1	2.65	0.1035
B1510	1	3.53	0.0603
B4201	1	0.05	0.8313
B4403	1	1.51	0.2192
B5702	1	0.31	0.5806
B5703	1	5.01	0.0252
B5801	1	0.07	0.7934
B5802	1	5.10	0.0240
B8100	1	3.59	0.0583
-592 Genotype	2	2.11	0.3485
-1082 Genotype	2	4.11	0.1283
B0801*Time	1	1.25	0.2629
B1503*Time	1	6.54	0.0105
B4403*Time	1	0.13	0.7142
B5702*Time	1	0.70	0.4037
B5703*Time	1	4.81	0.0283
B5801*Time	1	7.87	0.0050
B8100*Time	1	0.91	0.3388
-592 Genotype*Time	2	0.36	0.8353
-1082 Genotype*Time	2	2.01	0.3657
B5802*-592 Genotype	2	7.75	0.0208
B5702*-1082 Genotype	1	0.14	0.7108
B4201*-1082 Genotype*Time	3	4.39	0.2223
B4403*-1082 Genotype*Time	2	7.19	0.0274

At a 5% level of significance, the significance of model effects can be assessed in Table 23. The estimate for time shows that on average at the population level mean square root CD4+ count decreases at a rate of 0.0026 units per day (p-value<0.0001) as time elapses. Individuals who possess B*5802 have a mean square root CD4+ count 3.0974 units lower (p-value<0.0001) than individuals who do not have this HLA-B type. However, those with B*5703 have a mean square root CD4+ count 2.9480 units higher mean square root CD4+ count (p-value=0.0086) than individuals without B*5703. Considering the -1082 loci, individuals with the AA genotype have a mean square root CD4+ count 1.7561 units lower than individuals with the GG genotype. The interaction between HLA-B*1503 and time shows that when HLA-B*1503 is present, the mean square root CD4+ count increases at an additional rate of 0.0009 units per day than when HLA-B*1503 is absent (p-value=0.0079). The coefficient of the interaction

between HLA-B*4403 and time is -0.0016. This indicates that when HLA-B*4403 is present, the average square root CD4+ count decreases at an additional rate of 0.0016 units per day than when HLA-B*4403 is absent (p-value=0.0012). The interaction between HLA-B*5703 and time shows that when HLA-B*5703 is present, the mean square root CD4+ count increases at an additional rate of 0.0015 units per day than when HLA-B*5703 is absent (p-value=0.0018). Focusing on the -592 loci, the estimate of the interaction between HLA-B*5802 and CA genotype shows that the average square root CD4+ count is 3.0255 units more for individuals who have this HLA-B type than for individuals who do not (p-value=0.0082). Looking at the -1082 loci, the estimate of the interaction between HLA-B*4201 and AA genotype, the mean square root CD4+ count for such individuals is 0.0010 units higher when B*4201 is present, than when it is absent (p-value=0.0401). The estimate of the interaction between HLA-B*4403 and the AA genotype illustrates that the mean square root CD4+ count increases at a rate of 0.0025 units per day (p-value=0.0001) and when the -1082 AG genotype is present the mean square root CD4+ count increases at a rate of 0.0017 units per day (p-value=0.0073) for individuals who possess HLA-B*4403 compared to those who do not possess this HLA-B type.

Table 23: Analysis of GEE parameter estimates empirical standard error estimates for square root CD4+ count as response

Parameter	Category	Estimate	Std Error (Model-Based)	95% Confidence Limits		Z Value	Pr > |Z|
Intercept		20.8451	0.8114 (0.8382)	19.2548	22.4354	25.69	<0.0001
Time		-0.0026	0.0005 (0.0003)	-0.0035	-0.0017	-5.50	<0.0001
B0801		-0.2916	0.9380 (0.6968)	-2.1301	1.5469	-0.31	0.7559
B1503		-1.2047	0.7329 (0.6263)	-2.6411	0.2318	-1.64	0.1002
B1510		-1.4024	0.7411 (0.6407)	-2.8548	0.0501	-1.89	0.0584
B4201		-0.1379	0.6453 (0.6192)	-1.4026	1.1268	-0.21	0.8308
B4403		1.0163	0.8158 (0.6838)	-0.5826	2.6152	1.25	0.2128
B5702		-0.5142	4.0158 (2.3746)	-8.3850	7.3566	-0.13	0.8981
B5703		2.9480	1.1221 (1.1807)	0.7487	5.1473	2.63	0.0086
B5801		-0.2360	0.9028 (0.8117)	-2.0055	1.5334	-0.26	0.7937
B5802		-3.0974	0.7514 (0.8878)	-4.5702	-1.6246	-4.12	<0.0001
B8100		1.7568	0.9022 (0.8400)	-0.0114	3.5251	1.95	0.0515
-592 Genotype	AA	0.4018	1.3203 (0.9398)	-2.1860	2.9895	0.30	0.7609
-592 Genotype	CA	-0.7309	0.6907 (0.5847)	-2.0846	0.6228	-1.06	0.2899
-1082 Genotype	AA	-1.7561	0.8435 (0.8448)	-3.4092	-0.1029	-2.08	0.0373
-1082 Genotype	AG	-1.2293	0.7573 (0.8174)	-2.7135	0.2549	-1.62	0.1045
B0801*Time		0.0005	0.0004 (0.0002)	-0.0004	0.0013	1.15	0.2520
B1503*Time		0.0009	0.0003 (0.0002)	0.0002	0.0015	2.66	0.0079
B4403*Time		-0.0016	0.0005 (0.0006)	-0.0025	-0.0006	-3.25	0.0012
B5702*Time		0.0011	0.0010 (0.0006)	-0.0009	0.0030	1.06	0.2877
B5703*Time		0.0015	0.0006 (0.0004)	0.0003	0.0026	2.52	0.0116
B5801*Time		0.0013	0.0004 (0.0003)	0.0005	0.0021	3.12	0.0018
B8100*Time		0.0005	0.0005 (0.0003)	-0.0005	0.0015	0.97	0.3334
-592 Genotype*Time	AA	-0.0003	0.0005 (0.0003)	-0.0013	0.0007	-0.60	0.5485
-592 Genotype*Time	CA	-0.0001	0.0004 (0.0002)	-0.0008	0.0006	-0.26	0.7981
-1082 Genotype*Time	AA	-0.0002	0.0005 (0.0004)	-0.0013	0.0008	-0.43	0.6662
-1082 Genotype*Time	AG	0.0002	0.0005 (0.0003)	-0.0008	0.0012	0.43	0.6690
B5802*-592 Genotype	AA	-2.2947	2.9472 (2.2333)	-8.0712	3.4818	-0.78	0.4362
B5802*-592 Genotype	CA	3.0255	1.1445 (1.2316)	0.7824	5.2685	2.64	0.0082
B5702*-1082 Genotype	AA	-2.0668	5.4739 (4.0078)	-12.7955	8.6619	-0.38	0.7058
B4201*-1082 Genotype*Time	AA	0.0010	0.0005 (0.0003)	0.0000	0.0020	2.05	0.0401
B4201*-1082 Genotype*Time	AG	0.0003	0.0005 (0.0003)	-0.0008	0.0013	0.47	0.6359
B4201*-1082 Genotype*Time	GG	0.0008	0.0007 (0.0005)	-0.0006	0.0023	1.16	0.2450
B4403**-1082 Genotype*Time	AA	0.0025	0.0006 (0.0007)	0.0012	0.0038	3.83	0.0001
B4403*-1082 Genotype*Time	AG	0.0017	0.0006 (0.0007)	0.0005	0.0030	2.68	0.0073

5.6 Log Viral Load as the Response

We now consider log viral load as the response, and perform the analysis as done previously for square root CD4+ count.

5.6.1 Covariance Structure

Once again we considered the saturated or the full model. The covariance structures were then compared using the empirical standard error estimates and the model-based standard error estimates. The covariance structure with the least difference between the standard errors was taken as the best covariance structure. The compound symmetry covariance structure was found to be the best and will therefore be used in this analysis. This suggests that the correlation between measurements remains constant and is not dependent on how far apart the measurements are. The score statistics are shown in Table 2C in Appendix C. These statistics will be used to determine the best mean structure.

5.6.2 Mean Structure

Terms that are not significant were excluded or dropped systematically from the model and compared to the previous model using the QIC_u. The model with the smallest QIC_u is kept. The final model for log viral load is shown in Table 24. At a 5% significance level, gender is shown to be significantly associated with log viral load (p-value=0.0234), and time is shown not to be significantly associated with log viral load (p-value=0.4749). HLA-B*5703 (p-value=0.0164), the interaction between HLA-B*0801 and time (p-value=0.0383) and the interaction between HLA-B*1503, the -592 genotype and time (p-value=0.0446) were found to be significantly associated with log viral load.

Table 24: Score statistics for type III GEE analysis with viral load as response for final model

Source	DF	Chi-Square	Pr > Chi Sq
Gender	1	5.14	0.0234
Time	1	0.51	0.4749
B0801	1	0.06	0.8010
B1503	1	0.71	0.3998
B1510	1	2.12	0.1453
B4201	1	3.75	0.0527
B4403	1	0.06	0.8115
B5703	1	5.76	0.0164
B5801	1	1.47	0.2258
B5802	1	1.51	0.2199
B8100	1	3.54	0.0599
-592 Genotype	2	2.72	0.2570
-1802 Genotype	2	1.01	0.6036
B0801*Time	1	4.29	0.0383
B1503*Time	1	0.04	0.8466
B1510*Time	1	1.40	0.2360
B4201*Time	1	3.82	0.0506
B5703*Time	1	1.76	0.1848
B5801*Time	1	0.12	0.7277
B5802*Time	1	2.48	0.1153
B8100*Time	1	1.15	0.2825
-592 Genotype*Time	2	2.54	0.2813
-1082 Genotype*Time	2	2.70	0.2598
B1503*-592 Genotype	2	6.22	0.0446
B5802*-592 Genotype	2	1.44	0.4871
B0801*-1082 Genotype*Time	2	5.90	0.0522
B1510*-1082 Genotype*Time	2	3.26	0.1956

From Table 25 and it can be seen that females have mean log viral load 0.2345 units lower than males (p-value=0.0212). Individuals with HLA-B*5703 have a mean log viral load that is 0.7628 units lower than individuals without this HLA-B type (p-value=0.0029). The effect of HLA-B*4201 is marginally significant with an estimate of -0.2247 (p-value=0.0503). The interaction between HLA-B*4201 and time shows that when this HLA-B type is present, individuals have a mean log viral load that increases at a rate of 0.0002 units more per day compared to individuals who do not have HLA-B*4201 (p-value=0.0423). Looking at the -592 loci, when HLA-B*1503 and the CA genotype are present, such individuals have mean log viral loads 0.5387 units higher compared to those with the CC genotype was present (p-value=0.0114). Now considering the -1082 loci, when HLA-B*0801 and the AA genotype are present, the individuals mean log viral load increases at a rate of 0.0006 units more per day than when B*0801 was absent (p-value=0.0315).

Table 25: Analysis of GEE parameter estimates empirical standard error estimates for log viral load as response

Parameter	Genotype	Estimate	Std Error	95% Confidence Level		Z Value	Pr > |Z|
Intercept		4.8916	0.2490 (0.2526)	4.4035	5.3797	19.64	<0.0001
Gender		-0.2345	0.1018 (0.1057)	-0.4339	-0.0350	-2.30	0.0212
Time		0.0000	0.0001 (0.0001)	-0.0002	0.0002	0.21	0.8343
B0801		0.0328	0.1303 (0.1327)	-0.2225	0.2881	0.25	0.8012
B1503		-0.1172	0.1614 (0.1662)	-0.4336	0.1992	-0.73	0.4678
B1510		0.1947	0.1328 (0.1248)	-0.0656	0.4549	1.47	0.1427
B4201		-0.2247	0.1148 (0.1188)	-0.4497	0.0003	-1.96	0.0503
B4403		-0.0305	0.1273 (0.1258)	-0.2799	0.2190	-0.24	0.8109
B5703		-0.7628	0.2565 (0.2270)	-1.2655	-0.2601	-2.97	0.0029
B5801		-0.2083	0.1693 (0.1545)	-0.5401	0.1235	-1.23	0.2184
B5802		0.1110	0.1566 (0.1736)	-0.1959	0.4179	0.71	0.4783
B8100		-0.3452	0.1779 (0.1599)	-0.6938	0.0034	-1.94	0.0523
-592 Genotype	AA	0.0279	0.2177 (0.2144)	-0.3987	0.4545	0.13	0.8980
-592 Genotype	CA	-0.0000	0.1273 (0.1226)	-0.2495	0.2494	-0.00	0.9997
-1082 Genotype	AA	0.1477	0.1563 (0.1617)	-0.1587	0.4540	0.94	0.3447
-1082 Genotype	AG	0.1429	0.1507 (0.1563)	-0.1525	0.4383	0.95	0.3431
B0801*Time		-0.0005	0.0003 (0.0002)	-0.0011	0.0000	-1.92	0.0550
B1503*Time		-0.0000	0.0001 (0.0001)	-0.0002	0.0001	-0.19	0.8469
B1510*Time		-0.0006	0.0006 (0.0003)	-0.0017	0.0005	-1.06	0.2910
B4201*Time		0.0002	0.0001 (0.0001)	0.0000	0.0003	2.03	0.0423
B5703*Time		0.0002	0.0001 (0.0001)	-0.0001	0.0004	1.47	0.1427
B5801*Time		-0.0001	0.0002 (0.0001)	-0.0004	0.0003	-0.35	0.7279
B5802*Time		0.0002	0.0001 (0.0001)	-0.0000	0.0004	1.67	0.0949
B8100*Time		-0.0001	0.0001 (0.0001)	-0.0004	0.0001	-1.13	0.2578
-592 Genotype*Time	AA	0.0002	0.0001 (0.0001)	-0.0001	0.0004	1.48	0.1377
-592 Genotype*Time	CA	0.0000	0.0001 (0.0001)	-0.0002	0.0002	0.11	0.9085
-1082 Genotype*Time	AA	-0.0001	0.0001 (0.0001)	-0.0004	0.0001	-1.18	0.2386
-1082 Genotype*Time	AG	0.0000	0.0001 (0.0001)	-0.0002	0.0003	0.39	0.6996
B1503*-592 Genotype	AA	0.1228	0.3119 (0.3270)	-0.4884	0.7340	0.39	0.6938
B1503*-592 Genotype	CA	0.5387	0.2130 (0.2338)	0.1213	0.9561	2.53	0.0114
B5802*-592 Genotype	AA	0.3754	0.3856 (0.4364)	-0.3804	1.1312	0.97	0.3303
B5802*-592 Genotype	CA	-0.1261	0.2317 (0.2348)	-0.5803	0.3282	-0.54	0.5865
B0801*-1082 Genotype*Time	AA	0.0006	0.0003 (0.0002)	0.0001	0.0012	2.15	0.0315
B0801*-1082 Genotype*Time	AG	0.0003	0.0003 (0.0002)	-0.0003	0.0009	1.09	0.2270
B1510*-1082 Genotype*Time	AA	0.0006	0.0006 (0.0003)	-0.0005	0.0017	1.12	0.2624
B1510*-1082 Genotype*Time	AG	0.0002	0.0006 (0.0003)	-0.0010	0.0014	0.34	0.7322

5.7 Summary

A main advantage of using generalized estimating equations instead of other methods is that the correlated longitudinal data does not need to be normally distributed in order to fit the linear model. An assumed covariance structure is used for the response variables, with a specified variance function and a pair wise correlation pattern. This is done without assuming any specific multivariate distribution (Agresti, 2002, p. 467). Unlike when using linear mixed models, if the covariance structure is chosen incorrectly, the estimates of the model parameters will still be consistent (Agresti, 2002; Ghisletta & Spini, 2004). The GEE method offers two possible algorithms for estimating variance. The first is the model-based estimator. This requires the correct working covariance structure to be chosen. The second is the empirical estimator. This is purely based on the data, and is therefore the most trustworthy, since the model parameters will still be accurate even if the wrong covariance structure is used (Ghisletta & Spini, 2004). Since GEEs are an extension of GLMs for correlated data, they can be applied to a wide range of outcome variables. When dealing with incomplete data, GEEs are applicable when observations are missing completely at random. They are also advantageous since GEEs can be used to analyze unbalanced data easily (Ghisletta & Spini, 2004). A limitation of the GEE method is that it does not have a likelihood function, because the joint distribution of elements in Y_i cannot be completely specified (Agresti, 2002). If missing data is present, this method will assume the data is missing completely at random, and therefore bias will be introduced if this assumption is violated. Another limitation is that GEEs require a large sample size to ensure unbiased estimation (Ghisletta & Spini, 2004).

Chapter 6
Conclusion

Despite the intensive research about HIV, its pathogenesis is still not adequately understood. A thorough understanding of HIV and factors that influence or affect progression of the disease is required in order to help design control and treatment strategies to limit the further spread of the virus in the population. Modelling provides us with a means or tools to understand and predict the progression of the disease better. In this dissertation study, longitudinal data from the Sinikithemba cohort was used to understand the contribution of immunogenetic parameters (namely HLA-B types and IL-10 genotypes) to disease pathogenesis using biomarkers such as CD4+ counts and viral loads. In addition the data was further used to investigate the interaction between HLA-B alleles and IL-10 genotypes, and further to understand their interaction over time.

This study consisted of 450 individuals followed longitudinally over a 5 year period. With CD4+ counts being taken every 3 months and viral loads taken every 6 months, this resulted in having 4016 CD4+ count observations and 2007 viral load observations after excluding individuals who had missing data or who commenced antiretroviral treatment. CD4+ count and viral load did not conform to the normality assumptions and therefore needed to be transformed. For this reason, square root CD4+ count and log viral load were used for the analysis. From the exploratory analysis the most common HLA-B types were identified. These, in addition to the HLA-B types that were found to be significant in past research, enabled us to choose 10 HLA-B types to use in conducting the analysis. The IL-10 genotypes were compared using the Kruskal Wallis Test and pairwise comparisons. Looking at the -592 loci, it was found that there is a significant difference in median CD4+ counts across the CA and the CC genotypes, and a significant difference in median log viral load across the AA and CC genotypes, as

well as across the CA and CC genotypes. Examining the -1082 loci, it was evident that there was a significant difference in median CD4+ counts across all combinations of these genotypes (AA and AG, AA and GG, AG and GG), and a significant difference between median log viral loads across the AA and AG genotypes, as well as across the AG and GG genotypes.

Longitudinal data is commonly characterized with missing values, inconsistent timed observations and incomplete data which are problems that need to be accounted for during the analysis. Linear mixed models are useful for dealing with these problems. The maximum likelihood method and the restricted maximum likelihood method are two estimation methods applied when using linear mixed models. These likelihood based methods automatically allow the handling of missing data under the MAR assumption. Which of the two methods to use needs to be chosen correctly in order to model and assess effects in the data accurately. Another important consideration when dealing with longitudinal data is the type of covariance structure to use for observations from the same unit. Choosing a covariance structure that is too restrictive will invalidate inferences, while a structure that it too simple will lead to inefficient estimation and poor estimation of standard errors (Verbeke, et al., 1998). Although the data used in this dissertation is normally distributed after transformation, another method that was looked at is generalized estimating equations for marginal effects. This procedure is an extension of generalized linear models to account for correlated data, but can be used for non normal data. Another difference compared to LMMs is that GEEs do not take into account the random effects that the LMM does. Thus GEEs and linear or generalized linear mixed models are conceptually different. Model diagnostics is an important part of longitudinal analysis. An analysis of residuals is used to determine whether a model is adequate and to indicate any outliers. This can be done graphically by constructing scatter-plots of the residuals after model fitting.

From the linear mixed model analysis, we found that the model which most adequately fitted square root CD4+ count was that which excluded all random effects. It was established that the best model for log viral load was the model including a random intercept. This implies that there was no significant heterogeneity between individuals

for square root CD4+ count, but strong individual to individual variability at baseline for log viral load. It was shown that individuals who possess HLA-B*1510 and B*5802 have lower mean square root CD4+ counts than individuals who do not possess these HLA-B types. This suggests that these HLA-B types and this genotype are associated with a faster progression of HIV. Individuals that have the AA genotype on the -1082 loci or have B*5801 and the CA genotype on the -592 loci also have significantly lower mean square root CD4+ counts. This is consistent with the results found by Shin et al. (2000), where it was shown that individuals who carry the -592A promoter allele have an accelerated progression to AIDS. Individuals who possess HLA-B*5802 and the CA genotype on the -592 loci were however found to be significantly associated with a higher mean square root CD4+ count and therefore the presence of such genetic factors is considered as a relative controller of HIV. Individuals that have HLA-B*5703 have higher mean square root CD4+ counts and have lower mean log viral loads than individuals who do not have B*5703. This implies that B*5703 is associated with a slower progression of HIV. Other HLA-B types that are associated with a slower progression of HIV include HLA-B*5801 and B*8100 as individuals with these HLA-B types had significantly lower mean log viral loads. This is consistent with the results found by Kiepiela et al. (2004), where HLA-B*57 and B*5801 were found to be significantly associated with a lower viral load and therefore considered to be associated with a slower progression of HIV. HLA-B*5703 was found to be significantly associated with a higher mean square root CD4+ count and HLA-B*4201 was found to be significantly associated with a lower mean viral load over time, illustrating that individuals with either B*5703 and B*4201 have a slower progression of HIV than individuals without these HLA-B types. Including time in all the interactions and looking at the -592 loci, individuals with both HLA-B*1503 and the CA genotype, B*4201 and the CA genotype, B*4403 and the CA genotype, B*4403 and the CC genotype, B*5802 and the AA genotype, B*8100 and the AA genotype or B*8100 and the CA genotype were all significantly associated with a lower mean square root CD4+ count. Still considering time in each interaction and looking at the -592 loci, individuals with both HLA-B*5702 and the CA genotype, B*5802 and the CA genotype or B*8100 and the AA genotype are significantly associated with a higher mean log viral load. Looking at the -1082 loci and including time in the interactions, individuals with both

HLA-B*0801 and the AA genotype, B*0801 and the AG genotype, B*1510 and the AA genotype, B*1510 and the AG genotype or B*5801 and the AG genotype are also significantly associated with a lower mean square root CD4+ count. This implies that individuals with any of these interactions progress faster to HIV over time. In contrast, looking at the -592 loci and including time, an individual with both HLA-B*0801 and the CA genotype, B*0801 and the CC genotype, B*1510 and the CC genotype, B*5801 and the CA genotype or B*5801 and the CC genotype have a significantly higher mean square root CD4+ count. Looking at the -1082 loci, individuals with both HLA-B*4403 and the AA genotype, B*4403 and the AG genotype, B*8100 and the AA genotype or B*8100 and the AG genotype have significantly higher mean square root CD4+ counts, and individuals with both B*0801 and the CA genotype have significantly lower mean log viral loads. This indicates that individuals with a combination of these immunogenetic parameters will have a slower progression of HIV over time.

Using generalized estimating equations similar results were found. However, here it was found that HLA-B*4201 is associated with a lower mean log viral load, and B*8100 is associated with a higher mean square root CD4+ count compared to that for individuals who do not possess these HLA-B types. HLA-B*4201 and B*8100 are therefore associated with a slower progression of the disease. Individuals who possess HLA-B*4403 have mean square root CD4+ counts that decrease significantly faster than individuals who do not have B*4403 over time. The two-way interaction between B*1503 and the CA -592 genotype implies that these genetic factors are associated with a faster progression of HIV since possessing them indicates a significant association with a higher log viral load. The three-way interaction between B*4201, the AA -1082 genotype and time shows that an individual with these factors are significantly associated with an increasing mean square root CD4+ count over time, and can therefore be considered as controllers of the disease. The three-way interaction between HLA-B*0801, the AA -1082 genotype and time was found to be significantly associated with a significantly higher mean log viral load compared to individuals without B*0801 and therefore can be considered to be associated with a faster progression of HIV. These factors were not found to be significant when using linear mixed models.

Both methods found that time was significantly associated with square root CD4+ count, but not significantly associated with log viral load. This may be due to time being looked at in days. The mean square root CD4+ count decreases at a significant rate per day, whereas the mean log viral load does not. Gender is found to be significantly associated with log viral load, but not with square root CD4+ count. In this dissertation females were found to have a 0.2260 lower mean log viral load than men (when using linear mixed models) and 0.2345 lower mean log viral load than men (when using generalized estimating equations). This is consistent with results found in Gender Difference in Viral Load (1999). Results found from these studies found that women had viral loads that were approximately half those of men (Traub, 2002/2003). Since higher viral loads are associated with quicker progression to AIDS and poorer outcomes, this suggests that women may be able to suppress HIV infection better than men. However, although there are these differences in viral load, women still progress to AIDS at the same rate as men. This indicates that women progress to AIDS at much lower viral loads than men (Traub, 2002/2003).

Since this was a study conducted over a long period of time, there were many drop outs. These may have been due to individuals not returning for visits, or in some cases death. Another issue related to this was that the time periods between visits were considerably varied between individuals. This complicated the analysis since the data was not balanced. Due to the number of covariates and interactions included in the application, not all of the HLA-B types were included in the analysis. Thus there may have been other significant HLA-B types that were ignored. During the analysis, individuals were looked at over time only as HIV positive, not defined by which stage of HIV they were at. This causes complications since the individuals CD4+ counts and viral loads may have stabilized, whereas others may not. This data does not however permit us to look at these stages separately.

The findings from this study highlight the additive effects of two separate genetic loci in HIV pathogenesis. Previous studies have generally concentrated on analysis of only one specific gene or genetic loci to HIV pathogenesis and yet these genes likely interact and together contribute to the clinical outcome of HIV-1 disease. Therefore, this study

has extended our understanding of how combined genetic factors may contribute to HIV-1 disease progression. Further studies will need to be conducted to understand the mechanisms by which these genes may modulate HIV pathogenesis. However, previous studies have suggested that certain "protective" HLA alleles may modulate their effect by presenting conserved regions of the virus to the immune system as part of the immune recognition pathway, and immune escape in these epitopes is then associated with reduced replicative capacity of the virus (Wright et al, 2010 and Wright et al, 2011), thus conferring a clinical benefit to the infected patient. On the other hand, the IL-10 gene codes for interleukin-10, an important immunoregulatory cytokine. This cytokine has been shown to downregulate the expression of HLA molecules that are involved in antigen presentation (Matsuda, Salazar et al, 1994). IL-10 has been proposed to affect HIV disease pathogenesis according to the stage or phase of infection (Naicker et al, 2009). This is because IL-10 may dampen essential immune responses required to control the virus (Brooks et al, 2006) but it may also act as an anti-inflammatory agent that reduces immune activation and thus leading to better clinical outcomes (Naicker et al, 2009). Better understanding of how HLA and IL-10 interact to modulate disease pathogenesis may lead to better design of vaccines and immunotherapies which are urgently needed to stem the spread of AIDS. This study offers strong statistical genetics evidence for interaction between these two loci and should form a basis for functional and mechanistic studies.

This dissertation can be extended in many different aspects. Since this is a real-life study, issues such as missing data can be looked at in more depth. Methods for dealing with missing data such as simple and multiple imputations can be investigated. Only ten of the HLA-B types were chosen, therefore there are more that should be analyzed with respect to this research. Similar analysis could be conducted, but taking into account the stages of HIV that each individual is at when enrolled into the study, and comparing these groups accordingly. Another way to analyze this data would be to look at fitting a generalized linear mixed model using a Poisson distribution since CD4+ count is in fact a count.

References

Agresti, A. (2002). *Categorical Data Analysis* (Second ed.). John Wiley & Sons.

Averting HIV and AIDS. (2009). *History and science of HIV & AIDS*. Retrieved October 18, 2010, from http://www.avert.org/aids-timeline.htm

Brooks, D. L. (2008). IL-10 blockade facilitates DNA vaccine-induced T cell responses and enhances clearance of persistent virus infection. *The Journal of Experimental Medicine , 205*, 533-541.

David G Brooks, D. T. (2006). Interleukin-10 determines viral clearance or persistence in vivo. *Nature , 12*, 1301-1309.

Diggle, P. J., Heagerty, P., Liang, K.-Y., & Zeger, S. L. (2002). *Analysis of Longitudinal Data* (2nd Edition ed.). Oxford University Press.

Eskdale, J., Gallagher, G., Verweij, C. L., Keijsers, V., & Westerdorp, R. G. (1998). Interleukin 10 secretion in relation to human IL-10. *Proceedings of the National Academy of Sciences of the United States of America , 95*, 9465–9470.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied Longitudinal Analysis.* John Wiley & Sons.

Gender Difference in Viral Load. (1999, April). (Project Inform) Retrieved from The Body - Complete HIV/AIDS Resource: http://www.thebody.com/content/art5841.html

Ghisletta, P., & Spini, D. (2004). An Introduction to Generalized Estimating Equations and an Application to Assess Selectivity Effects in a Longitudinal Study on Very Old Individuals. *Journal of Educational and Behavioral Statistics , 29* (4), 421-437.

Hanley, J. A., Negassa, A., Edwardes, M. D., & Forrester, J. E. (2003). Statistical Analysis of Correlated Data Using Generalized Estimating Equations: An Orientation. *American Journal of Epidemiology , 157* (4), 364-375.

Hardin, J. W., & Hilbe, J. M. (2003). *Generalized Estimating Equations.* Chapman & Hall/CRC.

Jones, R. (1993). *Longitudinal Data with Serial Correlation: A State-space Approach.* Chapman & Hall.

Kiepiela, P., Leslie, A., Honeyborne, I., Ramduth, D., Thobakgale, C., Chetty, S., et al. (2004). Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. *Nature , 432*, 769-774.

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

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal Data Analysis using Generalized Linear Models. *Biometrika , 73*, 13-22.

Liu, C., Cao, D., Chen, P., & Zagar, T. (2007). RANDOM and REPEATED statements - How to Use Them to Model the Covariance Structure in Proc Mixed. Indianapolis: Eli Lilly & Company.

Matsuda, M. S. (December). Interleukln 10 Pretreatment Protects Target Cells from Tumor- and Allo-specific Cytotoxic T Cells and Downregulates HLA Class I Expression. *J. Exp. Med. , 180*, 2371-2376.

McCullagh, P., & Nelder, J. (1989). *Generalized Linear Models* (Second ed.). Chapman and Hall.

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

Muller, K. E., & Stewart, P. W. (2006). *Linear Model Theory: Univariate, Mulitvariate, and Mixed Models.* New Jersey: Wiley-Interscience.

Naicker, D. W. (2009). *Interleukin-10 Promoter Polymorphisms Influence HIV-1 Susceptibility and Primary HIV-1 Pathogenesis.* The Infectious Diseases Society of America.

National Department of Health South Africa. (2010). The South African Antiretroviral Treatment Guidelines .

Ngo, L., & Brand, R. (2002). Model Selection in Linear Mixed Effects Models Using SAS PROC MIXED.

Pan, W. (2001). Akaike's Information Criterian in Generalized Estimating Equations. *Biometrics, 57* (1), 120-125.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood.* Oxford University Press.

Photini Kipiela, A. J. (2004). Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. *Nature , 432*, 769-774.

Ramroop, S. (2008). Analysis of Longitudinal Binary Data: An Application to a Disease Process. PhD Thesis, The University of Kwa-Zulu Natal, 2008.

Shin, H. D., Winkler, C., Stephens, J. C., Bream, J., Young, H., & Goedert, J. J. (2000). Genetic restriction of HIV-1 pathogenesis to AIDS by promoter alleles of IL10. *Proceedings of the National Academy of Sciences of the United States of America , 97*, 14467–14472.

Terrazzano, G. R. (2000). HLA class I antigen downregulation by interleukin (IL)-10 is predominantly governed by NF-kB in the short termand by TAP1¹2 in the long term. *Tissue Antigens , 55*, 326–332.

Traub, O. (2002/2003). *Battle of the Sexes: Gender-Based Differences in HIV Viral Load*. Retrieved from The Body - Complete HIV/AIDS Resource: http://www.thebody.com/contnent/art1825.html

UNAIDS, Joint United Nations Programme on HIV/AIDS. (2009). *Report on the global AIDS 2009.*

Verbeke, G., & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* Springer.

Verbeke, G., Lesaffre, E., & Brant, L. J. (1998). The Detection of Residual Serial Correlation in Linear Mixed Models. *Statistics in Medicine , 17*, 1391-1402.

Werner, L. (2009). *Modelling Acute HIV Infection Using Longitudinally Measured Biomarker Data Including Informative Drop-out.* Masters Dissertation, The University of Kwa-Zulu Natal, 2009.

Wright, J. (2010). Gag-Protease-Mediated Replication Capacity in HIV-1 Subtype C Chronic Infection: Associations with HLA Type and Clinical Parameters. *Journal of Virology , 84* (20), 10820-10831.

Wright, J. (2011). Influence of Gag-Protease-Mediated Replication Capacity on Disease Progression in Individuals Recently Infected with HIV-1 Subtype C. *Journal of Virology , 85*, 3996-4006.

Zeger, S. L., Liang, K.-Y., & Albert, P. S. (1988). Models for Longitudinal Data: A Generalized Estimating Equation Approach. *Biometrics* , 1049-1060.

Appendix A: SAS Code

Linear Mixed Models

********** SQUARE ROOT CD4+ COUNT **********

```
/*FULL MODEL (MODEL IN TABLE 18)*/
proc mixed data=skfinal method=ml;
class pid gen B0801 B1503 B1510 B4201 B4403 B5702 B5703 B5801 B5802
B8100 F2 F3 timec;
model sqrtcd4 = gen B0801 B1503 B1510 B4201 B4403 B5702 B5703 B5801
B5802 B8100 F2 F3 time B0801*time B1503*time B1510*time B4201*time
B4403*time B5702*time B5703*time B5801*time B5802*time B8100*time
F2*time F3*time B0801*F2 B1503*F2 B1510*F2 B4201*F2 B4403*F2 B5702*F2
B5703*F2 B5801*F2 B5802*F2 B8100*F2 B0801*F3 B1503*F3 B1510*F3
B4201*F3 B4403*F3 B5703*F3 B5801*F3 B5802*F3 B8100*F3 B0801*F2*time
B1503*F2*time B1510*F2*time B4201*F2*time B4403*F2*time B5702*F2*time
B5703*F2*time B5801*F2*time B5802*F2*time B8100*F2*time B0801*F3*time
B1503*F3*time B1510*F3*time B4201*F3*time B4403*F3*time B5703*F3*time
B5801*F3*time B5802*F3*time B8100*F3*time / solution;
repeated timec/ subject=pid type=cs;
run;


/*FINAL MODEL WITH NO RANDOM EFFECTS (MODEL IN TABLE 19 AND 20)*/
proc mixed data=skfinal method=ml;
class pid gen B0801 B1503 B1510 B4201 B4403 B5702 B5703 B5801 B5802
B8100 F2 F3 timec;
model sqrtcd4 = B0801 B1503 B1510 B4201 B4403 B5702 B5703 B5801 B5802
B8100 F2 F3 time B1503*time B1510*time B5702*time B5703*time
B5801*time B5802*time B8100*time F2*time B1510*F2 B5801*F2 B5802*F2
B5801*F3 B0801*F2*time B1503*F2*time B1510*F2*time B4201*F2*time
B4403*F2*time B5702*F2*time B5801*F2*time B5802*F2*time B8100*F2*time
B0801*F3*time B1510*F3*time B4201*F3*time B4403*F3*time B5801*F3*time
B8100*F3*time /solution;
repeated timec/ subject=pid type=cs;
run;



/*FINAL MODEL WITH RANDOM INTERCEPT*/
proc mixed data=skfinal method=ml maxiter=10000;
class pid gen B0801 B1503 B1510 B4201 B4403 B5702 B5703 B5801 B5802
B8100 F2 F3 timec;
model sqrtcd4 = B0801 B1503 B1510 B4201 B4403 B5702 B5703 B5801 B5802
B8100 F2 F3 time B1503*time B1510*time B5702*time B5703*time
B5801*time B5802*time B8100*time F2*time B1510*F2 B5801*F2 B5802*F2
B5801*F3 B0801*F2*time B1503*F2*time B1510*F2*time B4201*F2*time
B4403*F2*time B5702*F2*time B5801*F2*time B5802*F2*time B8100*F2*time
```

```
B0801*F3*time B1510*F3*time B4201*F3*time B4403*F3*time B5801*F3*time
B8100*F3*time /solution;
random intercept/ subject=pid type=un;
repeated timec/ subject=pid type=cs;
run;
/*FINAL MODEL WITH RANDOM INTERCEPT AND SLOPE*/
proc mixed data=skfinal method=ml maxiter=10000;
class pid gen B0801 B1503 B1510 B4201 B4403 B5702 B5703 B5801 B5802
B8100 F2 F3 timec;
model sqrtcd4 = B0801 B1503 B1510 B4201 B4403 B5702 B5703 B5801 B5802
B8100 F2 F3 time B1503*time B1510*time B5702*time B5703*time
B5801*time B5802*time B8100*time F2*time B1510*F2 B5801*F2 B5802*F2
B5801*F3 B0801*F2*time B1503*F2*time B1510*F2*time B4201*F2*time
B4403*F2*time B5702*F2*time B5801*F2*time B5802*F2*time B8100*F2*time
B0801*F3*time B1510*F3*time B4201*F3*time B4403*F3*time B5801*F3*time
B8100*F3*time /solution;
random intercept time/ subject=pid type=un;
repeated timec/ subject=pid type=cs;
run;
```

********** LOG VIRAL LOAD **********

```
/*FULL MODEL WITH NO RANDOM EFFECTS (MODEL IN TABLE 24)*/
proc mixed data=skfinal method=ml;
class pid gen B0801 B1503 B1510 B4201 B4403 B5702 B5703 B5801 B5802
B8100 F2 F3 timec;
model logvl = gen B0801 B1503 B1510 B4201 B4403 B5702 B5703 B5801
B5802 B8100 F2 F3 time B0801*time B1503*time B1510*time B4201*time
B4403*time B5702*time B5703*time B5801*time B5802*time B8100*time
F2*time F3*time B0801*F2 B1503*F2 B1510*F2 B4201*F2 B4403*F2 B5702*F2
B5703*F2 B5801*F2 B5802*F2 B8100*F2 B0801*F3 B1503*F3 B1510*F3
B4201*F3 B4403*F3 B5703*F3 B5801*F3 B5802*F3 B8100*F3 B0801*F2*time
B1503*F2*time B1510*F2*time B4201*F2*time B4403*F2*time B5702*F2*time
B5703*F2*time B5801*F2*time B5802*F2*time B8100*F2*time B0801*F3*time
B1503*F3*time B1510*F3*time B4201*F3*time B4403*F3*time B5703*F3*time
B5801*F3*time B5802*F3*time B8100*F3*time ;
repeated timec/ subject=pid type=sp(pow)(sampledate);
run;


/*FINAL MODEL WITH NO RANDOM EFFECTS (MODEL IN TABLE 25 AND 26)*/
proc mixed data=skfinal method=ml;
class pid gen B0801 B1503 B1510 B4201 B4403 B5702 B5703 B5801 B5802
B8100 F2 F3 timec;
model logvl = gen B1503 B4201 B5702 B5703 B5802 B8100 F2 F3 time
B4201*time
B1503*F2 B8100*F2 B4201*F3 B5802*F3 B5702*F2*time B5703*F2*time
/solution;
repeated timec/ subject=pid type=sp(pow)(sampledate);
run;
```

```
/*FINAL MODEL WITH RANDOM INTERCEPT (MODEL IN TABLE 29 AND 30)*/
proc mixed data=skfinal method=ml;
class pid gen B0801 B1503 B1510 B4201 B4403 B5702 B5703 B5801 B5802
B8100 F2 F3 timec;
model logvl = gen B1503 B4201 B5702 B5703 B5802 B8100 F2 F3 time
B4201*time
B1503*F2  B8100*F2  B4201*F3  B5802*F3  B5702*F2*time  B5703*F2*time
/solution;
random intercept / subject=pid type=un;
repeated timec/ subject=pid ;
run;



/*FINAL MODEL WITH RANDOM INTERCEPT AND SLOPE*/
proc mixed data=skfinal method=ml;
class pid gen B0801 B1503 B1510 B4201 B4403 B5702 B5703 B5801 B5802
B8100 F2 F3 timec;
model logvl = gen B1503 B4201 B5702 B5703 B5802 B8100 F2 F3 time
B4201*time
B1503*F2  B8100*F2  B4201*F3  B5802*F3  B5702*F2*time  B5703*F2*time
/solution;
random intercept time / subject=pid type=un;
repeated  timec/ subject=pid;
run;
```

Generalized Estimating Equations

********** SQUARE ROOT CD4+ COUNT **********

```
/*FULL MODEL (MODEL IN TABLE 31)*/
PROC GENMOD data=skfinal;
class pid gen B0801 B1503 B1510 B4201 B4403 B5702 B5703 B5801 B5802
B8100 F2 F3 timec;
model sqrtcd4 = gen B0801 B1503 B1510 B4201 B4403 B5702 B5703 B5801
B5802 B8100 F2 F3 time B0801*time B1503*time B1510*time B4201*time
B4403*time B5702*time B5703*time B5801*time B5802*time B8100*time
F2*time F3*time B0801*F2 B1503*F2 B1510*F2 B4201*F2 B4403*F2 B5702*F2
B5703*F2  B5801*F2  B5802*F2  B8100*F2  B0801*F3  B1503*F3  B1510*F3
B4201*F3  B4403*F3    B5702*F3  B5703*F3  B5801*F3  B5802*F3  B8100*F3
B0801*F2*time B1503*F2*time B1510*F2*time B4201*F2*time B4403*F2*time
B5702*F2*time B5703*F2*time B5801*F2*time B5802*F2*time B8100*F2*time
B0801*F3*time B1503*F3*time B1510*F3*time B4201*F3*time B4403*F3*time
B5702*F3*time B5703*F3*time B5801*F3*time B5802*F3*time B8100*F3*time
/ type3;
repeated subject=pid / withinsubject=timec type=cs covb corrw modelse;
run;
```

```
/*FINAL MODEL (MODEL IN TABLE 32 AND 33)*/
PROC GENMOD data=skfinal descending;
class pid gen B0801 B1503 B1510 B4201 B4403 B5702 B5703 B5801 B5802
B8100 F2 F3 timec;
model sqrtcd4 = time B0801 B1503 B1510 B4201 B4403 B5702 B5703 B5801
B5802 B8100 F2 F3 B0801*time B1503*time B4403*time B5702*time
B5703*time B5801*time B8100*time F2*time F3*time B5802*F2 B5702*F3
B4201*F3*time B4403*F3*time / type3;
repeated subject=pid / withinsubject=timec type=cs covb corrw modelse;
run;
```

********** **LOG VIRAL LOAD** **********

```
/*FULL MODEL (MODEL IN TABLE 34)*/
PROC GENMOD data=skfinal;
class pid gen B0801 B1503 B1510 B4201 B4403 B5702 B5703 B5801 B5802
B8100 F2 F3 timec;
model logvl = gen B0801 B1503 B1510 B4201 B4403 B5702 B5703 B5801
B5802 B8100 F2 F3 time B0801*time B1503*time B1510*time B4201*time
B4403*time B5702*time B5703*time B5801*time B5802*time B8100*time
F2*time F3*time B0801*F2 B1503*F2 B1510*F2 B4201*F2 B4403*F2 B5702*F2
B5703*F2 B5801*F2 B5802*F2 B8100*F2 B0801*F3 B1503*F3 B1510*F3
B4201*F3 B4403*F3 B5702*F3 B5703*F3 B5801*F3 B5802*F3 B8100*F3
B0801*F2*time B1503*F2*time B1510*F2*time B4201*F2*time B4403*F2*time
B5702*F2*time B5703*F2*time B5801*F2*time B5802*F2*time B8100*F2*time
B0801*F3*time B1503*F3*time B1510*F3*time B4201*F3*time B4403*F3*time
B5702*F3*time B5703*F3*time B5801*F3*time B5802*F3*time B8100*F3*time
/ type3;
repeated subject=pid / withinsubject=timec type=cs covb corrw modelse;
run;
```

```
/*FINAL MODEL (MODEL IN TABLE 35 AND 36)*/
PROC GENMOD data=skfinal;
class pid gen B0801 B1503 B1510 B4201 B4403 B5702 B5703 B5801 B5802
B8100 F2 F3 timec;
model logvl = gen B0801 B1503 B1510 B4201 B4403 B5703 B5801 B5802
B8100 F2 F3 time B0801*time B1503*time B1510*time B4201*time
B5703*time B5801*time B5802*time B8100*time F2*time F3*time B1503*F2
B5802*F2 B0801*F3*time B1510*F3*time / type3;
repeated subject=pid / withinsubject=timec type=cs covb corrw modelse;
run;
```

Appendix B: Linear Mixed Models

Table 1B: Type III tests of fixed effects for square root CD4+ count as response for full model

Effect	Num DF	Den DF	F Value	Pr>F
Gender	1	373	1.23	0.2688
Time	1	3583	22.47	<.0001
B0801	1	373	0.03	0.8665
B1503	1	373	0.94	0.3322
B1510	1	373	0.05	0.8219
B4201	1	373	0.15	0.6944
B4403	1	373	3.48	0.0628
B5702	1	373	0.00	0.9536
B5703	1	373	7.53	0.0064
B5801	1	373	1.03	0.3101
B5802	1	373	4.74	0.0301
B8100	1	373	1.24	0.2666
-592 Genotype	2	373	0.16	0.8536
-1082 Genotype	2	373	0.60	0.5484
B0801*Time	1	3583	0.07	0.7957
B1503*Time	1	3583	1.88	0.1699
B1510*Time	1	3583	13.62	0.0002
B4201*Time	1	3583	0.92	0.3372
B4403*Time	1	3583	2.13	0.1449
B5702*Time	1	3583	16.44	<.0001
B5703*Time	1	3583	12.17	0.0005
B5801*Time	1	3583	16.81	<.0001
B5802*Time	1	3583	8.35	0.0039
B8100*Time	1	3583	4.62	0.0316
-592 Genotype*Time	2	3583	4.86	0.0078
-1082 Genotype*Time	2	3583	0.57	0.5658
B0801*-592 Genotypes	2	373	0.07	0.9323
B1503*-592 Genotypes	2	373	0.71	0.4901
B1510*-592 Genotypes	2	373	1.59	0.2058
B4201*-592 Genotypes	2	373	1.07	0.3431
B4403*-592 Genotypes	2	373	0.72	0.4888
B5702*-592 Genotypes	2	373	0.11	0.8977
B5703*-592 Genotypes	1	373	0.00	0.9451
B5801*-592 Genotypes	2	373	2.69	0.0689
B5802*-592 Genotypes	2	373	2.94	0.0539
B8100*-592 Genotypes	2	373	0.00	0.9998
B0801*-1082 Genotypes	2	373	1.90	0.1514
B1503*-1082 Genotypes	2	373	1.77	0.1726
B1510*-1082 Genotypes	2	373	0.20	0.8200
B4201*-1082 Genotypes	2	373	0.19	0.8231
B4403*-1082 Genotypes	2	373	1.12	0.3267
B5703*-1082 Genotypes	2	373	0.97	0.3809
B5801*-1082 Genotypes	2	373	1.32	0.2691
B5802*-1082 Genotypes	2	373	0.01	0.9902
B8100*-1082 Genotypes	2	373	0.28	0.7596
B0801*-592 Genotypes*Time	2	3583	1.91	0.1485
B1503*-592 Genotypes*Time	2	3583	4.62	0.0099
B1510*-592 Genotypes*Time	2	3583	11.55	<.0001

B4201*-592 Genotypes*Time	2	3583	10.75	<.0001
B4403*-592 Genotypes*Time	2	3583	9.78	<.0001
B5702*-592 Genotypes*Time	2	3583	17.06	<.0001
B5703*-592 Genotypes*Time	1	3583	0.43	0.5102
B5801*-592 Genotypes*Time	2	3583	7.43	0.0006
B5802*-592 Genotypes*Time	2	3583	2.53	0.0795
B8100*-592 Genotypes*Time	2	3583	21.46	<.0001
B0801*-1082 Genotypes*Time	2	3583	5.63	0.0036
B1503*-1082 Genotypes*Time	2	3583	0.89	0.4106
B1510*-1082 Genotypes*Time	2	3583	2.12	0.1208
B4201*-1082 Genotypes*Time	2	3583	2.58	0.0763
B4403*-1082 Genotypes*Time	2	3583	10.05	<.0001
B5703*-1082 Genotypes*Time	2	3583	0.32	0.7283
B5801*-1082 Genotypes*Time	2	3583	5.56	0.0039
B5802*-1082 Genotypes*Time	2	3583	0.65	0.5202
B8100*-1082 Genotypes*Time	2	3583	14.58	<.0001

Table 2B: Type III tests of fixed effects for log viral load as response for full model

Effect	Num DF	Den DF	F Value	Pr>F
Gender	1	371	5.07	0.0249
Time	1	1696	2.46	0.1169
B0801	1	371	0.04	0.8385
B1503	1	371	0.10	0.7534
B1510	1	371	0.00	0.9806
B4201	1	371	0.42	0.5190
B4403	1	371	0.35	0.5531
B5702	1	371	0.72	0.3964
B5703	1	371	15.93	<.0001
B5801	1	371	2.44	0.1194
B5802	1	371	1.38	0.2401
B8100	1	371	2.56	0.1103
-592 Genotype	2	371	0.94	0.3905
-1082 Genotype	2	371	0.39	0.6774
B0801*Time	1	1696	0.41	0.5196
B1503*Time	1	1696	0.02	0.8904
B1510*Time	1	1696	1.26	0.2613
B4201*Time	1	1696	2.40	0.1215
B4403*Time	1	1696	0.34	0.5611
B5702*Time	1	1696	2.39	0.1222
B5703*Time	1	1696	3.76	0.0528
B5801*Time	1	1696	0.59	0.4416
B5802*Time	1	1696	0.59	0.4428
B8100*Time	1	1696	1.37	0.2415
-592 Genotype*Time	2	1696	1.74	0.1762
-1082 Genotype*Time	2	1696	0.44	0.6448
B0801*-592 Genotypes	2	371	1.73	0.1783
B1503*-592 Genotypes	2	371	0.62	0.5374
B1510*-592 Genotypes	2	371	1.17	0.3113
B4201*-592 Genotypes	2	371	0.90	0.4074
B4403*-592 Genotypes	2	371	0.59	0.5568
B5702*-592 Genotypes	2	371	0.38	0.6832
B5703*-592 Genotypes	1	371	0.16	0.6861
B5801*-592 Genotypes	2	371	0.01	0.9893
B5802*-592 Genotypes	2	371	1.99	0.1384
B8100*-592 Genotypes	2	371	0.42	0.6574
B0801*-1082 Genotypes	2	371	1.96	0.1429
B1503*-1082 Genotypes	2	371	0.09	0.9108
B1510*-1082 Genotypes	2	371	0.15	0.8574
B4201*-1082 Genotypes	2	371	1.08	0.3397
B4403*-1082 Genotypes	2	371	0.17	0.8435
B5703*-1082 Genotypes	2	371	1.55	0.2147
B5801*-1082 Genotypes	2	371	0.34	0.7108
B5802*-1082 Genotypes	2	371	1.05	0.3508
B8100*-1082 Genotypes	2	371	0.16	0.8548
B0801*-592 Genotypes*Time	2	1696	0.05	0.9511
B1503*-592 Genotypes*Time	2	1696	1.45	0.2353
B1510*-592 Genotypes*Time	2	1696	2.07	0.1261
B4201*-592 Genotypes*Time	2	1696	0.91	0.4031
B4403*-592 Genotypes*Time	2	1696	1.40	0.2477
B5702*-592 Genotypes*Time	2	1696	2.63	0.0724

B5703*-592 Genotypes*Time	1	1696	3.83	0.0505
B5801*-592 Genotypes*Time	2	1696	0.12	0.8908
B5802*-592 Genotypes*Time	2	1696	2.02	0.1327
B8100*-592 Genotypes*Time	2	1696	3.22	0.0401
B0801*-1082 Genotypes*Time	2	1696	1.34	0.2629
B1503*-1082 Genotypes*Time	2	1696	0.38	0.6830
B1510*-1082 Genotypes*Time	2	1696	0.50	0.6085
B4201*-1082 Genotypes*Time	2	1696	0.11	0.8920
B4403*-1082 Genotypes*Time	2	1696	0.63	0.5349
B5703*-1082 Genotypes*Time	2	1696	1.39	0.2484
B5801*-1082 Genotypes*Time	2	1696	0.45	0.6408
B5802*-1082 Genotypes*Time	2	1696	0.14	0.8727
B8100*-1082 Genotypes*Time	2	1696	2.58	0.0764

Table 3B: Type III tests of fixed effects for full model with log viral load as the response

Effect	Num DF	Den DF	F Value	Pr>F
Gender	1	1696	5.61	0.0180
Time	1	1696	3.52	0.0609
B0801	1	1696	0.02	0.8972
B1503	1	1696	0.01	0.9040
B1510	1	1696	0.05	0.8300
B4201	1	1696	0.64	0.4244
B4403	1	1696	0.21	0.6463
B5702	1	1696	0.60	0.4402
B5703	1	1696	14.70	0.0001
B5801	1	1696	1.89	0.1698
B5802	1	1696	0.85	0.3558
B8100	1	1696	2.18	0.1399
-592 Genotype	2	1696	0.75	0.4728
-1082 Genotype	2	1696	0.31	0.7347
B0801*Time	1	1696	1.94	0.1635
B1503*Time	1	1696	0.13	0.7236
B1510*Time	1	1696	0.86	0.3531
B4201*Time	1	1696	4.86	0.0276
B4403*Time	1	1696	0.64	0.4254
B5702*Time	1	1696	2.44	0.1188
B5703*Time	1	1696	5.41	0.0201
B5801*Time	1	1696	0.24	0.6238
B5802*Time	1	1696	0.32	0.5705
B8100*Time	1	1696	2.27	0.1317
-592 Genotype*Time	2	1696	2.85	0.0583
-1082 Genotype*Time	2	1696	0.30	0.7420
B0801*-592 Genotypes	2	1696	1.92	0.1466
B1503*-592 Genotypes	2	1696	0.56	0.5717
B1510*-592 Genotypes	2	1696	0.96	0.3815
B4201*-592 Genotypes	2	1696	0.71	0.4930
B4403*-592 Genotypes	2	1696	0.71	0.4922
B5702*-592 Genotypes	2	1696	0.24	0.7899
B5703*-592 Genotypes	1	1696	0.08	0.7715
B5801*-592 Genotypes	2	1696	0.03	0.9682
B5802*-592 Genotypes	2	1696	1.85	0.1569
B8100*-592 Genotypes	2	1696	0.31	0.7305
B0801*-1082 Genotypes	2	1696	2.23	0.1075
B1503*-1082 Genotypes	2	1696	0.11	0.8945
B1510*-1082 Genotypes	2	1696	0.19	0.8303
B4201*-1082 Genotypes	2	1696	0.87	0.4189
B4403*-1082 Genotypes	2	1696	0.16	0.8544
B5703*-1082 Genotypes	2	1696	1.41	0.2451
B5801*-1082 Genotypes	2	1696	0.41	0.6643
B5802*-1082 Genotypes	2	1696	1.40	0.2466
B8100*-1082 Genotypes	2	1696	0.26	0.7703
B0801*-592 Genotypes*Time	2	1696	1.24	0.2888
B1503*-592 Genotypes*Time	2	1696	1.87	0.1546
B1510*-592 Genotypes*Time	2	1696	0.95	0.3870
B4201*-592 Genotypes*Time	2	1696	2.71	0.0667
B4403*-592 Genotypes*Time	2	1696	2.21	0.1102
B5702*-592 Genotypes*Time	2	1696	6.25	0.0020
B5703*-592 Genotypes*Time	1	1696	5.01	0.0253
B5801*-592 Genotypes*Time	2	1696	0.14	0.8660
B5802*-592 Genotypes*Time	2	1696	3.37	0.0346
B8100*-592 Genotypes*Time	2	1696	3.74	0.0240

B0801*-1082 Genotypes*Time	2	1696	3.11	0.0449
B1503*-1082 Genotypes*Time	2	1696	1.12	0.3274
B1510*-1082 Genotypes*Time	2	1696	1.87	0.1539
B4201*-1082 Genotypes*Time	2	1696	0.68	0.5048
B4403*-1082 Genotypes*Time	2	1696	0.52	0.5971
B5703*-1082 Genotypes*Time	2	1696	1.87	0.1543
B5801*-1082 Genotypes*Time	2	1696	0.40	0.6679
B5802*-1082 Genotypes*Time	2	1696	0.01	0.9930
B8100*-1082 Genotypes*Time	2	1696	3.14	0.0434

Appendix C: Generalized Estimating Equations

Table 1C: Score statistics for type III GEE analysis with square root CD4+ count as response for full model

Source	DF	Chi-Square	Pr > Chi-Sq
Gender	1	1.20	0.2726
Time	1	11.13	0.0008
B0801	1	0.02	0.8881
B1503	1	1.17	0.2788
B1510	1	0.02	0.8989
B4201	1	0.31	0.5801
B4403	1	4.46	0.0347
B5702	1	0.00	0.9458
B5703	1	4.29	0.0383
B5801	1	1.86	0.1726
B5802	1	5.32	0.0210
B8100	1	1.67	0.1958
-592 Genotype	2	0.35	0.8408
-1082 Genotype	2	2.47	0.2903
B0801*Time	1	0.10	0.7573
B1503*Time	1	2.20	0.1380
B1510*Time	1	1.89	0.1687
B4201*Time	1	0.47	0.4939
B4403*Time	1	0.24	0.6275
B5702*Time	1	1.08	0.2981
B5703*Time	1	4.53	0.0333
B5801*Time	1	5.64	0.0176
B5802*Time	1	2.80	0.0945
B8100*Time	1	0.01	0.9303
-592 Genotype*Time	2	3.44	0.1791
-1082 Genotype*Time	2	0.77	0.6801
B0801*-592 Genotype	2	0.12	0.9428
B1503*-592 Genotype	2	1.20	0.5498
B1510*-592 Genotype	2	2.27	0.3210
B4201*-592 Genotype	2	2.32	0.3140
B4403*-592 Genotype	2	1.42	0.4918
B5702*-592 Genotype	1	1.37	0.2414
B5703*-592 Genotype	1	0.01	0.9219
B5801*-592 Genotype	2	5.36	0.0687
B5802*-592 Genotype	2	5.96	0.0509
B8100*-592 Genotype	2	0.00	0.9997
B0801*-1082 Genotype	2	4.55	0.1026
B1503*-1082 Genotype	2	4.07	0.1306
B1510*-1082 Genotype	2	0.41	0.8127
B4201*-1082 Genotype	2	1.00	0.6063
B4403*-1082 Genotype	2	3.11	0.2109
B5703*-1082 Genotype	2	2.32	0.3131
B5801*-1082 Genotype	2	4.22	0.1209

B5802*-1082 Genotype	2	0.06	0.9697
B8100*-1082 Genotype	2	0.51	0.7745
B0801*-592 Genotype*Time	2	0.95	0.6205
B1503*-592 Genotype*Time	2	2.57	0.2769
B1510*-592 Genotype*Time	2	2.40	0.3008
B4201*-592 Genotype*Time	2	4.36	0.1128
B4403*-592 Genotype*Time	2	4.96	0.0846
B5702*-592 Genotype*Time	1	1.06	0.3030
B5703*-592 Genotype*Time	1	0.10	0.7464
B5801*-592 Genotype*Time	2	3.99	0.1358
B5802*-592 Genotype*Time	2	1.50	0.4730
B8100*-592 Genotype*Time	2	8.53	0.0140
B0801*-1082 Genotype*Time	2	3.49	0.1743
B1503*-1082 Genotype*Time	2	0.83	0.6595
B1510*-1082 Genotype*Time	2	1.75	0.4169
B4201*-1082 Genotype*Time	2	1.33	0.5151
B4403*-1082 Genotype*Time	2	10.93	0.0042
B5703*-1082 Genotype*Time	2	0.55	0.7604
B5801*-1082 Genotype*Time	2	3.09	0.2128
B5802*-1082 Genotype*Time	2	0.32	0.8502
B8100*-1082 Genotype*Time	2	8.49	0.0143

Table 2C: Score statistics for type III GEE analysis with log viral load as response for full model

Source	DF	Chi-Square	Pr > Chi-Sq
Gender	1	5.28	0.0216
Time	1	1.54	0.2145
B0801	1	0.02	0.8785
B1503	1	0.00	0.9622
B1510	1	0.06	0.8050
B4201	1	0.97	0.3240
B4403	1	0.19	0.6619
B5702	1	2.10	0.1468
B5703	1	5.18	0.0228
B5801	1	2.30	0.1293
B5802	1	0.46	0.4983
B8100	1	1.95	0.1627
-592 Genotype	2	1.19	0.5519
-1082 Genotype	2	0.83	0.6598
B0801*Time	1	3.07	0.0796
B1503*Time	1	0.32	0.5694
B1510*Time	1	0.75	0.3866
B4201*Time	1	4.95	0.0260
B4403*Time	1	0.58	0.4477
B5702*Time	1	1.09	0.2962
B5703*Time	1	3.46	0.0627
B5801*Time	1	0.20	0.6544
B5802*Time	1	1.22	0.2688
B8100*Time	1	3.00	0.0832
-592 Genotype*Time	2	6.98	0.0305
-1082 Genotype*Time	2	0.93	0.6277
B0801*-592 Genotype	2	4.66	0.0973
B1503*-592 Genotype	2	1.19	0.5529
B1510*-592 Genotype	2	1.39	0.5001
B4201*-592 Genotype	2	1.14	0.5643
B4403*-592 Genotype	2	1.59	0.4507
B5702*-592 Genotype	1	0.07	0.7850
B5703*-592 Genotype	1	0.05	0.8194
B5801*-592 Genotype	2	0.24	0.8849
B5802*-592 Genotype	2	3.29	0.1928
B8100*-592 Genotype	2	0.33	0.8482
B0801*-1082 Genotype	2	5.30	0.0708
B1503*-1082 Genotype	2	0.25	0.8822
B1510*-1082 Genotype	2	0.47	0.7903
B4201*-1082 Genotype	2	1.41	0.4943
B4403*-1082 Genotype	2	0.47	0.7924
B5703*-1082 Genotype	2	4.34	0.1144
B5801*-1082 Genotype	2	1.21	0.5473
B5802*-1082 Genotype	2	3.04	0.2190
B8100*-1082 Genotype	2	0.66	0.7178
B0801*-592 Genotype*Time	2	4.29	0.1173
B1503*-592 Genotype*Time	2	4.25	0.1196
B1510*-592 Genotype*Time	2	0.49	0.7818
B4201*-592 Genotype*Time	2	4.49	0.1057
B4403*-592 Genotype*Time	2	7.01	0.0300
B5702*-592 Genotype*Time	1	1.02	0.3118
B5703*-592 Genotype*Time	1	4.63	0.0314
B5801*-592 Genotype*Time	2	0.61	0.7382
B5802*-592 Genotype*Time	2	6.73	0.0346
B8100*-592 Genotype*Time	2	4.70	0.0955

B0801*-1082 Genotype*Time	2	6.10	0.0474
B1503*-1082 Genotype*Time	2	3.22	0.1995
B1510*-1082 Genotype*Time	2	2.63	0.2689
B4201*-1082 Genotype*Time	2	2.39	0.3021
B4403*-1082 Genotype*Time	2	0.62	0.7321
B5703*-1082 Genotype*Time	2	3.27	0.1953
B5801*-1082 Genotype*Time	2	1.30	0.5214
B5802*-1082 Genotype*Time	2	0.66	0.7202
B8100*-1082 Genotype*Time	2	3.70	0.1570