Modelling Acute HIV Infection Using Longitudinally Measured Biomarker Data Including Informative Drop-out

A thesis presented to

The University of KwaZulu-Natal

in fulfilment of the requirement for the degree

of

Master of Science in Statistics

by

Lise Werner

School of Statistics and Actuarial Science

December 2009

Abstract

Background

Numerous methods have been developed to model longitudinal data. In HIV/AIDS studies, HIV markers, CD4+ count and viral load are measured over time. Informative drop-out and the lower detection limit of viral load assays can bias the results and influence assumptions of the models.

Objective

The objective of this thesis is to describe the evolution of HIV markers in an HIV-1 subtype C acutely infected cohort of women from the CAPRISA 002: Acute Infection Study in Durban, South Africa. They were HIV treatment naïve.

Methods

Various linear mixed models were fitted to both CD4+ count and viral load, adjusting for repeated measurements, as well as including intercept and slope as random effects. The rate of change in each of the HIV markers was assessed using weeks post infection as both a linear effect and piecewise linear effects. Left-censoring of viral load was explored to account for missing data resulting from undetectable measurements falling below the lower detection limit of the assay. Informative drop-out was addressed by using a method of joint modelling in which a longitudinal and survival model were jointly linked using a latent Gaussian process. The progression of HIV markers were described and the effectiveness and usefulness of each modelling procedure was evaluated.

Results

62 women were followed for a median of 29 months post infection (IQR 20-39). Viral load increased sharply by 2.6 log copies/mℓ per week in the first 2 weeks of infection and decreased by 0.4 log copies/mℓ per week the next fortnight. It decreased at a slower rate thereafter. Similarly CD4+ count fell in the first 2 weeks by 4.4 square root cells/$\mu\ell$ per week then recovered slightly only to decrease again. Left-censoring was unnecessary in this acute infection cohort as few viral load measures were below the detection limit and provided no improvement on model fit.

Conclusion

Piecewise linear effects proved to be useful in quantifying the degree at which the HIV markers progress during the first few weeks of HIV infection, whereas modelling time as a linear effect was not very meaningful. Modelling HIV markers jointly with informative drop-out is shown to be necessary to account for the missing data incurred from participants leaving the study to initiate ARV treatment. In ignoring this drop-out, CD4+ count is estimated to be higher than what it actually is.

Declaration 1 - Plagiarism

I, Lise Werner, declare that

1. The research reported in this thesis, except where otherwise indicated, is my original research.
2. This thesis has not been submitted for any degree or examination at any other university.
3. This thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.
4. This thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:
   (a) Their words have been re-written but the general information attributed to them has been referenced
   (b) Where their exact words have been used, then their writing has been placed in italics and inside quotation marks, and referenced.
5. This thesis does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the thesis and in the References sections.

Signed

Lise Werner (Student Number 202518360)	Date
Prof Henry Mwambi	Date

Notes

This thesis, as well as results and analysis related to this thesis, has been presented at various conferences.

- A preliminary analysis on piecewise linear mixed models and a conceptual discussion on joint modelling was presented in an oral presentation at a conference on TB and HIV Modelling (6 to 8 November 2006) at the University of Stellenbosch, hosted by South African Centre for Epidemiological Modelling and Analysis (SACEMA). The title of the presentation was *Joint modelling of CD4+ T-cell counts and HIV-RNA to describe the evolution of HIV markers.*

- The results from piecewise linear mixed models, as well as joint modelling analysis on a simulated data set was presented at the 50th Annual Conference of the South African Statistical Association (SASA) in Muldersdrift (29 October to 2 November 2007) as an oral presentation, titled *Joint Modelling of CD4+ cell counts and HIV-RNA.*

- The analysis on the data used in this thesis, specifically on describing the evolution of the HIV markers, was presented at the UKZN Postgraduate Research Day for the Faculty of Science and Agriculture, hosted by UKZN Westville campus on 25 May 2009. This oral presentation won third prize and was titled *Modelling Acute HIV Infection Using Longitudinally Measured Biomarker Data Including Informative Drop-out.*

- A poster presentation on piecewise linear mixed models and joint modelling, as presented in this thesis, was exhibited at the 30th annual conference of the International Society for Clinical Biostatistics (ISCB) after receiving a conference award for scientists. The conference was hosted by the University of Economics in Prague, Czech Republic (23 to 27 August 2009) and the poster was titled *Exploring CD4+ count and viral load evolution in an acutely infected cohort using joint modelling.*

Acknowledgements

I want to thank Prof Henry Mwambi for his thorough feedback and support. It has been a tough four years, doing this thesis part-time, whilst working full-time; something I could not have done without Prof Mwambi as a supervisor. I'm very appreciative of his careful dedication to my work. I also want to thank CAPRISA for the use of their data and a special thanks for the hard work put in by the CAPRISA 002: Acute Infection Cohort Study Team; without them, this would not have been possible. I'm grateful for the wonderful opportunity of working for CAPRISA and I have learnt so much in these past few years while working in research.

I want to thank my family and friends for their love and support, especially Mom, Dad and Wesley; and also Scott for his encouragement and love. Thanks for believing in me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The Human Immunodeficiency Virus (HIV) has become a global epidemic with an estimated 33 million people (30-36 million) living with HIV in 2007 (UNAIDS, 2008). In 2007, there were a total of 2.7 million new HIV infections and 2 million HIV-related deaths (UNAIDS, 2008). Sub-Saharan Africa has carried the biggest burden and accounts for 67% of those living with HIV and for 75% of AIDS deaths in 2007 (UNAIDS, 2008). Much research has been done and is still ongoing, on various aspects of the disease, including how the virus replicates, what effect this has on the human body and how the virus can be controlled. The most common surrogate biomarkers used for determining clinical progress in someone infected with HIV is viral load and CD4+ count. These HIV markers react differently when someone first gets infected with the virus, compared to chronic infection, and compared to when antiretroviral treatment is taken. Understanding these key markers do not only let health providers monitor a patient's health, but also allows researchers to discover essential information regarding the mechanisms of the virus and the human immune system. In turn this allows researchers to develop HIV vaccines, new efficient treatments and preventative measures such as microbicides and pre-exposure prophylaxis (PrEP).

This thesis aims to describe the evolution of these HIV biomarkers in a cohort of acutely infected women who were followed up longitudinally over time. However, because of the way the virus and the immune system reacts during the initial few weeks of HIV infection, a simple linear regression is not adequate in describing the evolution of the disease over time. The history and biology behind HIV, as well as treatment, diagnosis, different HIV types and the reaction by the human body will be discussed in Section 1.1. Another issue that will be addressed is missing data brought on by informative drop-out. This is study withdrawal by a participant for a reason related to one or more of the responses being modelled.

In the case of the data used in this thesis as part of an ongoing study, women are being followed immediately after infection and they are not yet ready to be initiated on antiretroviral therapy. However, once their body shows signs of being unable to adequately control the virus and their CD4+ count falls below a certain threshold, they are initiated on treatment and are no longer followed up in this specific cohort. Informative drop-out will be accounted for by using a statistical technique called joint modelling, which aims to combine a longitudinal model, describing the CD4+ count or viral load evolution, to a survival model which accounts for time to informative drop-out. This is important, as ignoring informative drop-out can lead to inaccurate statements about the HIV markers, because those who drop out are weighted less in the study. Another problem occurs when there are viral load observations which fall below the detection limit of the assay used to quantify the measurement. This thesis will address this issue by using a parametric approach to apply left-censoring to viral load.

1.1 The Human Immunodeficiency Virus

1.1.1 History

In 1981 the Centre for Disease Control (CDC), based in Atlanta, United States of America (USA), published a report on a pattern of *Pneumocystis carinii pneumonia* that occurred in homosexual men in Los Angeles (CDC, 1981). Around the same time a few aggressive cases of Kaposi Sarcoma, usually a rare and benign cancer of the skin, mouth, gastrointestinal and respiratory tract, was found in young homosexual men in New York (Hymes et al., 1981). Later this would be recognised as Acquired Immune Deficiency Syndrome (AIDS). The CDC had set up a task force to investigate these unusual events, which seemed to appear more and more. By the end of 1981, the first case of AIDS was reported in the United Kingdom (du Bois et al., 1981). It seemed that AIDS was an appropriate name because it was an acquired condition and not inherited. It was a deficiency within the immune system which allowed opportunistic diseases to easily infect the host. It was also seen as a syndrome with various manifestations instead of a single disease. Throughout 1982 and 1983 it became clear that this new disease was spreading through sex with infected people, as well as blood transfusions and injectable drug use. In 1983, French scientist Luc Montagnier and in 1984 American Robert Gallo, isolated what is now known as the Human Immunodeficiency Virus (HIV) (Barre-Sinoussi et al., 1983; Marx, 1984).

After discovering a similar virus in a chimpanzee (Cohen, 2000; Gao et al., 1999), the Simian

Immunodeficiency Virus (SIV), it has been generally accepted that HIV is a descendant of this virus because of the close resemblance in genetic structure and biological properties. Other strains of SIV have been found in other primate species and many theories exist on how the virus could have crossed over to humans (Cohen, 2000; Hooper, 1999; Blancou et al., 2001). One such theory proposes that SIV was transferred after humans hunted and killed infected primates, and the blood from the monkeys or apes came into contact with cuts or wounds from the humans hunting them. The issue remains a controversial one up to the present time.

1.1.2 HIV and the Human Body

The Human Immunodeficiency Virus (HIV) is a retrovirus that infects bodily fluids in humans and remains in the immune cells in these fluids. HIV targets these immune cells in order to replicate, killing them in the process. These immune cells, CD4+ cells and macrophages, play an important role in the body's immune system. The CD4+ cells are mature T helper cells, a type of white blood cell, which expresses a surface protein CD4. T-cells cannot kill infected cells or invading pathogens and without other immune cells, they cannot fight infection in the human body. Their purpose is to activate and direct other immune cells which play a major role in fighting off disease. Macrophages are another type of white blood cell within tissues and they are also known as `eater cells ´since they remove dead cell material and pathogens. They also stimulate other immune cells to respond to the pathogen and are vital to the regulation of immune responses. The CD4+ cells are the primary entry point for HIV into the host. The virus attaches itself to the CD4 receptor via its own surface protein when exposed to the CD4+ cells. The outer membrane of the virus fuses to the cell membrane and the virus enters the CD4+ cell. Every retrovirus has a reverse transcriptase enzyme which copies genetic information from ribonucleic acid (RNA) to deoxyribonucleic acid (DNA). When attached to the CD4+ cell, the virus encodes the enzyme reverse transcriptase and allows a DNA copy to be made from viral RNA. However, this process is prone to error and accounts for the genetic variability that HIV is known for. The DNA molecule is then transported to the nucleus where it is incorporated into human DNA. This proviral DNA can remain dormant for a long time or become active, especially when there is inflammation present. The virus makes use of the host cell to replicate itself and destroys it, crippling the functionality of the immune system. Hence this is why medical professionals rely on the CD4+ count to decide on the state of the immune system and to decide when the patient needs to be initiated on HIV treatment.

There are three main routes of transmission of HIV, i.e. sexual, blood transfusions and mother-to-child transmission. Majority of HIV infections are acquired through unprotected sex with an

infected person. The transmission can occur through infected secretions coming in contact with oral, genital or rectal membranes. Studies have shown that women are significantly more likely to contract HIV through heterosexual intercourse compared to men (Padian, Shiboski and Jewell, 1991; European Study Group on Heterosexual Transmission of HIV, 1992). The presence of other sexually transmitted infections increases the risk of contracting HIV through further sexual contact (Fleming and Wasserheit, 1999).

If infected blood comes into contact with any open wound, HIV will be transmitted. Thus HIV is easily spread through blood transfusions and intravenous drug use where infected blood is involved. HIV transmission between a mother and infant can happen while pregnant, during childbirth, or whilst breastfeeding. However, preventative antiretroviral treatment can be used by the mother to protect the child from getting infected, such as Nevirapine which will be discussed below.

Within a few weeks of infection, there is a high level of replication in the blood that can exceed ten million viral particles per millilitre of blood (Abdool Karim, 2005). This rapid replication of viral particles is followed by a decline of CD4+ cells in the body. However, after a few weeks the body develops its own immune response to the HIV which stops the viral replication and the viral load declines and the number of CD4+ cells increase again to levels which are near normal. Thus infected individuals can remain asymptomatic for many years. However, it has been shown that during this time in which the person is feeling well, the body destroys up to a billion HIV particles and produces up to two billion CD4+ cells a day (Abdool Karim, 2005). The virus continues to replicate, causing a gradual decline in CD4+ count, which in turn makes the individual susceptible to various opportunistic diseases such as TB and pneumonia. This signals the onset of AIDS and in the absence of treatment the immune system fails to protect the person from invading opportunistic infections and the viral load increases. The natural history of HIV is depicted in Figure 1.1 from DeFranco, Locksley and Robertson (2007). In the absence of treatment, the time from HIV infection to AIDS diagnosis is between 5 and 10 years. Death will occur between eighteen months and two years after the onset of AIDS (Abdool Karim, 2005). These figures came from studies done in Europe and North America, and the natural history of HIV/AIDS in Africa is actually shorter by one or two years (Abdool Karim, 2005). A healthy person has between 500 and 1500 CD4+ cells/$\mu\ell$ blood (Abdool Karim, 2005) or even higher, but if the CD4+ count falls to low levels then the immune system can no longer fight disease properly. The CD4+ cells are counted from a blood sample with a method called laser flow cytometry. Flow cytometry is a technique used for counting, examining and sorting microscopic particles suspended in fluid. Through the binding of monoclonal antibodies, which recognise specific surface structures on these cells, the

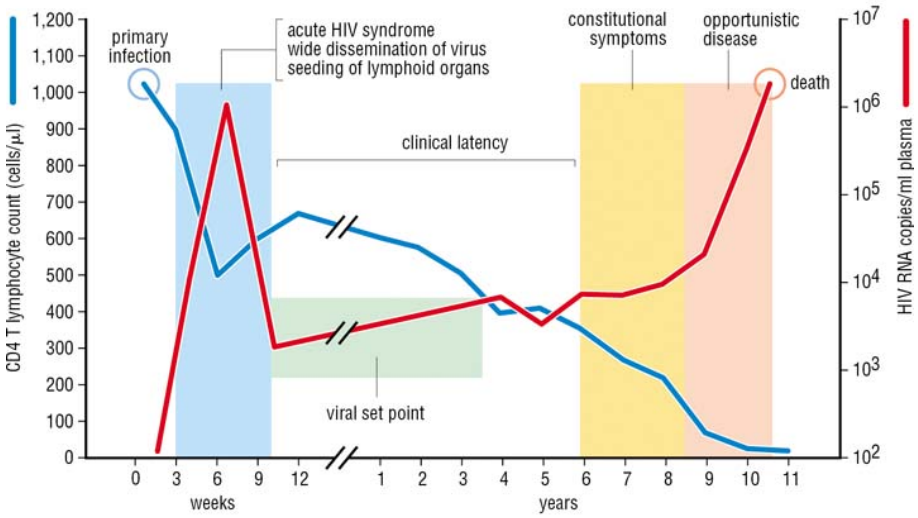

Figure 1.1: Natural History of HIV

CD4+ cells are labelled with special fluorescent markers. This allows one to distinguish the CD4+ cells in the blood sample and they are counted while passing a detector.

The HIV-RNA (HIV ribonucleic acid) viral load measurement determines the amount of HIV particles in the individual's blood, the number of copies of HIV-RNA per millilitre blood. One method that is used is a direct amplification of the viral load RNA, exactly defined multiplication of the RNA copies in the sample, and is called the Polymerase Chain Reaction (PCR) technique. Another method used is a high amplification of the measured signal (branched DNA assay, bDNA). With development of such technologies, scientists have been able to achieve a lower detection limit of 50 copies/mℓ. These are quite low numbers since, more than 30 000 viral copies/mℓ plasma is regarded as a high viral load.

1.1.3 HIV Diagnosis

There are two diagnostic approaches when it comes to testing for HIV, i.e. detecting the virus itself (rapid tests, ELISA) and detecting an immunological response to the virus, such as testing for HIV antibodies. The choice of the diagnostic test depends on the situation. A very early infection can only be diagnosed using tests that detect the virus, as the body would not have developed antibodies or cellular responses yet to the virus. Common tests for detecting the virus itself would include p24 antigen and PCR tests, and tests for detecting an immunological response would be rapid antibody tests or Enzyme-Linked ImmunoSorbent Assay (ELISA). If someone had to test PCR positive and antibody negative on a sample of blood taken on the same day then

they would be considered in the process of seroconverting (i.e. developing antibodies) and thus must have been infected recently. This stage of being antibody negative and PCR positive is called the window period. The distinction between these two methods has been exploited to estimate incidence rates for HIV by several researchers, among them Kaplan and Brookmeyer (1999).

The p24 antigen is a protein that is part of HIV and as the virus replicates, more p24 is produced and can be detected in the blood. Thus PCR is used to test for the genetic material of the virus itself. The process itself involves extracting and amplifying genetic material of an organism and testing for it using an Nucleic-acid Amplification Testing (NAT). These can test for either HIV-DNA or HIV-RNA. HIV-DNA is mainly used on newborn babies born to HIV positive mothers, as the HIV-RNA from the mother might still be in the baby's system and will give a false impression of the status of the infant. PCR testing for HIV usually involves testing for HIV-RNA. Tests for the virus itself would be able to detect an HIV infection much sooner than an antibody test, and a PCR-RNA test would produce a positive test result within 2 to 3 weeks after infection (Abdool Karim, 2005).

Antibody tests are the most common test used to determine whether someone has HIV, as it is also the least expensive in some cases. The test determines whether a person has developed specific antibodies against HIV. Rapid antibody tests are available that can deliver a result in under 30 minutes and are therefore used in the public clinic setting when testing a patient for HIV. The ELISA antibody test may take a few days to produce results, but it is more accurate as it is sensitive and reliable. An antibody test will produce a positive result after 6 weeks of infection.

1.1.4 Treatment of HIV

With the CD4+ count at dangerously low levels, an infected person's immunity is compromised and this person becomes prone to acquiring opportunistic infections, such as Tuberculosis, Cryptococcal Meningitis, Kaposi's Sarcoma, Peripheral Neuropathy, etc. The best way to prevent these opportunistic infections is to improve the level of immune function through highly active antiretroviral therapy (HAART), a combination of three or four different antiretroviral (ARV) drugs. When ARVs were first developed and used to treat HIV, only one drug was prescribed as treatment. Later, as different ARVs were developed and the medical community realised that patients were developing resistance to these ARVs, they started prescribing three or four concurrent ARVs as treatment and found this to be more effective in controlling HIV. HAART has now become standard treatment. Today the terms HAART and ARVs are used interchangeably to refer to HIV

treatment. The main aim of HAART is to delay or prevent the progression to AIDS and death of those infected with HIV by suppressing and slowing down the replication of the virus. HAART maintains the reproductive number (Anderson and May, 1991) of the viral population below a threshold that cannot allow the viral population to increase and dominate. The World Health Organization (WHO) have recommended guidelines as when to start antiretroviral therapy (Panel on Antiretroviral Guidelines for Adults and Adolescents, 2008). Unfortunately, if a person is at an advanced stage of HIV/AIDS, when the CD4+ count is less than 50 cells/$\mu\ell$, then starting therapy would not be always successful. Once an HIV infected person initiates HAART, he or she has to take it for the rest of his or her life in order to control the virus.

There are different opinions on when HAART therapy should be initiated. Since the therapy will have to be continued for the rest of the infected person's life and thus many years, it is not advisable to start HAART immediately after testing HIV positive. Another reason for postponing treatment until it is absolutely necessary, is that most of the ARV drugs have side-effects. There are some official guidelines regarding the initiation of HAART therapy. In particular, the US Department of Health and Human Services (Panel on Antiretroviral Guidelines for Adults and Adolescents, 2008) recommend that HAART should be started if someone has an AIDS-defining illness or if their CD4+ count falls below 350 cells/$\mu\ell$. They also state that certain groups of people should be initiated on HAART regardless of their CD4+ count, for example pregnant women, patients with HIV-related nephropathy (kidney disease) or patients co-infected with hepatitis B virus. Guidelines for initiating therapy can differ between countries. In South Africa, the criteria for initiating HAART, according to the National Department of Health, is that a person must either have a CD4+ count under 200 cells/$\mu\ell$ or a WHO stage IV, regardless of the CD4+ count (National Department of Health South Africa, 2004). In 1990 the World Health Organization (WHO) developed a staging system for people infected with HIV. This system uses conditions and infections to classify someone with HIV into a particular stage, ranging from stage I to IV. The staging increases as the severity of the diseases increases with stage IV corresponding to full-blown AIDS (WHO, 1990).

There are three broad types of antiretroviral drugs that have been developed:

- Nucleoside Reverse Transcriptase Inhibitors (NRTI), such as Zidovudine (AZT) or Lamivudine (3TC), which mimics the natural building blocks of DNA and act as chain terminators.
- Non-nucleoside Reverse Transcriptase Inhibitors (NNRTI), like Nevirapine (NVP) or Efavirenz (EFV), which bind directly to the enzyme reverse transcriptase and inhibits its activity; and

- Protease Inhibitors which prevents cleavage of viral proteins resulting in the production of immature viral particles.

Since the introduction of HAART, there has been dramatic decrease in rates of mortality due to HIV/AIDS. It has changed the perceptions of the HIV/AIDS epidemic from it being viewed as a death sentence to be seen as just a manageable chronic illness.

A main concern in the treatment of individuals with HAART is the fact that the HIV has the characteristic ability to mutate and develop resistance to the drugs. Another reason why initiation of ARV therapy is delayed, until deemed necessary, is to decrease the chances of someone developing resistance to treatment. The mutation of HIV and resistance to antiretroviral therapy threatens the usefulness of the treatments available on the market. Although HAART is able to control viral replication, it cannot completely eradicate HIV which persists in the host cells. This storage of infected cells allows the virus to replicate when HAART is discontinued or when the therapy can no longer suppress the virus.

Currently new treatments are being studied, the most advanced of those are entry inhibitors which prevent the interaction between the virus and the host cell. These class of drugs will be particularly useful in preventative agents such as microbicides and pre-exposure prophylaxis (PrEP). In HIV research, microbicides is a substance used vaginally or rectally to reduce the infectivity and replication of HIV and other sexually transmitted diseases. It can come in various forms, including gels, creams, sponges, rings or suppositories. PrEP is the long term use of an oral antiretroviral treatment for HIV prior to exposure, which is already being implemented in mother-to-child transmission of HIV. Currently studies are being done with PrEP in healthy adults to determine whether it is protects against HIV acquisition.

1.1.5 Types of HIV

There are two main types of HIV, namely HIV-1 (including three lineages called the M, N, and O groups) and HIV-2 (with three lineages A, B and G). The viruses in the HIV-1 M group are mainly responsible for the AIDS epidemic. The HIV-1 M group is divided further into 11 subtypes (A1, A2, B, C, D, F1, F2, G, H, J and K). Figure 1.2 (Kuiken et al., 1999) offers a graphical representation of some of the different HIV subtypes that have been classified. HIV-1 subtype C is the virus that dominates the epidemic in South Africa, as well as in Nepal and India, and in 2000 accounted for 47% of all HIV infections in the world (Osmanov et al., 2002). Studies have shown that subtype C dominates the HIV infections in South Africa with one study showing a

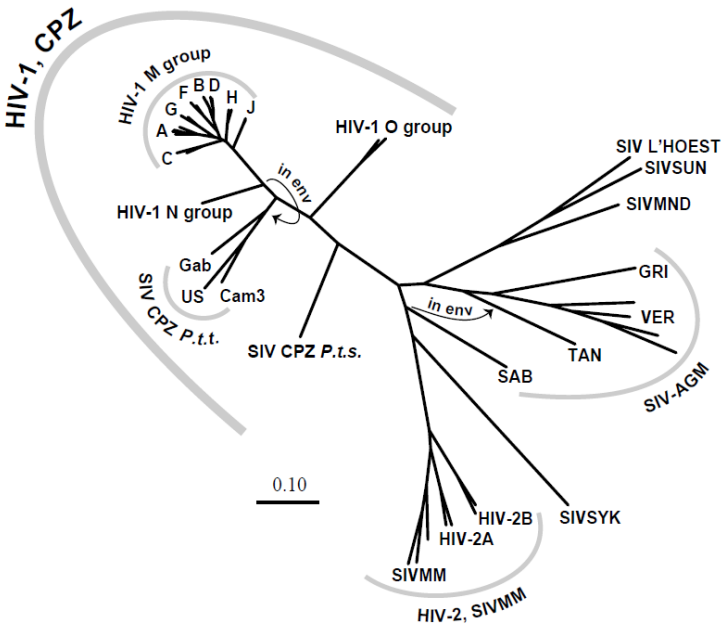

Figure 1.2: HIV Subtypes

92% prevalence of subtype C (Van Harmelen et al., 1999). HIV-2 is less harmful and is restricted to West Africa. HIV-1 subtype B is dominant in the Americas, Europe, Japan, Thailand and Australia, while subtype D is limited to East and Central Africa. Figure 1.3 is from Osmanov et al. (2002) and shows the prevalence of HIV-1 subtypes in 2000. A study presented in 2007 found that women infected with HIV-1 subtype D in Kenya had more than double the risk of death compared to women infected with HIV-1 subtype A over six years (Baeten et al., 2007). Another study of sex workers in Senegal found that women infected with HIV-1 subtype C, D or G were more likely to progress to AIDS within five years of infection compared to those infected with subtype A (Kanki et al., 1999). Thus there are different subtypes of HIV which have been shown to be more virulent then other subtypes.

1.1.6 Markers for HIV

CD4+ cell count provided the first reliable marker for disease progression since it gives an indication of how the immune system is doing. The viral load is also a reliable marker as it is the most important indicator for the effectiveness of HAART. Both these markers are the best available predictors of progression to disease and death (Lyles et al., 2000). The aim of HAART is to lower the viral load in the infected person's blood to a point where the virus is undetectable and in turn, allow the CD4+ cells to increase and the immune system to recover. Thus reduction of viral load

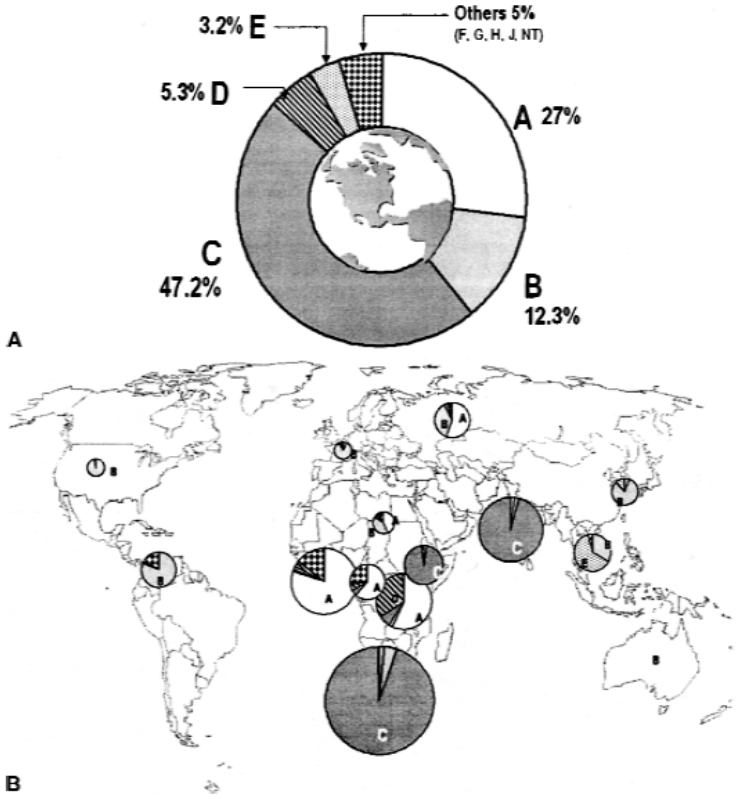

FIG. 1. **(A)** Estimated incidence of HIV-1 *env* subtypes in 2000. **(B)** Estimated distribution of new HIV-1 infections by *env* subtypes and regions in 2000.

Figure 1.3: HIV-1 subtype prevalence in 2000

is associated with a delay of disease progression and lower rates of transmission. However, if a person's viral load is undetectable, it does not mean that he or she no longer has the virus. Rather, it means that the virus is under control and the individual would have to be on HAART for the rest of his or her life. Newer assays lead to much lower detection limits of the viral load. An increase in CD4+ count and a decrease in viral load, a suppression of viral load below 50 copies/mℓ are standard endpoints of HAART trials. However, it should be noted that these surrogate markers do not truly represent clinical outcomes.

The current thesis is aimed at applying statistical methods useful in understanding the evolution of CD4+ count and viral load for newly infected individuals, before they need to be initiated on HAART. Such work is important because it provides investigators with a tool to understand the dynamics of HIV/AIDS before and after HAART initiation.

A review of publications on acute HIV infection and joint modelling will be follow in Section 1.2 and an exploratory analysis on the data used for this thesis is described in Chapter 2. The

theory on the statistical methods used will be covered in Chapters 3 and 4, while the application will be in presented in Chapter 5. Finally the conclusion and future work will be discussed in Chapter 6.

1.2 Literature Review

1.2.1 Acute HIV Infection

Currently there are not many studies aimed at acute HIV infection. Many HIV studies look at treatment, disease progression, opportunistic infections and genetic factors influencing HIV acquisition and progression. It is difficult to study acute HIV infection for logistical reasons. Individuals need to be recruited early enough after acquiring HIV to make sure they are still in the acute infection stage of the disease. Thus their date of infection needs to be known and the only effective way of doing this is to run large cohort studies following HIV negative subjects at risk of getting HIV, testing them regularly and study the subjects who acquire HIV.

Goujard et al. (2006) followed 552 patients who were enrolled into the French PRIMO cohort during primary HIV infection and majority of the patients were male (80.8%). Patients were enrolled if they had become HIV infected less than 6 months prior to enrolment. Primary HIV infection was diagnosed with one of the following criteria: (i) having an incomplete Western Blot result (with the absence of anti-p68 and anti-p34), (ii) a positive HIV-RNA test result and a negative ELISA (antibody test) result or (iii) an interval of less than 6 months between last negative and first positive ELISA.

Goujard et al. (2006) estimated date of infection as either the date of the incomplete Western Blot result minus one month, or the midpoint between last negative and first positive ELISA. If the patient had experienced symptoms of HIV seroconversion (and subsequent tests later proved a positive HIV diagnosis), then the date of infection was estimated to be 15 days prior to onset of the symptom(s). Patients were ARV treatment naïve and were followed for a median of 30 months (interquartile range 16 to 54 months). Clinical and laboratory investigations were performed at months 1, 3, 6 and six-monthly thereafter. Goujard et al. (2006) analysed the risk factors associated with disease progression, which they defined by the occurrence of death or an AIDS defining illness or event. Disease progression was also defined as a patient having a CD4+ count measurement of less than 350 cells/$\mu\ell$ after 3 months of follow-up. This value was chosen because it is the threshold which is recommended at which ARV treatment should be initiated (Panel on

Antiretroviral Guidelines for Adults and Adolescents, 2008). Goujard et al. (2006) found that low initial CD4+ count and high viral load are predictive of rapid disease progression and their conclusion was that HIV infected patients will benefit from regular clinical and immunological monitoring.

Lyles et al. (2000) looked at the Multicenter AIDS Cohort Study (MACS), which is an ongoing study and at the time of their analysis had a total of 5622 homosexual or bisexual men. Of these, 3427 were HIV antibody negative at baseline and 511 of the 3427 men seroconverted during study follow-up. The MACS cohort enrolled men from centres in four US cities and they were followed semi-annually. Physical examinations, laboratory testing and questionnaires were administered at each visit. There were 269 of the 511 men who fulfilled the criteria which Lyles et al. (2000) set, which was to have at least 2 plasma specimens after seroconversion available. All available samples at the last negative and the first positive visits and for 2 years thereafter, were assayed for viral load. Beyond 2 years after the first positive visit, one sample per year from each participant was assayed. The total number of samples in which viral load was measured for these 269 men was 2527 and this stretched over a period of 10 years after seroconversion. For 90% of the men, the dates of their last negative and first positive HIV test result was less than 7 months apart. According to Lyles et al. (2000) the last negative and the first positive visits took place on average at -3 and +3 months from seroconversion. Lyles et al. (2000) described the association between CD4+ count and viral load at the first positive visit using a multiple linear regression model. They also described trends in viral load and CD4+ count over time by fitting random-effects linear models, providing estimates of the average linear marker trajectories over time, while taking into account repeated measurements. Since the data was restricted to samples taken prior to 1990, none of the men were on ARV treatment at the time the data was analysed, thus they could model the natural history of HIV over this long period of time.

Lyles et al. (2000) found in this specific cohort that CD4+ count decreased by 249 cells/$\mu\ell$ per year within the first 3 years of HIV infection. Within the same time period, viral load increased by 0.18 log copies/mℓ per year. Between 3 and 7 years after seroconversion, CD4+ count decreased at a slower rate of 89 cells/$\mu\ell$ per year while viral load continued to increase at a rate of 0.09 log copies/mℓ per year. After 7 years of HIV infection, CD4+ count continued to decrease at a similar rate of 87 cells/$\mu\ell$ per year. Interestingly, their model showed that viral load was decreasing in the period after 7 years after seroconversion with a rate of 0.01 log copies/mℓ per year. However, they found that the slope during the 3 years prior to progression to AIDS was 0.14 to 0.20 log copies/mℓ per year and the estimated level at the time of progression to AIDS was 5.2 to 5.3 log copies/mℓ. Lyles et al. (2000) and the MACS cohort provided information on the CD4+ count and viral load

levels and slopes in the first few years after HIV infection, specifically on men who were not on any ARV treatment. Lyles et al. (2000) depicted the average viral load and CD4+ count trajectories in the first 3 years after the first HIV-seropositive visit for 218 participants a plot as shown in Figure 1.4. The first HIV-seropositive visit was approximately 3 months after seroconversion and they only included measurements before 1 January 1990 to avoid potential confounding effects of antiretroviral therapy.

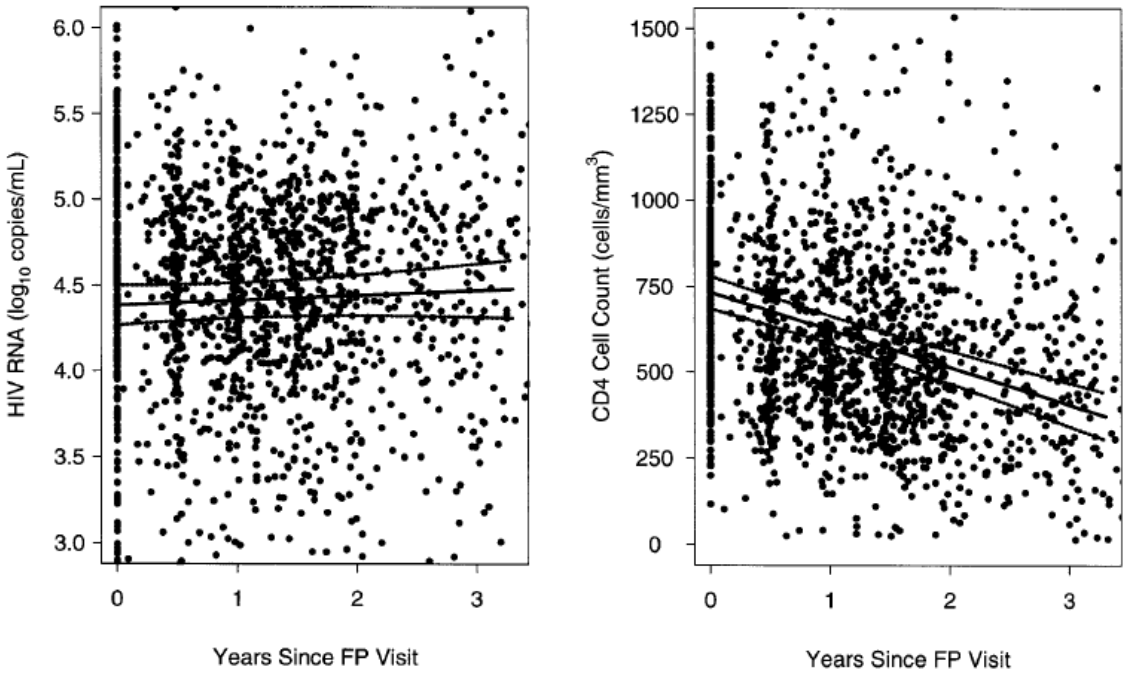

Figure 1.4: Average trajectories of viral load and CD4+ count during the first 3 years after the first HIV-seropositive visit (MACS Cohort)

1.2.2 Joint Modelling

This thesis also investigates the joint modelling approach of Henderson, Diggle and Dobson (2000) which enables one to combine longitudinal and survival data. Henderson, Diggle and Dobson (2000) applied the method to data from a clinical trial on the treatment of schizophrenia. The interest was in the effect of the three treatments and this effect was measured with a particular scoring system called the Positive And Negative Symptom Score (PANSS), a measure of psychiatric disorders. The PANSS score was measured at entry into the study (-1), baseline (0), 1, 2, 4, 6 and 8 weeks. Of the 523 patients on the trial, over a third dropped out because of `inadequate response ´and these were treated as informative drop-outs. Their joint modelling strategy was based on the specification of two linked latent Gaussian processes. They used a linear random effects model

for the longitudinal PANSS data and the EM estimation algorithm was their preferred method of estimation. They used the semi-parametric Cox regression model to fit the informative drop-out process. They also used various models for the latent Gaussian processes, looking at different types of associations between the longitudinal and time-to-event models.

Henderson, Diggle and Dobson (2000) found that, depending on the specification of the latent Guassian random processes, there can be difficulties with identifiability and sensitivity. They found several models to be a good fit to the data, if the patients do not drop out, but the models give very different results for the predicted drop-out-free population profiles.

Guo and Carlin (2004) revisited the method proposed by Henderson, Diggle and Dobson (2000). They looked at jointly modelling longitudinal and survival data from an AIDS clinical trial which compared two ARV treatments, Didanosine (ddI) and Zalcitabine (ddC). This trial specifically looked at patients who had failed or were intolerant to another antiretroviral treatment, Zidovudine (AZT). Their longitudinal outcome was CD4+ count, to which they applied a square root transformation in order to ensure normality. These CD4+ counts were measured at irregular intervals, i.e. at study entry and at 2-, 6-, 12- and 18-month visits. Time to death was also recorded and deaths were deemed as informative drop-outs. They had various explanatory variables in their model, namely current ARV treatment (ddI or ddC), gender, previous opportunistic infection or AIDS diagnosis at study entry and AZT failure or intolerance. They used a linear random effects model for square root CD4+ count, and both the exponential survival model (parametric) and the Cox proportional hazards model (semi-parametric) to analyse the time-to-event data. They fitted the models separately and then jointly and then proceeded to fit the joint model using Bayesian methods. Guo and Carlin (2004) found that the joint analysis increased the estimated median survival times by approximately 50% due to the model's correct accounting for the correlation between the longitudinal and survival data.

Pantazis and Touloumi (2005) also modelled AIDS data, and specifically fitted a bivariate repeated measures model to CD4+ count and viral load using a random effects model. They also added piecewise linear effects for the time component as they found that there was a sharp decline in viral load with the first year of seroconversion and slow increase thereafter. They adjusted for informative drop-out, which was due to disease progression or death, using a parametric log-normal survival model. The results of their analysis on simulated data confirmed that in the presence of informative drop-out, conventional analyses such as random-effects models lead to biased estimates. The method they applied performed well, giving model estimates with negligible bias.

Their two independent models (adjusting for informative drop-out but not for correlated markers) gave slightly larger biases for the fixed effects than the bivariate model, but a much smaller bias compared with the two independent or the bivariate random effects models.

Thiébaut et al. (2005) looked at the data from the CASCADE project, which followed a group of 994 patients on HAART. They restricted their analysis to 494 patients who had a CD4+ count and viral load measurement done at treatment initiation and at least one measurement thereafter. They analysed the CD4+ count and viral load in a bivariate mixed model, with two linear piecewise effects which allowed them to model the rate of change of the HIV markers in the first 1.5 months and after 1.5 months. They accounted for left-censoring of viral load measurements, using a parametric approach, as 57% of the viral load measurements were below the lower limit of detection. Thiébaut et al. (2005) used joint modelling to combine the longitudinal bivariate mixed model to a log-normal survival model in order to take into account the informative drop-out. They found that modelling the longitudinal model as bivariate, i.e. having both CD4+ count and viral load in the outcome variable, had changed the fixed effects significantly compared to just modelling each of these HIV markers in a univariate model. It was also useful that the bivariate model provided estimates of the correlation between CD4+ count and viral load throughout follow-up. Thiébaut et al. (2005) also found that the piecewise modelling was useful in the interpretation of the HIV markers, but they found its major drawback was the inability to detect a viral rebound. This would certainly be the case as simple piecewise modelling requires pre-specified intervals. However they admit that in their case it would probably not have been possible to witness this viral rebound because of the restrictions they had placed on the data they analysed. With regards to their joint modelling, they found a significant correlation between the random slope in CD4+ count and time to treatment modification. In their models accounting for informative drop-out, the latter slope of CD4+ count was lower compared to the models where informative drop-out was not modelled. Similarly in the viral load model which accounted for both drop-out and left-censoring, the slope for viral load was actually predicted to be lower compared to the crude model not accounting for left-censoring or drop-out.

Chapter 2

Exploratory Data Analysis

2.1 CAPRISA 002: Acute Infection Study

In August 2004, the Centre for the AIDS Programme of Research in South Africa (CAPRISA) initiated a cohort study, enrolling high risk HIV negative women (Van Loggerenberg et al., 2008). The objective of the CAPRISA 002: Acute Infection (AI) Study was to describe demographic and clinical characteristics, and measure HIV incidence of a cohort at high risk for HIV infection in South Africa. A total of 245 women were enrolled, the mean age of the cohort being 34.2 years (range 18-58), and majority of the women (78.8%) were self-reported sex workers (Van Loggerenberg et al., 2008). Various clinical assessments and behavioural questions were asked during enrolment into the study and CD4+ count measurements of these HIV negative women were also measured. These women were followed monthly for up to two years and received regular clinical examinations, including monthly HIV tests. During this follow-up period, 245 women completed 4784 visits, and from these, 28 women acquired HIV infection. A second phase of the study followed these newly infected individuals closely to determine clinical characteristics of HIV during early infection. Thirty four acutely infected women from other CAPRISA research cohorts in urban and rural were enrolled into this phase of the study, bringing the total of acutely infected individuals up to 62. The mean age of the 62 acutely infected women was 28.1 (SD=9.08). The acutely HIV infected women were accrued over approximately four years with the first HIV infected woman enrolled into this phase of the study in October 2004 and the last woman enrolled in September 2008.

One of the criteria for enrolling these HIV infected women into the second phase of the AI study was that the time between last negative and first positive HIV test had to be no more than 5 months apart. This was to ensure that participants are enrolled as early as possible into their infection to

be classified as acutely infected. Since each woman had a last negative and first positive test date their possible date of infection could only be approximated. For the purposes of this thesis the midpoint method was used to determine the possible date of infection assuming event times are uniformly distributed between the two dates. Participants who tested HIV negative on an antibody test and positive on a PCR on the same day, were those who had recently been infected and they have not developed antibodies yet and were still in their window period. Their date of infection was estimated to be 14 days prior to when they tested antibody negative and PCR positive. There were 17 women who tested PCR positive and antibody negative. Estimating the date of infection allows for analysis of the natural history of HIV with regards to the surrogate markers (CD4+ count and viral load), since time post infection can be calculated and the evolution of the disease can be described. Of course a more precise way of analysing such data would be to use methods which take into account the uncertainty in the date of infection data, such as interval censoring or the measurement error approach (Carroll et al., 2006). However, this is not the aim of the current study.

After identifying the women as HIV positive they were enrolled into the second phase of the study and were followed up thereafter weekly for three weeks, fortnightly for two months, monthly for nine months and from there onwards quarterly until the end of the study. At each visit various clinical examinations were performed, including measuring CD4+ count and viral load. It is important to note that these women were not treated for HIV and thus were ARV naïve. The reason for this is that these women, being acutely infected with HIV, are not yet eligible for HIV treatment.

Since the participants had their CD4+ count regularly measured, their disease progression could be assessed. When their CD4+ count fell below 350 cells/$\mu\ell$ for more than two consecutive visits, they would be referred to a public government clinic for ARV treatment. However, these participants would only start HAART once their CD4+ count falls below 200, according to South African government policy (National Department of Health South Africa, 2004). These participants would still be followed up in the AI study and only terminated once they initiate ARV therapy. These terminations would be classified as informative drop-outs since their exit from the study is related to the outcome that will be analysed. Out of the 62 women, 11 (17.7%) had to be initiated on treatment and the median month post infection these women were at initiated on ARV therapy was 27.5 months (interquartile range 15 - 40 months). Figure 2.1 shows the Kaplan-Meier graph of time to ARV initiation from date of infection. Participants who were not initiated on treatment were censored at their last HIV positive follow-up visit.

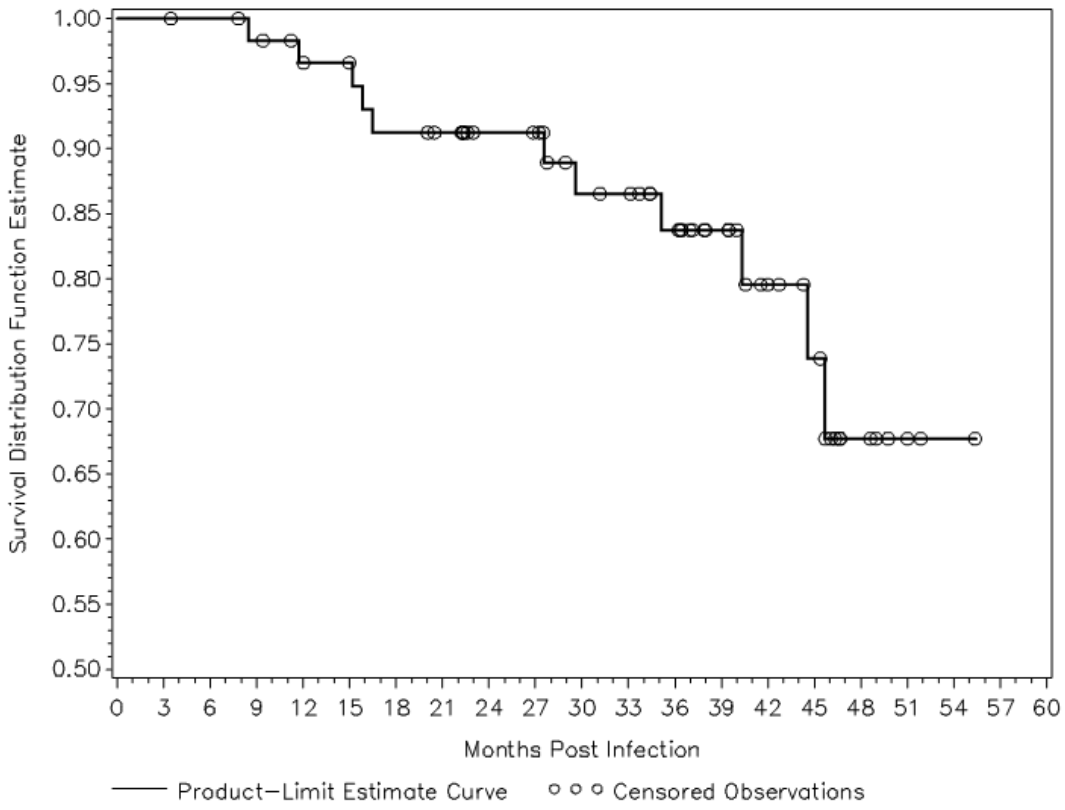

Figure 2.1: Kaplan-Meier graph of time to ARV initiation

2.2 Viral Load and CD4+ Count Data

Out of the 245 HIV negative women who were enrolled initially, 28 seroconverted and three of these 28 women were actually HIV positive at entry into the study, as PCR tests later showed. Hence, only 25 women were truly negative at study entry and had a healthy (HIV negative) CD4+ count. The HIV infected women in the second phase of the study were followed for a median of 33.2 months (range 2.2 to 53.6 months). At entry into the second phase of the study, the women had been infected for a median of 4.5 weeks (range 1 to 15). For the 28 women who were enrolled into the original HIV negative cohort, this time post infection at the second phase was a mere 2 weeks (range 1 to 4), illustrating how soon after being infected with HIV that these women were studied. There was a total of 1384 viral load and 1378 CD4+ count measurements taken at various time points during follow-up. A further 62 viral load measurements were assumed to be 0 at week 0 when participants were still HIV negative (giving a total of 1446 viral load measurements). By plotting the CD4+ count and viral load observations over time post infection a pattern in the HIV markers can be established and the effect of HIV among those newly infected can be studied.

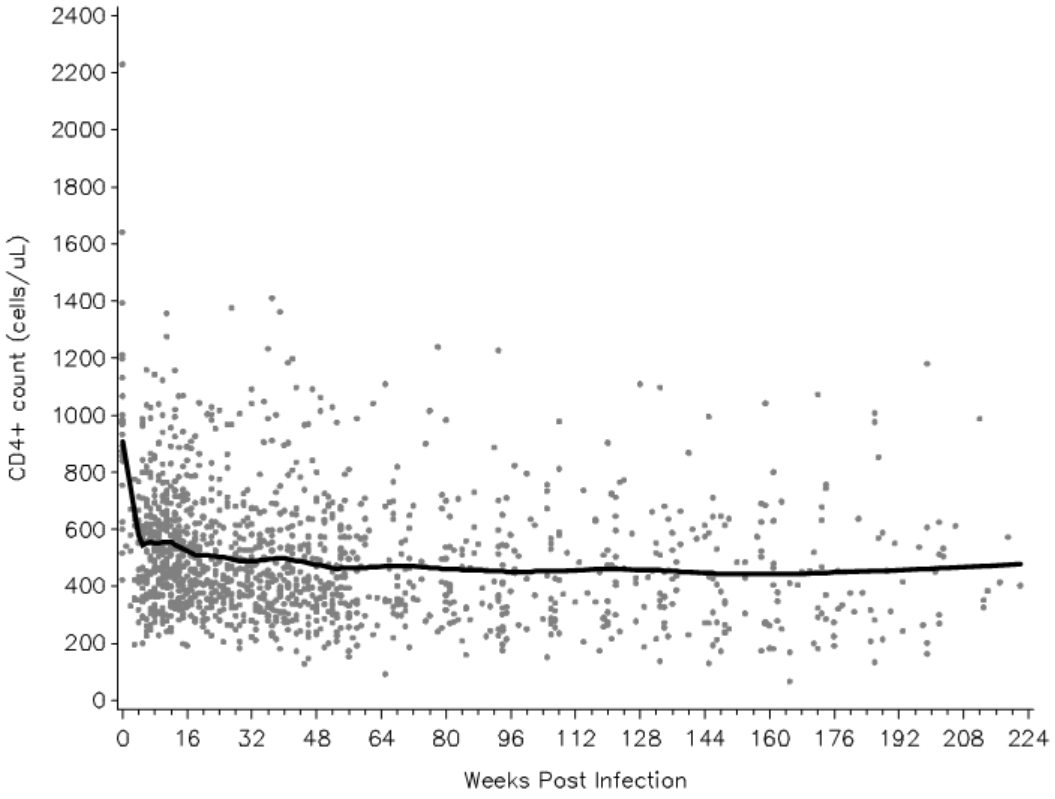

Figure 2.2: Scatter plot of all CD4+ count (cells/$\mu\ell$) measurements with a Loess smoothing regression line

Figure 2.2 and 2.3 illustrates the actual CD4+ count and viral load measurements plotted over weeks post infection with a Loess smoothing line overlaid in each. Note that week 0 represents an HIV negative state. The CD4+ count plot (Figure 2.2) shows an initial drop followed by a return to near stable level. However, the CD4+ count never reaches its initial HIV-uninfected state. The viral load plot (Figure 2.3) indicates a sharp rise followed by a drop to a near stable state as expected for HIV infected individuals who are still in the acute phase of the disease.

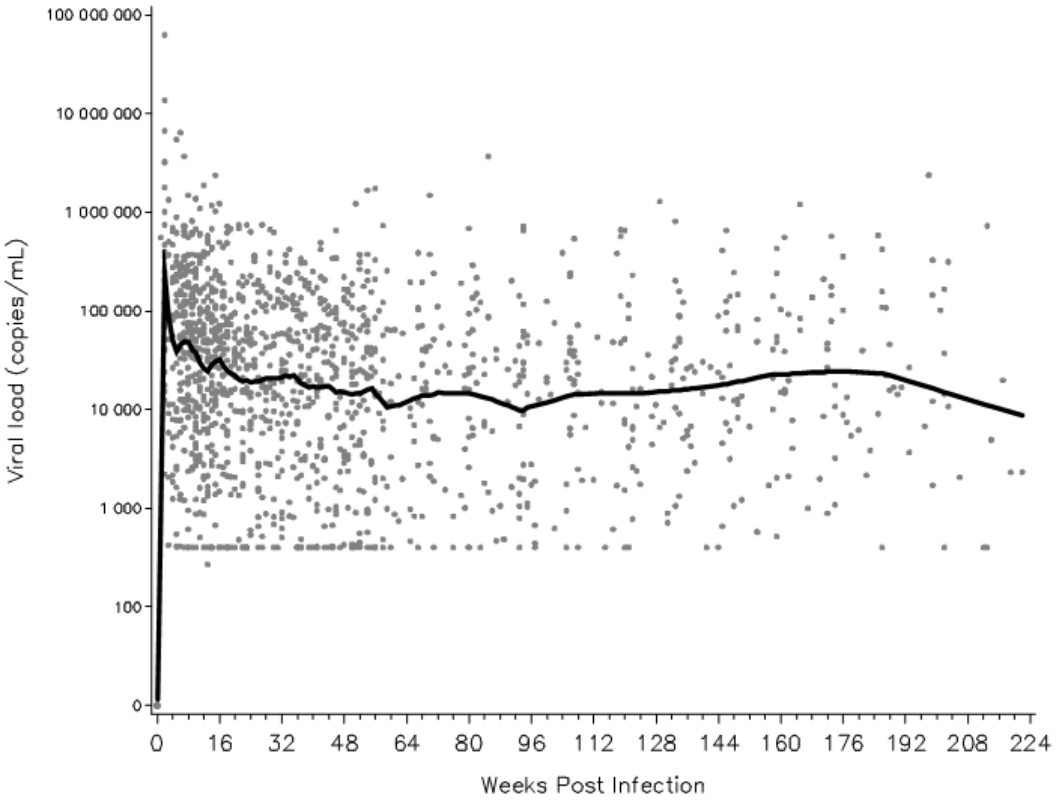

Figure 2.3: Scatter plot of all viral load (log copies/mℓ) measurements with a Loess smoothing regression line

Table 2.1 and 2.2 show the distribution of the number of repeated measures for CD4+ count and viral load respectively. There were a total of 1378 CD4+ count and 1446 viral load measurements for 62 participants. From these tables it can be seen that 98.4% (61/62) of individuals had 11 or more CD4+ count measurements, with 79.0% (49/62) of individuals having at least 20 CD4+ count measurements. Table 2.2 shows that again, 79.0% (49/62) of individuals had at least 20 viral load measurements.

No. of observations per subject	Number of subjects	% of subjects (cumulative %)	Total number of observations (%)
4	1	1.6 (1.6)	4 (0.3)
11	1	1.6 (3.2)	11 (0.8)
14	1	1.6 (4.8)	14 (1.0)
15	2	3.2 (8.1)	30 (2.2)
17	2	3.2 (11.3)	34 (2.5)
18	4	6.5 (17.7)	72 (5.2)
19	2	3.2 (21.0)	38 (2.8)
20	4	6.5 (27.4)	80 (5.8)
21	5	8.1 (35.5)	105 (7.6)
22	5	8.1 (43.5)	110 (8.0)
23	4	6.5 (50.0)	92 (6.7)
24	2	3.2 (53.2)	48 (3.5)
25	3	4.8 (58.1)	75 (5.4)
26	6	9.7 (67.7)	156 (11.3)
27	4	6.5 (74.2)	108 (7.8)
28	5	8.1 (82.3)	140 (10.2)
29	5	8.1 (90.3)	145 (10.5)
30	3	4.8 (95.2)	90 (6.5)
31	3	4.8 (100.0)	93 (6.7)

Table 2.1: Repeated Measurements of CD4+ count

2.2.1 Individual Profiles

Figures 2.4 and 2.5 illustrate the actual CD4+ count and viral load measurements for each participant over time. There is clearly a lot of variability between individuals throughout follow-up in both CD4+ count and viral load. These graphs also show that there is much variability within individuals. In Figure 2.4 it can be seen that a decrease in CD4+ count is experienced after HIV infection, but again with evident variability in this drop in CD4+ count between individuals. It is important to note that CD4+ count was only measured within the HIV negative women at enrolment into the study and it is assumed that this was their healthy CD4+ count the day before infection. Over the same time interval in Figure 2.5, these women experienced a sharp rise in viral load, assuming that viral load is zero at week 0.

No. of observations per subject	Number of subjects	% of subjects (cumulative %)	Total number of observations (%)
4	1	1.6 (1.6)	4 (0.3)
11	1	1.6 (3.2)	11 (0.8)
14	1	1.6 (4.8)	14 (1.0)
15	2	3.2 (8.1)	30 (2.1)
17	2	3.2 (11.3)	34 (2.4)
18	4	6.5 (17.7)	72 (5.0)
19	2	3.2 (21.0)	38 (2.6)
20	4	6.5 (27.4)	80 (5.5)
21	5	8.1 (35.5)	105 (7.3)
22	5	8.1 (43.5)	110 (7.6)
23	4	6.5 (50.0)	92 (6.4)
24	2	3.2 (53.2)	48 (3.3)
25	3	4.8 (58.1)	75 (5.2)
26	6	9.7 (67.7)	156 (10.8)
27	4	6.5 (74.2)	108 (7.5)
28	5	8.1 (82.3)	140 (9.7)
29	5	8.1 (90.3)	145 (10.0)
30	3	4.8 (95.2)	90 (6.2)
31	3	4.8 (100.0)	93 (6.4)

Table 2.2: Repeated Measurements of viral load

Four participants' CD4+ count and viral load data are depicted in Figure 2.6. The solid line represents viral load and the dotted line, CD4+ count. Participant 1 in Figure 2.6 (a) is a rapid progressor as her CD4+ count is consistently below 350 cells/$\mu\ell$. She has been initiated on ARVs. In contrast, Participant 5 in Figure 2.6 (b), who maintains a high CD4+ count and a low viral load is controlling the retro virus very well, is considered to be a slow progressor. Participant 9 in Figure 2.6 (c) is one of the infected who was initiated on HAART as her CD4+ count fell below 200 cells/$\mu\ell$ and this is why there is not as much CD4+ count and viral load data as she had left the study for treatment. Participant 11 in Figure 2.6 (d) represents a typical case, and her data clearly shows the interactive relationship between viral load and CD4+ count, however her latest measurements show she is progressing towards AIDS and she has also been initiated on HAART.

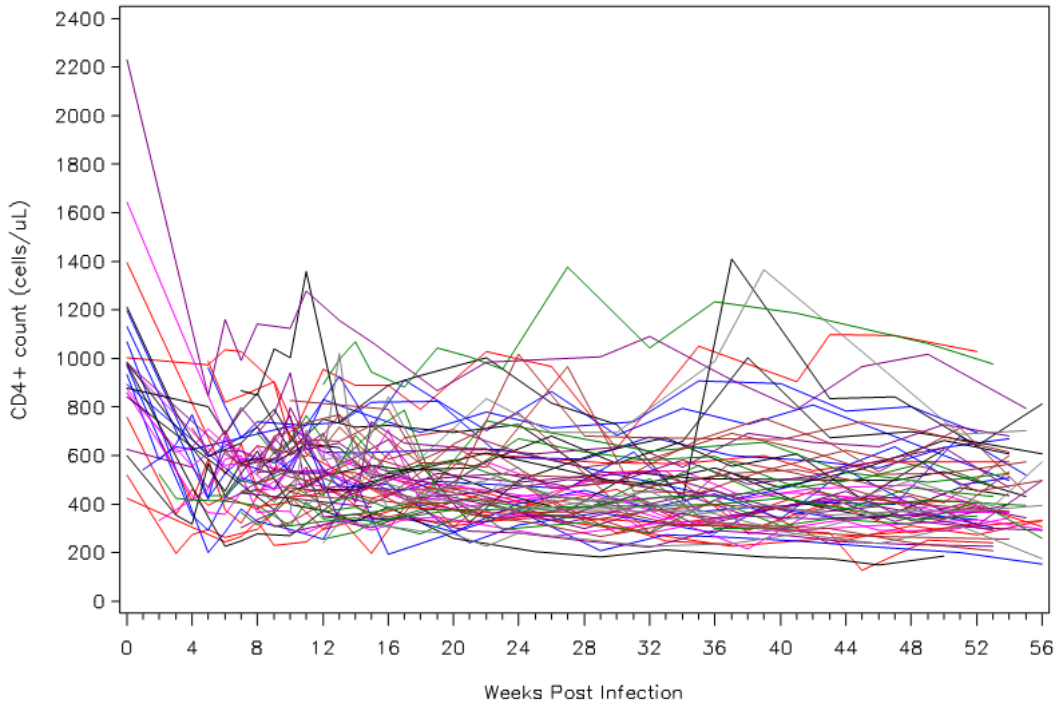

Figure 2.4: CD4+ count (cells/$\mu\ell$) for the first year of infection, each line representing an individual

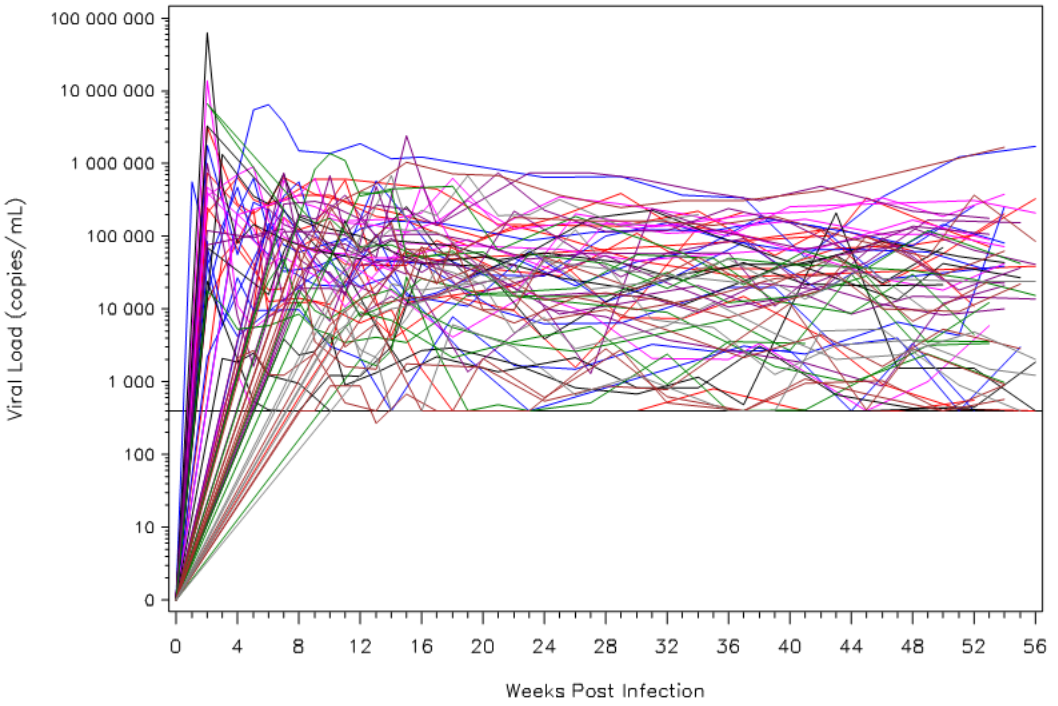

Figure 2.5: Viral load (log copies/mℓ) for the first year of infection, each line representing an individual

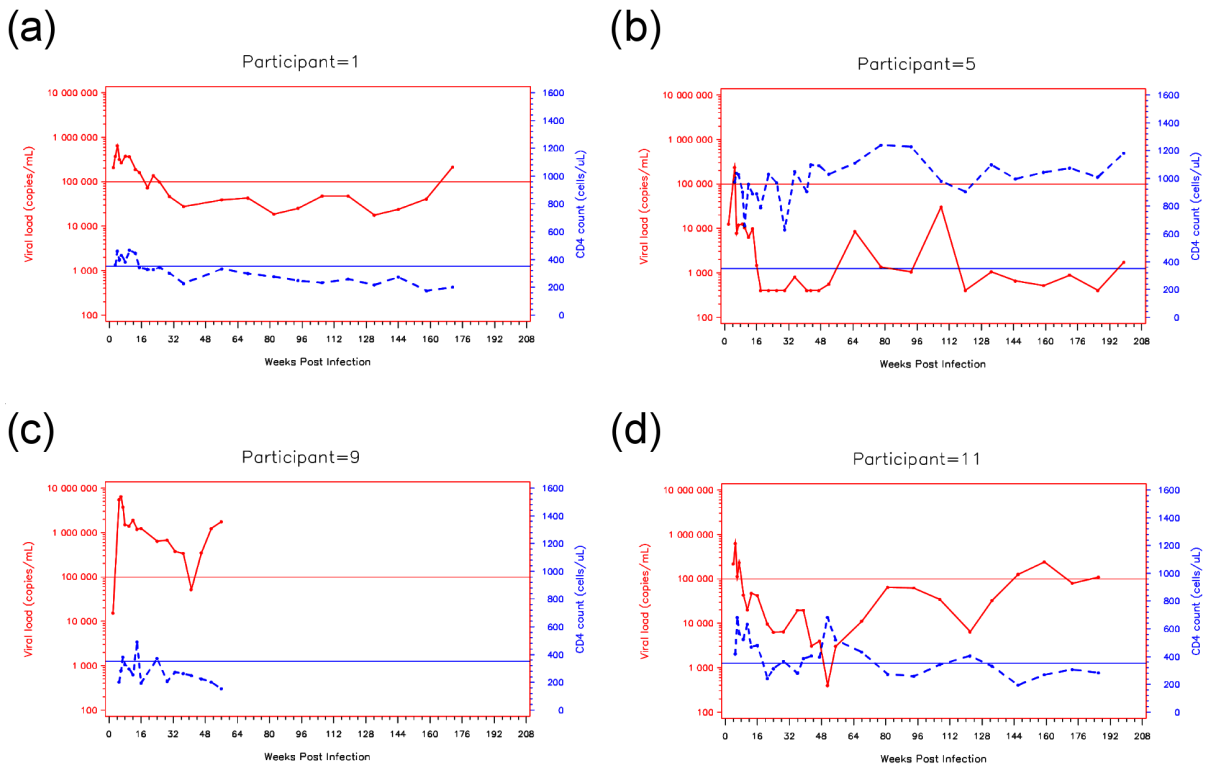

Figure 2.6: Examples of four individual profiles for viral load and CD4+ count

2.2.2 Basic Descriptive Statistics

Using time post infection, the closest measurement to months 1, 3, 6, 9, 12, 15, 18 and 24 were determined and basic summary measures for CD4+ count and viral load were calculated at these months post infection. Table 2.3 shows a summary of these basic statistics at the time points indicated.

From the box plots in Figure 2.7 and 2.8, it is seen how CD4+ count and viral load change at the specified time points after infection. Note that month 0 represents pre-infection, i.e. HIV negative. These box plots also show that the data, especially viral load, is not normally distributed as the mean and median measures at each time point tend to be quite different. The two sets of box plots both suggest variables which are clearly skewed to the right. The normality assumption can also be checked by plotting a histogram of CD4+ count and viral load separately. These are presented in Figures 2.9 and 2.10 for CD4+ count and Figures 2.11 and 2.12 for viral load.

Time	CD4+ count			Viral Load		
	n	Mean (SD)	Median (Range)	n	Mean (SD)	Median (Range)
Pre-Infection	25	993.1 (366.75)	969 (424 - 2231)	-	-	-
1 Month Post Inf	19	574.7 (210.32)	557 (275 - 989)	21	162 118 (217 418)	55 900 (547 - 698 000)
3 Months Post Inf	62	555.0 (223.6)	512 (242 - 1358)	62	123 954 (277 022)	39 100 (<400 - 1 890 000)
6 Months Post Inf	59	507.0 (218.18)	431 (206 - 1378)	59	85 201 (146 059)	21 100 (<400 - 750 000)
9 Months Post Inf	59	494.1 (266.48)	404 (182 - 1411)	59	57 092 (79 238.)	20 500 (<400 - 310 000)
12 Months Post Inf	58	461.6 (181.02)	404 (188 - 1030)	58	105 516 (275 576)	22 000 (<400 - 1 680 000)
15 Months Post Inf	50	476.8 (204.64)	435 (94 - 1110)	50	78 437 (222 601)	15 100 (<400 - 1 500 000)
18 Months Post Inf	43	483.4 (224.73)	416 (210 - 1240)	43	61 160 (117 917)	22 800 (<400 - 689 000)
24 Months Post Inf	35	456.5 (201.83)	423 (153 - 979)	35	62 764 (116 221)	21 100 (<400 - 541 000)

Table 2.3: Basic summary statistics of CD4+ count and viral load at pre-infection, and months 1, 3, 6, 9, 12, 15, 18 and 24 post infection

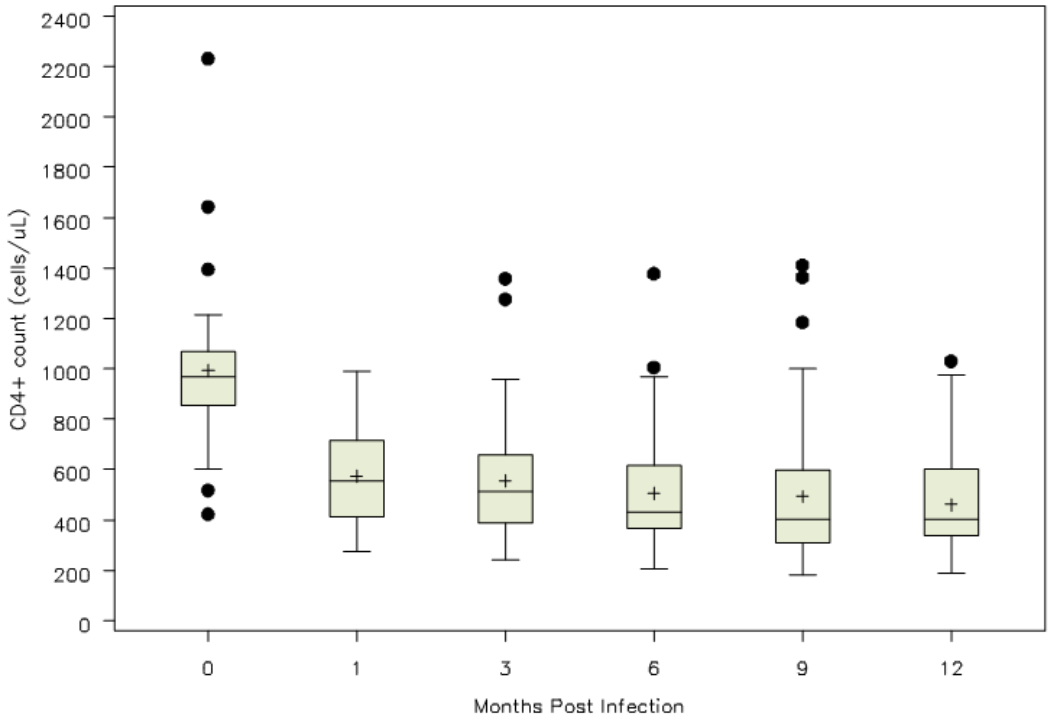

Figure 2.7: Boxplot of CD4+ count (cells/$\mu\ell$) by months post infection for the first year of infection

The histogram of actual CD4+ count shows non-normality with a right-skewed histogram. By applying a square root transformation to CD4+ count, the data is normalised, as can be seen in Figure 2.10. This justifies the use of square root CD4+ count rather than actual CD4+ count as the response variable in the modelling processes in Chapter 5.

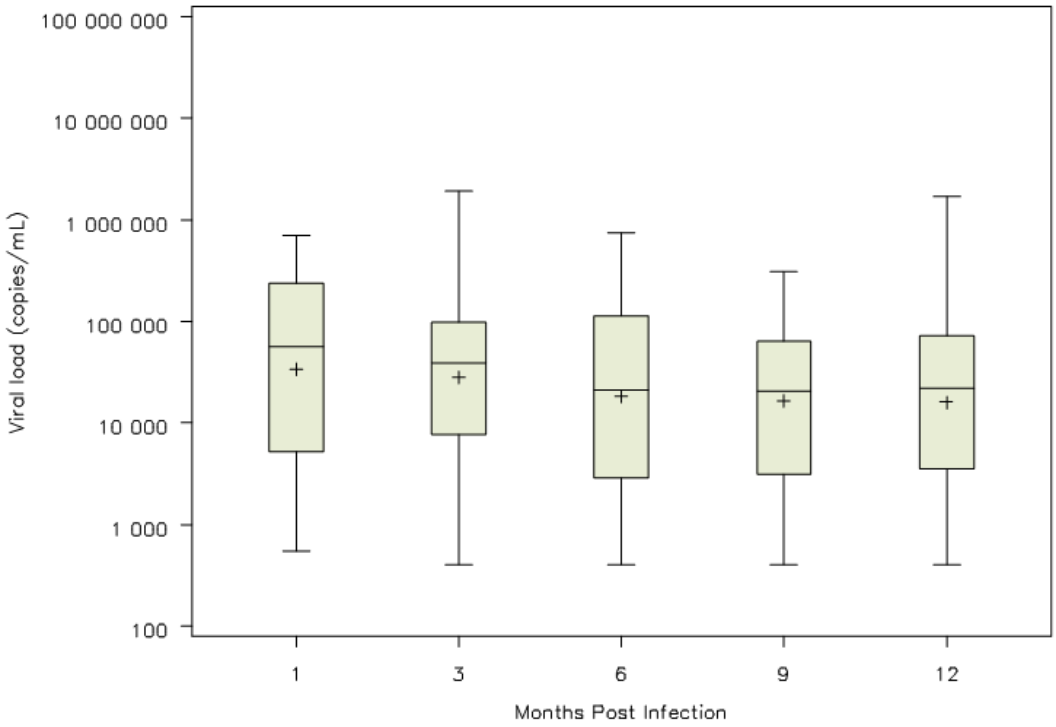

Figure 2.8: Boxplot of viral load (copies/mℓ) by months post infection for the first year of infection

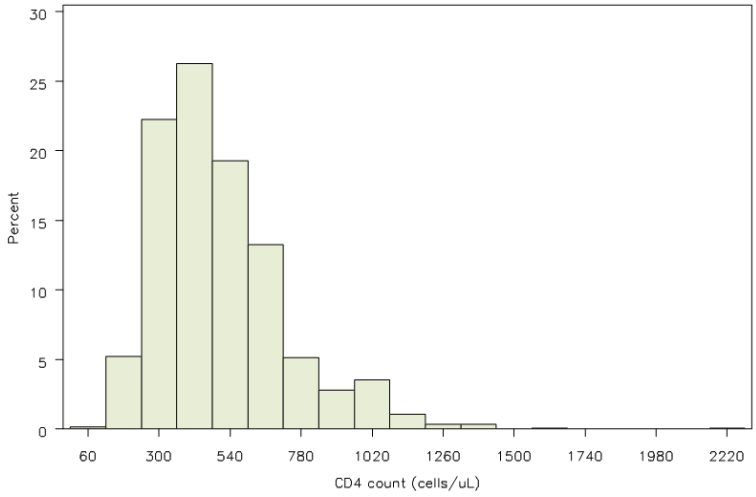

Figure 2.9: Histogram of CD4+ count (cells/$\mu\ell$)

Figure 2.11 is the histogram of all viral load measurements. The first histogram (a) was limited to viral load below 2 million copies/mℓ since the nine measurements that were above this threshold skewed the histogram in such a way that it was difficult to discern what was going on. Figure (b) looks more closely at viral load below 10 000 copies/mℓ. A spike around 400 copies/mℓ can

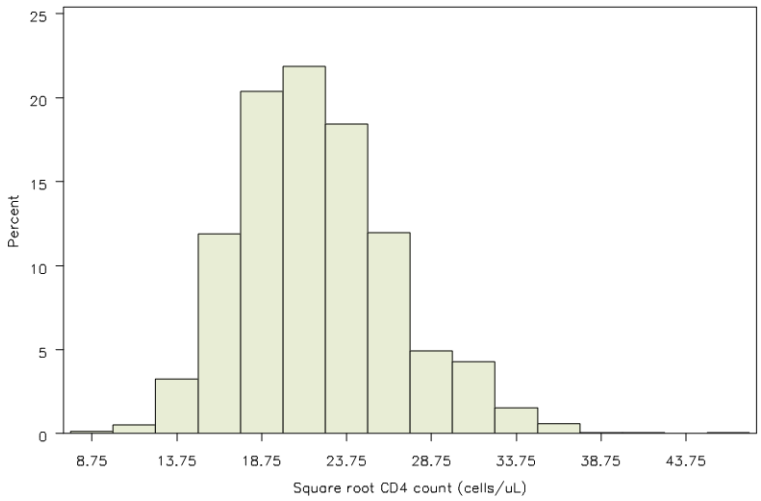

Figure 2.10: Histogram of square root of CD4+ count (cells/$\mu\ell$)

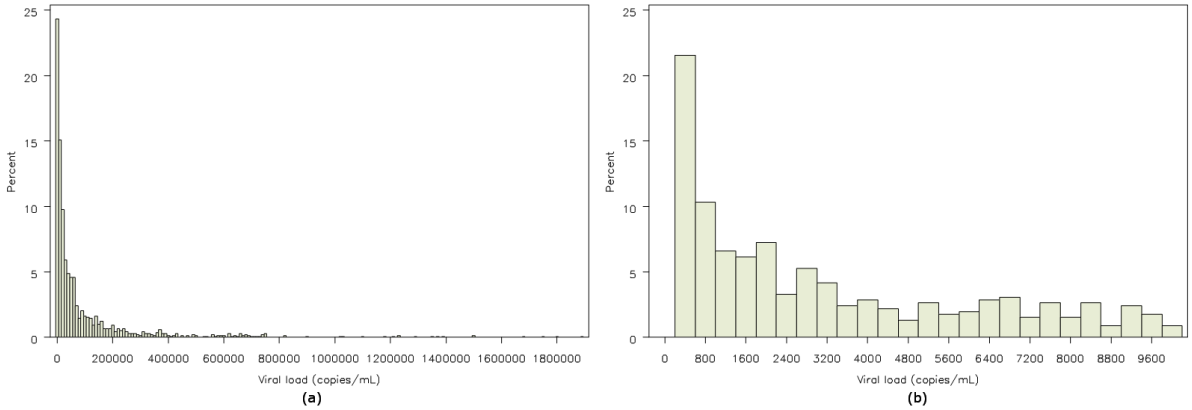

Figure 2.11: Histogram of viral load (copies/mℓ), including only viral load below 2 000 000 copies/mℓ (a) and then viral load below 10 000 copies/mℓ (b)

be seen in these histograms, as this was the lower detection limit of the viral load assays. With viral load being extremely right skewed, a log transformation was performed to normalise the data. Looking at the histogram of log transformed viral load, the data appears more normal after transforming it. However, there is a spike around 2.7 log copies. This is again due to so many viral load measurements being 400 copies/mℓ (which is equivalent to 2.6 log copies/mℓ), because of the lower detection limit of the viral load assay. This highlights the importance of left-censoring. Of the 1384 viral load measurements which were taken at various time points for different individuals throughout the study, 71 measurements were undetectable, accounting for 5.1% of the viral load data.

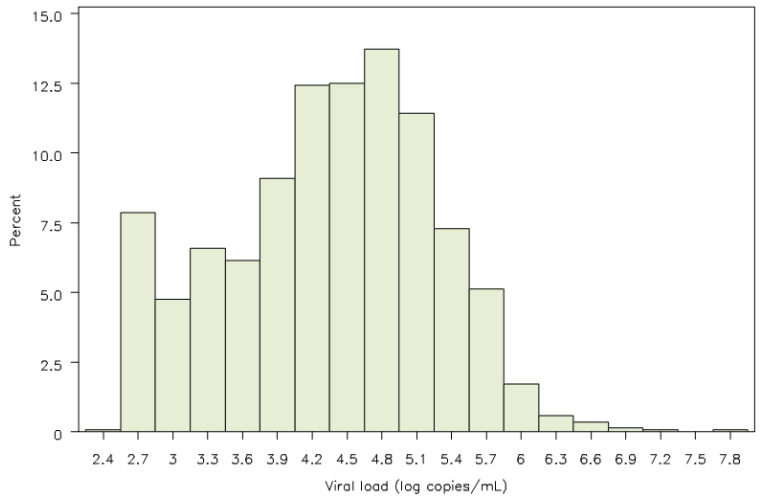

Figure 2.12: Histogram of viral load (log copies/mℓ)

2.2.3 Semi-variograms

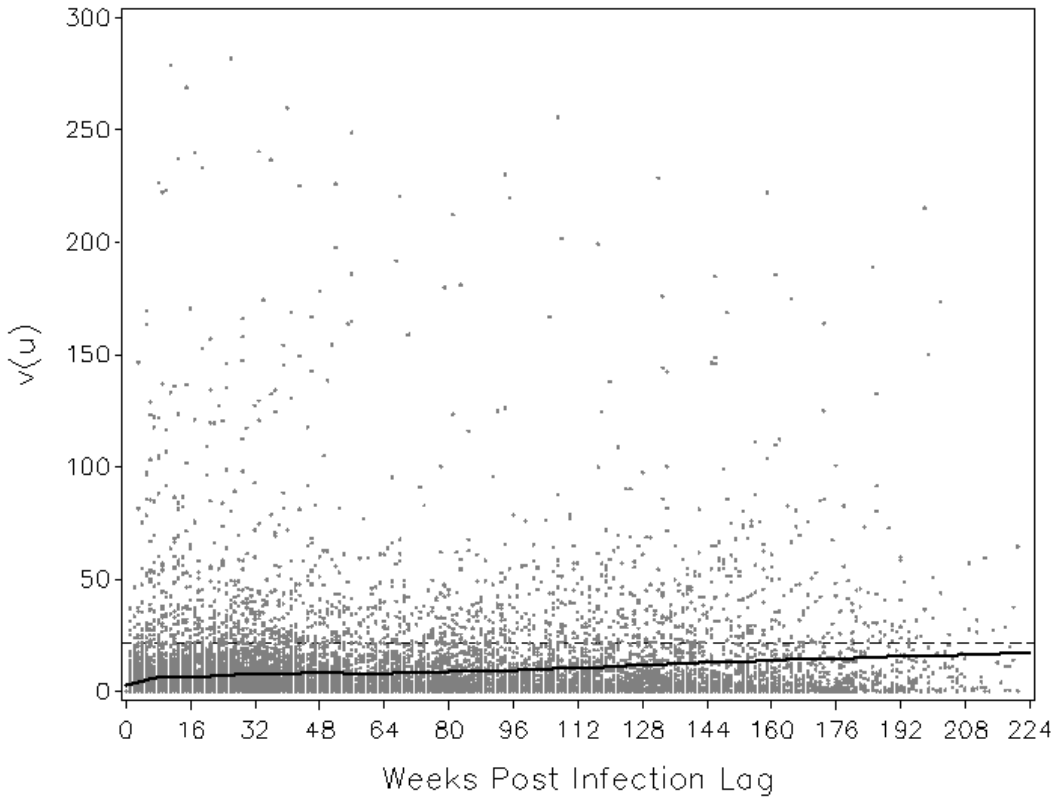

Figure 2.13: Semi-variogram of square root transformed CD4+ count

Figure 2.13 is the semi-variogram of all the CD4+ count measurements. Since the Loess regression

line does not converge to zero as time approaches zero this is an indication that measurement error is present in the data. Since the slope of the Loess line is not zero there is indication of serial correlation. Lastly, the Loess line does not reach the process variance of 21.17 cells/$\mu\ell$ (the horizontal line) and this is further indication of the presence of random or subject-specific effects. Figure 2.14 is the variogram of all log viral load measurements. The Loess regression line crosses

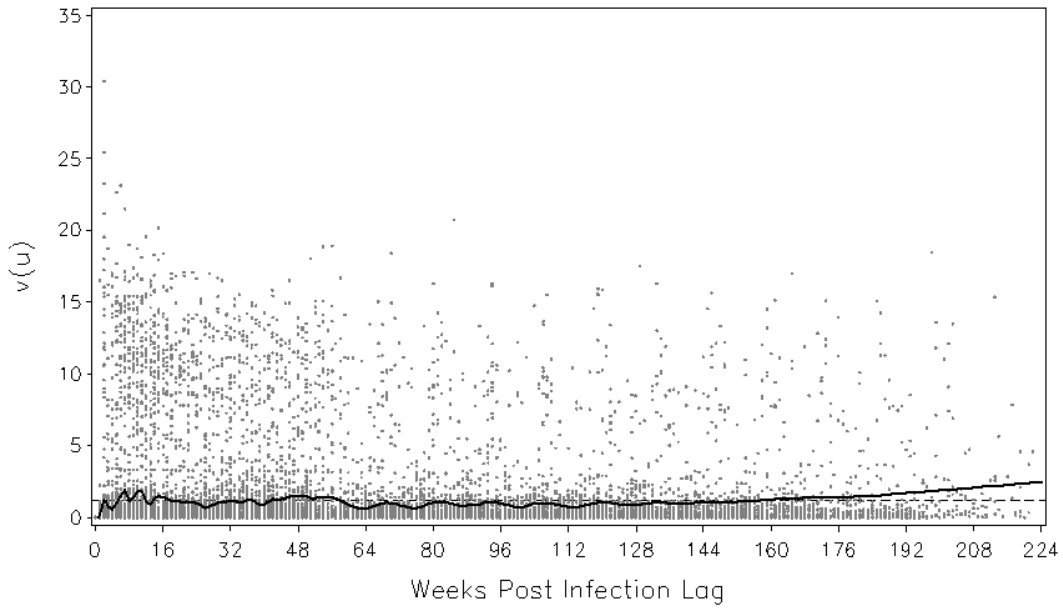

Figure 2.14: Semi-variogram of log transformed viral load measurements

the process variance of 1.18 log copies/mℓ (horizontal line), which indicates that there are no random effects associated with viral load. It reaches zero as time approaches zero, implying zero measurement error. The slope of this line varies over time and this could be an indication of serial correlation.

2.2.4 Sample Correlations

Because of the natural interaction between the immune system and the HIV virions, a correlation between these two markers, CD4+ count and viral load, is expected.

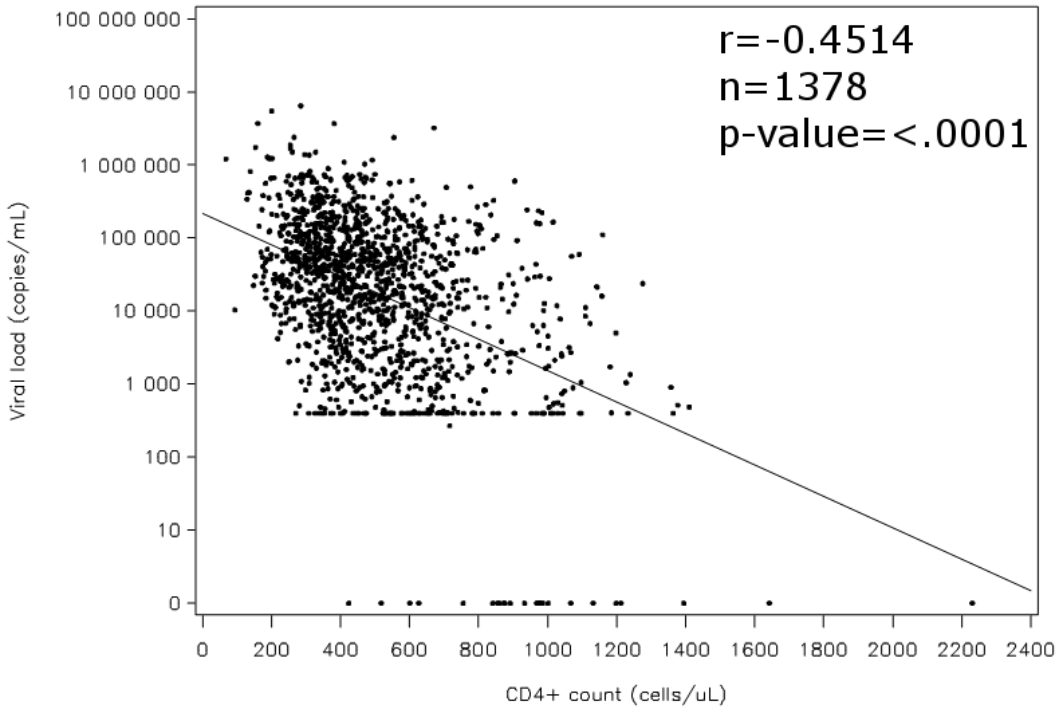

Figure 2.15: Scatterplot of all viral load (on a log scale) and CD4+ count measurements

Figure 2.15 shows a negative correlation between viral load and CD4+ count. Hence, as viral load increases, CD4+ count decreases, which is exactly what is expected given the relationship between these two variables in the absence of treatment. The Pearson's correlation coefficient was -0.4514 (n=1378) and this was statistically significant (p-value (p)<0.0001). When looking at the correlation between CD4+ count and viral load within specified four-weekly intervals, it can be seen that the highest correlation is within the first four weeks of infection, with a correlation coefficient of -0.6418 (p<0.0001). This explains the observed initial behaviour between the two markers where viral load exhibits a sharp increase and decline while CD4+ count exhibits a sharp decline and rebound a few weeks post infection. Figure 2.16 shows that the correlation between CD4+ count and viral load is higher in the first four weeks of infection compared to all later time periods. After the initial month of infection, the correlation between CD4+ count and viral load drops almost 50% but remains significant in all intervals post infection (p<0.0001).

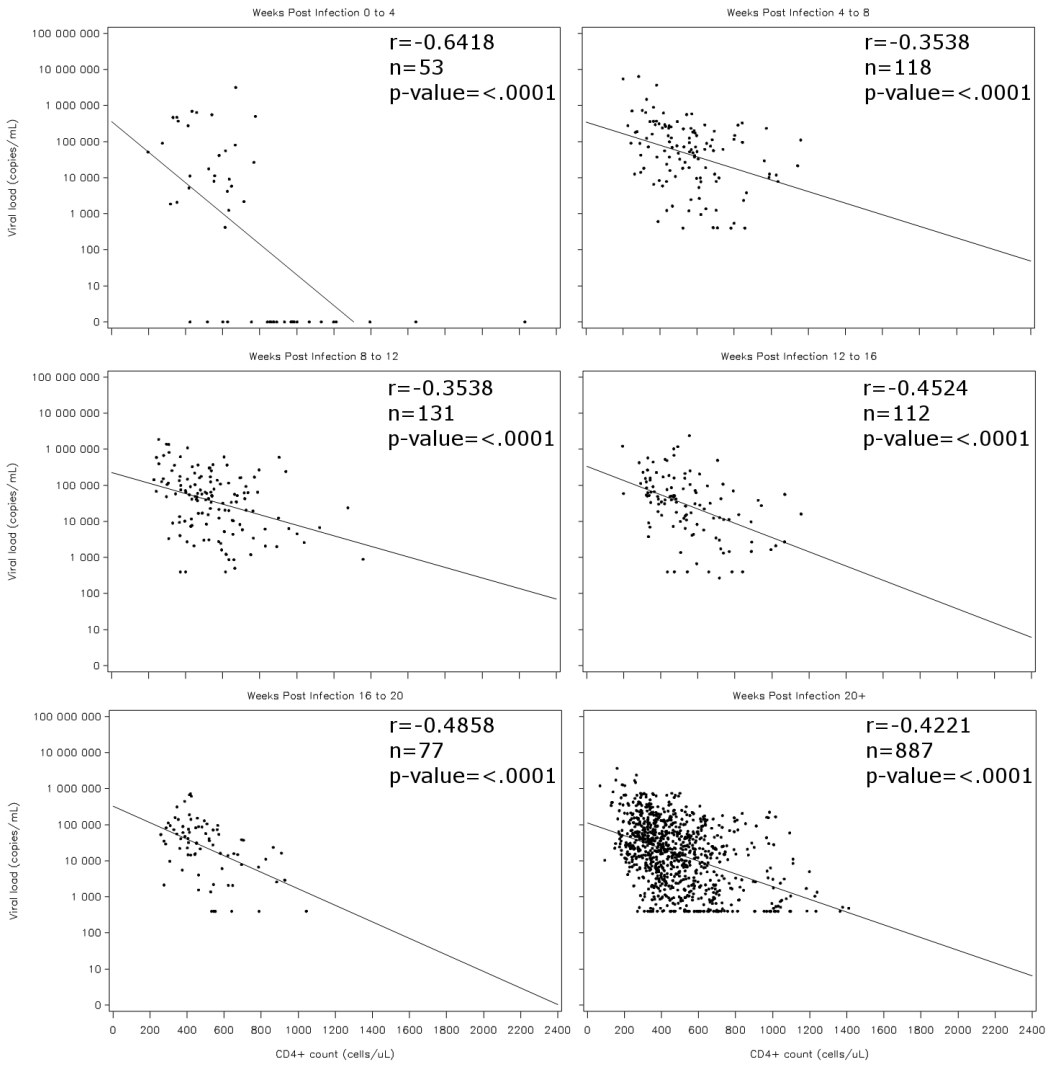

Figure 2.16: Scatter plot of all viral load (on a log scale) and CD4+ count measurements, separated by different intervals weeks post infection

2.3 Discussion

The CAPRISA 002: Acute Infection (AI) Study is a unique cohort since the women infected with HIV are enrolled into the study early, followed intensely and monitored closely. When comparing this cohort to the studies reviewed in section 1.2.1, the AI study proves itself to be invaluable data. It is important to note the differences between the AI study and the other Acute Infection cohorts (Goujard et al., 2006; Lyles et al., 2000). Although the PRIMO cohort described in Goujard et al. (2006) is large (n=552), the AI study offers more precise estimates of infection dates. The PRIMO cohort also follows HIV infected participants, but only implement clinical and laboratory investigations at follow-up month 1, 3 and 6 (and six-monthly thereafter). From what can be seen

in the AI cohort, the most important and interesting changes in viral load and CD4+ count occurs within the first eight weeks post infection, with viral load peaking at approximately two weeks post infection. The MACS cohort also follows participants after infection but again, samples are not taken at such close and intense intervals. The participants enrolled into the AI cohort were followed in CAPRISA 002 and other CAPRISA cohort studies and tested for HIV monthly. Because of this, it is highly likely that a sample could be collected around two weeks post infection since they are enrolled so early after getting infected. The same can be said when comparing the AI study to the MACS cohort (Lyles et al., 2000), which on average enrolled participants around 3 months post infection. Another difference to note is that the AI cohort is infected with HIV-1 subtype C, as opposed to the American or European cohorts which are mainly HIV-1 subtype B. This is very important since it is generally thought that subtype C is a more virulent strain.

Chapter 3

Mixed Models

3.1 Linear Mixed Model

Consider the normal linear regression model

$$Y = X\beta + \varepsilon \tag{3.1}$$

where Y is a vector of observations from a continuous response or dependent variable and X is the design matrix of the independent variables. The usual assumptions are that ε is normally distributed with a mean of zero and variance $\sigma^2\mathrm{I}$, while β is a vector of fixed unknown regression coefficients which explain the dependence of Y on the independent variables in X. The linear regression model is used to construct a simple estimation method in order to predict the outcome of interest and to test whether a given predictor variable in the model has a significant effect on the outcome. In normal linear regression least squares estimation or maximum likelihood are both used to obtain effect estimates. The model can be extended to include categorical explanatory variables, giving rise to the general linear model. Suppose this model is fitted to data where there are a number of categorical factors or variables which will contribute in explaining the response Y, with each factor consisting of a number of levels. It should be noted that each level of a factor can have a different linear effect on the value of the dependent variable. If the interest was restricted to the factor levels included only in this study, the model would be a fixed effects model. However, if the levels of the factor in the data were randomly selected from a population of all possible factor levels then the factor would be called a random effect. A model can contain both fixed and random effects in order to explain the outcome and this would then be a *Linear Mixed Model*. The Linear

Mixed Model has the following general form:

$$Y = X\beta + Zb + \varepsilon \tag{3.2}$$

where Z represents the design matrix for the random effects and b is the vector of the random effect coefficients. Thus model (3.2) has three discernible components, namely the fixed component ($X\beta$), the random component (Zb) and the error components. In contrast, model (3.1) does not explicitly model the random component hence the major distinction between the two. Including random effects in the model is useful since it explains the excess variability in the dependent variable that is not accounted for by the measured covariates. Readily available statistical software such as SAS has a specific procedure, `PROC MIXED`, which can be used to specify the relationship between the response Y and the levels of the random effects. If the covariance structure is not specified, it assumes by default that the levels of the random effects are uncorrelated and have the same variance. This model is explained in detail in sub-section 3.2.1.

In many clinical research studies, data for a continuous response are accrued repeatedly over time. The linear mixed model extends naturally to accommodate repeated or longitudinal data and expands on the general linear mixed model, allowing for the error terms and random effects to be correlated and also allows for possible non-constant variability. Thus a complex covariance structure for Y can be specified and still maintain the normality assumption. This makes it much more flexible to model the mean of the dependent variable and its covariance structure and also to model the relationship between the levels of the repeated effects. The model structure in (3.2) was first formulated by Laird and Ware (1982).

The details of how the linear mixed model is derived follows in Section 3.2. The linear mixed model is a general model which accommodates many linear models as a special case. One such special case is the *general linear longitudinal model* which models data where a number of subjects are followed over time or prospectively. The data from such a study is also known as repeated measurements or longitudinal data. Clearly observations from the same individual are bound to be more correlated than observations from two different individuals. Two ways to model such a dependence is by either (1) modelling the correlation between the repeated measurements leading to the Generalised Estimating Equation (GEE) approach of Liang and Zeger (1986) or (2) by simply adding subject-specific effects in the form of random effects as in Laird and Ware (1982).

Data is considered longitudinal when the same response is repeatedly measured or observed on

the same subject for a given number of time points, resulting in a vector of measurements per individual. Because these measurements are repeatedly observed, the individual effect can be analysed and assessed. This vector of measurements in this thesis is often denoted by Y and it is naturally ordered by time. Thus such data is amenable to several, generally non-equivalent, extensions of univariate models.

One of the aims of fitting a longitudinal model to a data set is to model how the response variable evolves over time for each subject, taking into account several independent variables which affect the response in some way. The effect of the independent variables can also be estimated and their significance in the model can be determined. From a designed experiment consideration, in agriculture an individual or subject can be viewed as a whole plot and the measurement time occasions within an individual sequence of observations as the split plots. Thus an individual effect is not of prime importance and this effect can be viewed as a blocking factor which can be random or fixed while the treatments applied at the split plots are of main interest and whose effects need to be estimated more accurately. In this thesis the time evolution of the HIV disease, using CD4+ count and viral load measurements, is of prime interest while at the same time accounting for individual variability.

3.2 Modelling Longitudinal Data

Longitudinal data is unbalanced when the number of measurements for all subjects are not equal and/or measurements are not taken at fixed time points. This is definitely the case in the CAPRISA 002: Acute Infection Study data since not all the participants have measurements taken at the exact same time points in the study. In addition, information between two time points is unknown because of the discrete nature on how the observations are taken. In such a case it should be clear that some kind of incomplete data is being dealt with.

Some participants arrive late for their scheduled visits or do not arrive at all until the next study visit. This results in missing data which presents an added complexity in modelling longitudinal data. There are several suggested methods of modelling missing data (Molenberghs and Kenward, 2007; Molenberghs and Verbeke, 2005; Verbeke and Molenberghs, 2000). In the case of the AI (Acute Infection) study, the phase II participants (those who are HIV positive), visit the clinic every week after being infected and enrolled into the study, for 2 weeks, then attend bi-weekly for 2 months, and then monthly. They are followed-up for 2 years after seroconverting, until they reach their end point in the study or until they drop out. Under such an intense visit schedule, some

participants are bound to be less adherent to the intended schedule leading to missing information. There are different types of missing data.

Data can be missing completely at random (MCAR) meaning that when an observation is missing it does not depend on observed (Y_i^o) or missing (Y_i^m) responses. On the other hand data can be missing at random (MAR) meaning the probability of missing data is not related to any variable that is unmeasured but the pattern of missing data is predictable from observed data (Y_i^o). When data is not missing at random (NMAR) then the missing data are not random and are related to the values which are missing and possibly also related to observed responses.

Because of the data being unbalanced, it cannot be analysed using standard multivariate regression methods and since there are repeated measures for each subject, both the subject-specific effects as well as the effect of other measured covariates can be modelled. That is, both marginal and random effects can be assessed. To derive the model, a two stage formulation is adopted for the longitudinal data model (Verbeke and Molenberghs, 2000). In the first stage the vector of the repeated measurements is summarized for each subject by a vector of estimated subject-specific regression coefficients. Then in the second stage, multivariate regression methods are used to relate the estimates to measured covariates. In other words, it will be demonstrated how model (3.2) is derived. Since the information is not the same for all individuals, the appropriate statistical methods need to be used to be able to recover inter-individual information in order to estimate the fixed effects.

3.2.1 The General Linear Mixed Effects Model

The General Linear Mixed Model is an extension of the linear model in equation (3.1) to include both fixed and random effects and further to accommodate both cross-sectional and longitudinal data settings. The response vector of observations within an individual is assumed to be multivariate normal. The two stage formulation of such a model is presented below.

Stage 1

Let Y_{ij} denote the j^{th} observation for the i^{th} participant made at time t_{ij}, $i = 1, 2, ..., N$ and $j = 1, ..., n_i$. The response vector Y_i is an n_i-dimensional vector of repeated measurements taken from the i^{th} subject. In the first stage of model building, the individual linear regression model can be written as

$$Y_i = Z_i\beta_i + \varepsilon_i \tag{3.3}$$

where Z_i is a design matrix of dimension $(n_i \times q)$, of known covariates, β_i is a q-dimensional vector of unknown subject-specific regression coefficients which are to be estimated and ε_i is a vector of residual components ε_{ij}, $j = 1, ..., n_i$.

It is assumed that the components of ε_i are normally and independently distributed with a mean vector of zero and a covariance matrix $\sigma^2 I_{n_i}$ (where I_{n_i} is the identity matrix). This model describes a specific regression model and in this way, modelling subject-specific effects.

Stage 2

In stage 2 model formulation, a multivariate regression model for the N regression vector parameters β_i is given by

$$\beta_i = K_i\beta + b_i \tag{3.4}$$

is used to explain the observed variability between the subjects with respect to their subject-specific regression coefficients β_i, where K_i is a matrix of known covariates of dimension $(q \times p)$, β is a p-dimensional vector of unknown regression coefficients and b_i is a q-dimensional vector of subject specific effects, where b_i is assumed to be normally distributed with mean vector zero and general covariance matrix D. It is assumed that $b_i \sim N(0, D)$ where D is the variance-covariance matrix of the random effects. Model 3.4 can be viewed as a means of pooling information from the different subjects in order to estimate population-based regression parameters in β.

Finally, to complete the formulation, β_i in (3.4) is substituted into (3.3) leading to the following derivation.

$$\begin{aligned} Y_i &= Z_i(K_i\beta + b_i) + \varepsilon_i \\ &= X_i\beta + Z_i b_i + \varepsilon_i \end{aligned} \tag{3.5}$$

where Y_i is a vector of responses from individual i of length n_i generally assuming not all scheduled measurements for an individual are necessarily available. The $(n_i \times p)$ design matrix $X_i = Z_i K_i$ is the matrix of p known covariates, or fixed effects, measured alongside the response (assuming no measurement error). The fixed effects are denoted by β, a p-dimensional vector corresponding to the columns of X_i. The subject-specific effects (or random effects) are denoted by b_i, a q-dimensional vector corresponding to Z_i (Note that $q<p$). The n_i-dimensional vector ε_i is a vector of residuals accounting for any unaccounted variation in the model. This means that further structure can be

imposed on ε_i if there is evidence to do so, such as splitting ε_i further into a residual component and serial correlation. The overall distributional assumptions are:

$$b_i \sim N(0, D), \qquad \varepsilon_i \sim N(0, \Sigma_i) \tag{3.6}$$

where D is the covariance matrix of the elements in b_i and Σ_i is the covariance matrix of the elements in ε_i. Most frequently the assumption $\Sigma_i = \sigma^2 I_{n_i}$ is used such that in the absence of or conditional on random effects, this assumption implies the components in $Y_i = (y_{i1}, ..., y_{in_i})'$ measured at times $t_i = (t_{i1}, ..., t_{in_i})'$ are independent. A somewhat unrealistic assumption. Model (3.5) assumes that the vector of repeated measurements on each subject follows a linear regression model where some of the regression parameters are population-averaged, i.e. the same for all subjects, and other regression parameters are subject-specific.

The above model has two important interpretations namely, the marginal and hierarchical interpretation. Marginally, the model for Y_i is

$$Y_i \sim N(X_i\beta, Z_i D Z_i' + \Sigma_i) \tag{3.7}$$

where the mean of Y_i is $X_i\beta$, also called the mean structure in general. The variance of Y_i is $V_i = Z_i D Z_i' + \Sigma_i$, also called the overall covariance structure. The diagonal elements, or parameters, in D and Σ_i are called variance components. It should be noted that under the marginal interpretation of the model, negative variance components are admissible provided the overall variance-covariance matrix V_i is positive definite.

Under the hierarchical interpretation, the model can be specified as

$$Y_i | b_i \sim N(X_i\beta + Z_i b_i, \Sigma_i) \tag{3.8}$$

where $b_i \sim N(0, D)$ and $\varepsilon_i \sim N(0, \Sigma_i)$. Thus under this hierarchical interpretation of the model, negative variance components do not make sense at all.

The diagonal elements of D show how much the individual regression coefficients, the b_{si} vector of random effects ($s = 1, ..., q$) vary from subject to subject, after adjusting for the covariates in the fixed effects β. The random effects can also be thought of as subject-specific regression coefficients, accounting for the natural heterogeneity in the study population. Random effects model the between-subject variability, while Σ_i models the within-subject variability or residual variance.

The structure of D should be specified before the model is fitted and the type of covariance structure often depends on the type of data that is being modelled.

There are several covariance structures that can be assigned to the structure of the covariance matrix for the random effects and the repeated observations within an individual. These covariance structures can be specified in the SAS software in the `RANDOM` and `REPEATED` statements in `PROC MIXED` under the `TYPE=` setting, respectively. Note that the covariance structure for the random effects accounts for the variation in effects between subjects, while for the repeated measures it accounts for the within-subject variation. The most common covariance structures are: variance components (under the independence assumption), autoregressive structure, the compound symmetry and unstructured covariance. The order of the autoregressive structure has to be specified in most statistical packages, but the default being order one or the AR(1) structure.

Assuming a 4x4 variance-covariance matrices, the different forms of the covariance structures can be specified as follows. The variance components (VC) structure is the default in SAS `PROC MIXED` and it assumes that different measurements are independent. It is also known as the SIMPLE structure and has the following form if used for describing the covariance structure for random effects where there are, for example, four random effects in the model.

$$\begin{pmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & 0 \\ 0 & 0 & 0 & \sigma_4^2 \end{pmatrix}$$

The VC or SIMPLE structure has the following form if it is used to describe how the repeated measurements are related if there are, for example, a maximum of four measurements per subject.

$$\begin{pmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{pmatrix}$$

The AR(1) covariance structure has homogeneous variances and correlations that decline with time or distance. It assumes that the variance of any measurements is constant, regardless of when you measure it and also that measurements that are closer to each other in time are more correlated than measurements further away. The AR(1) structure also assumes that data are equally spaced

and has the following structure:

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

The compound symmetry (CS) structure has homogeneous variances and the correlation between any two measurements is assumed to be constant regardless of how far apart they are. Simply put, this covariance structure specifies that all pairs of measurements on the same individual have the same correlation. The CS covariance structure is given by:

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

The unstructured (UN) covariance structure does not assume any particular pattern about the variance and covariance between measurements and allows every variance and covariance to be different. Although this structure seems like the most desirable to fit, it requires the most number of parameters to estimate and can cause computational difficulties. The unstructured covariance matrix has the following form where $\sigma_{ij} = \sigma_{ji}$.

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{pmatrix}$$

Other covariance structures are the so-called spatial covariance structures. The correlations are positive and decreasing functions of the Euclidean distances between observations. Thus when specified for the relationship between repeated measurements, they take into account the distance between the observations within each subject. The coordinates of the measurements are given by a set of variables which are specified in a list written next to the SP option. There are three common types of spatial structures - power, exponential and Gaussian - each one specifying and defining how fast the correlations decrease as functions of the distances between measurements. The power

spatial correlation structure is called by using SP(POW)(*list*) in the `TYPE=` setting, with $d_{ij} = d_{ji}$.

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

The exponential spatial covariance structure, specified with SP(EXP)(*list*) has the following structure:

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

The Gaussian spatial covariance structure, specified with SP(GAU)(*list*) has the following structure:

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

In some literature the covariance structures are presented in terms of the correlation matrices instead of the actual variances and covariances. Note that the spatial structure specifications do not require that the t_{ij}'s be equally spaced. There are a few factors which will influence the decision on which covariance structure to use for the random effects or between observations within an individual. These factors include the number of parameters, the interpretation of the structure, diagnostic results and the effects on the fixed effects. Often it is possible to fit the unstructured covariance structure, but in the expense of more parameters to estimate. Thus it is always better and more efficient to have less parameters to estimate in the model. However, choosing a structure that is too simple, such as the independence and common variance assumption, increases the fixed effects type I error rate.

Another strategy of choosing the appropriate covariance structure is to fit a couple of candidate covariance structures and then use information criteria, such as Akaike's Information Criteria (AIC) or others, to compare the different models and to determine which model has a better fit to the data. However this might not be the best method for selecting the covariance structure because information criteria, such as AIC and Bayesian Information Criteria (BIC), have been shown to

predict the correct model poorly (Ferron, Dailey and Yi, 2002; Keselman et al., 1998).

Alternatively, a graphical method of deciding on the covariance structure can be used. Here one fits the model with an unstructured covariance matrix, using say SAS `PROC MIXED` and make the procedure output the residual correlations and covariances. Then plot the covariances separately for each starting time, i.e. plot lag 1 covariance, lag 2 covariance, etc. for errors starting at time 0. Then the same should be done for errors starting at time 1, then errors at time 2, etc. If there are linearly declining covariances with increasing lags then a AR(1) structure can be fitted and if the lines cross each other then this shows that the random effects have a constant variance (Kincaid, 2005).

3.3 Estimation Procedures

3.3.1 Estimation of the Marginal Model

Recall the structure for the general linear mixed model was derived as

$$Y_i = X_i\beta_1 + Z_i b_i + \varepsilon_i \tag{3.9}$$

where it was assumed that $b_i \sim N(0, D)$, $\varepsilon_i \sim N(0, \Sigma_i)$, and that b_i and ε_i are mutually independent. Recall, the marginal interpretation of the model implies that

$$Y_i \sim N(X_i\beta, Z_i D Z_i' + \Sigma_i) \tag{3.10}$$

where $V_i = Z_i D Z_i' + \Sigma_i$ is the variance-covariance matrix of the vector Y_i, which can be found by setting up the random-effects design matrix Z and specifying the covariance structures for D and Σ. For simplicity, the subscript of i will be ignored. It can be shown mathematically that

$$\begin{aligned} V &= Cov(Y) \\ &= E[Cov(Y|b)] + Cov(E[Y|b])] \\ &= \Sigma + Cov(X\beta + Zb) \\ &= \Sigma + Cov(Zb) \\ &= \Sigma + ZDZ' \end{aligned} \tag{3.11}$$

The unconditional mean of Y is $E(Y) = X\beta$. The unknown parameters to estimate in the model are β, b_i, D and Σ_i. It should be noted that the marginal model does not explicitly assume the

presence of random effects representing the natural heterogeneity between subjects. The following notation will be used in this and subsequent developments. Namely β denotes the vector of fixed effects, α the vector of all variance-covariance parameters, or variance components, in D and Σ_i. Further let $\theta = (\beta',\alpha')'$ denote the vector of all parameters in the marginal model.

An advantage of the linear mixed model is that covariate interactions can be specified which will allow one to estimate the different linear effects allowing for a combination of factor levels that may have an effect on the dependent variable. A major advantage of using the mixed model approach with subject specific effects is that the same number of observations per subject are not required, rather it allows for subjects with missing data. In addition, as in any regression model, time can be treated as a continuous variable instead of having a set of fixed time points.

3.3.2 Maximum Likelihood Estimation of Parameters

3.3.2.1 Estimation of Fixed Effects

The maximum likelihood method is one of the most commonly applied methods of estimation in statistics. This method calculates the values of the model parameters by maximising the likelihood of the data with respect to the model parameters which are to be estimated.

For example, consider the normal distribution whose probability density function is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(- \frac{(x-\mu)^2}{2\sigma} \right) \tag{3.12}$$

Suppose you have n normally, and independently distributed variables $Y_1, Y_2, ..., Y_n$, each with mean μ and variance σ^2. The likelihood of these variables can be written as the continuous joint probability density function:

$$\begin{aligned} L(\mu, \sigma^2) &= f(y_1, y_2, ..., y_n) \\ &= \prod_{i=1}^{n} f(y_i) \\ &= \frac{1}{(\sqrt{2\pi\sigma^2})^n} \prod_{i=1}^{n} \exp\left(- \frac{(y_i-\mu)^2}{2\sigma} \right) \end{aligned} \tag{3.13}$$

Following a similar argument as for the multivariate regression model and using matrix notation, the marginal likelihood function for the full longitudinal data is given by

$$L_{ML} = \prod_{i=1}^{N} \left[(2\pi)^{-\frac{n_i}{2}} |V_i(\alpha)|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(Y_i - X_i\beta)' V_i^{-1}(\alpha)(Y_i - X_i\beta) \right) \right] \tag{3.14}$$

In order to estimate the fixed parameters the log of the likelihood $l = \log L$ needs to be maximized by differentiating it with respect to β and solving the equation for $\frac{\partial l}{\partial \beta} = 0$:

$$\begin{aligned}
\frac{\partial l}{\partial \beta} &= \frac{\partial}{\partial \beta}\left(-\frac{1}{2}(Y_i - X_i\beta)' V_i^{-1}(\alpha)(Y_i - X_i\beta) \right) \\
&= \frac{\partial}{\partial \beta}\left(-\frac{1}{2}(V_i^{-1}(\alpha)Y - V_i^{-1}(\alpha)X_i\beta)'(Y_i - X_i\beta) \right) \\
&= \frac{\partial}{\partial \beta}\left(-\frac{1}{2}(Y_i' V_i^{-1}(\alpha)Y_i - Y_i' V_i^{-1}(\alpha)X_i\beta - \beta' X_i' V_i^{-1}(\alpha)Y_i + \beta' X_i' V_i^{-1}(\alpha)X_i\beta) \right) \\
&= -(X_i' V_i^{-1}(\alpha)X_i\beta - X_i' V_i^{-1}(\alpha)Y_i)
\end{aligned} \tag{3.15}$$

Now equating the derivative to zero implies

$$\begin{aligned}
\frac{\partial l}{\partial \beta} &= 0 \\
X_i' V_i^{-1}(\alpha)X_i\beta - X_i' V_i^{-1}(\alpha)Y_i &= 0 \\
X_i' V_i^{-1}(\alpha)X_i\beta &= X_i' V_i^{-1}(\alpha)Y_i \\
\widehat{\beta} &= (X_i' V_i^{-1}(\alpha)X_i)^{-1} X_i' V_i^{-1}(\alpha)Y_i
\end{aligned} \tag{3.16}$$

The above estimate is only based on information from a single individual. Combining information from all the N profiles, the revised estimate is given by:

$$\hat{\beta}(\alpha) = \left(\sum_{i=1}^{N} X_i' W_i X_i \right)^{-1} \sum_{i=1}^{N} X_i W_i Y_i \tag{3.17}$$

where W_i equals V_i^{-1}. It thus follows, that $\mathrm{E}(\hat{\beta}(\alpha))=\beta(\alpha)$ and $\mathrm{var}(\hat{\beta}(\alpha))=(X_i' W_i X_i)^{-1}$. Note that the estimate above is the same as the weighted generalised least squares (WGLS) estimate. However, in most cases α is unknown, therefore it needs to be replaced by an estimate $\hat{\alpha}$. Two frequently used estimates for α are the maximum likelihood (ML) and restricted maximum likelihood (REML) estimates.

Let the ML estimate of α be denoted by $\hat{\alpha}_{ML}$ for a fixed β. Then $\hat{\alpha}_{ML}$ is obtained by maximising

$$L_{ML}(\alpha) = L_{ML}(\alpha, \beta(\alpha)) \tag{3.18}$$

with respect to α. The resulting estimate $\hat{\beta}(\hat{\alpha}_{ML})$ for β is denoted by $\hat{\beta}_{ML}$, obtained by maximising its profile likelihood $L(\beta_{\hat{\alpha}})$. Alternatively, $\hat{\alpha}_{ML}$ and $\hat{\beta}_{ML}$ can also be obtained from the direct maximisation of L_{ML} with respect to θ, that is with respect to α and β simultaneously (where θ=(α',β')).

3.3.2.2 Prediction of Random Effects

Under the longitudinal mixed effect model structure, individual predictions can be made even if the number of observation points for the specific individual is less than the number of estimated parameters or fixed effects. This comes from the main assumption that each individual has its own subject-specific parameter $\beta_i = \beta + b_i$, whose expectation is the population mean parameter β. The advantage with repeated measurements is that individual observations n_i may be small, but since there are several individuals, this enhances information required to estimate the required model parameters. Random effects can be predicted under maximum likelihood estimation. If $\text{cov}(b, Y)$=DZ', where Y is the response vector from a typical individual and b the vector of individual specific parameters, then

$$\begin{pmatrix} Y \\ b \end{pmatrix} \sim N\left\{ \begin{pmatrix} X\beta \\ 0 \end{pmatrix}, \begin{pmatrix} V & ZD \\ DZ' & D \end{pmatrix} \right\}$$

and the random effects can be predicted as

$$\begin{aligned} \hat{b} &= E(b|Y) \\ &= E(b) + cov(b|Y)[var(Y)]^{-1}[Y - E(Y)] \\ &= DZ'V^{-1}(Y - X\beta) \end{aligned} \tag{3.19}$$

It can be shown that

$$E(\hat{b}) = 0 \tag{3.20}$$

and

$$var(\hat{b}) = DZ'V^{-1}D - D(Z'V^{-1}X)(X'V^{-1}X)^{-1}(X'V^{-1}Z)D' \tag{3.21}$$

where X is the design matrix for fixed effects which are common to all individuals. Sometimes the fact that $\hat{b} = E(b|Y)$ estimation of b is referred to as prediction of random effects to emphasise the conditional posterior structure of the equation (3.19), similar to Bayesian estimation. It should be noted that since the methods for predicting random effects are non-Bayesian in formulation, the random effect estimates are empirical Bayes (EB), which are the means of the conditional random effects distribution, given the overall observed data and plug-in estimates of hyper-parameters.

The hyper-parameters are estimated marginally through likelihood-based methods from the data; thus the EB methods underestimate the variability in the random effects since uncertainty in the hyper-parameters is not allowed. This results in narrow confidence intervals from EB methods, which in turn leads to apparent significant results.

3.3.3 Restricted Maximum Likelihood Estimation

To develop the concept of REML estimation, consider a sample of N observations or measurements $Y_1, Y_2, ..., Y_N$ from $N(\mu, \sigma^2)$. Given the mean μ is known the maximum likelihood estimate, or MLE, of σ^2 will now be given by

$$\hat{\sigma}^2_{ML} = \sum_{i=1}^{N} (Y_i - \mu)^2 / N \tag{3.22}$$

In this case, $\hat{\sigma}^2_{ML}$ is an unbiased estimator for σ^2. However if μ is unknown it is replaced by $\bar{Y}$, the sample mean. The MLE of σ^2 is now given by

$$\hat{\sigma}^2_{ML} = \sum_{i=1}^{N} (Y_i - \bar{Y})^2 / N \tag{3.23}$$

Now $\hat{\sigma}^2_{ML}$ is a biased estimator of σ^2, because $\mathrm{E}(\hat{\sigma}^2_{ML}) = (1 - N^{-1})\sigma^2$, leading to a bias of $-N^{-1}\sigma^2$ and is thus biased downwards. The biased expectation leads to the conclusion that an unbiased estimate for σ^2 when μ is unknown should be

$$s^2 = \sum_{i=1}^{N} (Y_i - \bar{Y})^2 / (N - 1) \tag{3.24}$$

The above discussion shows that having to estimate μ introduces bias into the maximum likelihood estimation of σ^2. Thus one way to circumvent this problem is to find a way of estimating σ^2 without having to estimate μ first. Note that all the data can be combined into one distributional model

$$Y = \begin{pmatrix} Y_1 \\ : \\ Y_N \end{pmatrix} \sim \left(\begin{pmatrix} \mu \\ : \\ \mu \end{pmatrix}, \sigma^2 I_N \right) \tag{3.25}$$

or $Y \sim N(\mu 1_N, \sigma^2 I_N)$ where 1_N is a N-dimensional vector full of 1's and I_N the N-dimensional identity matrix. One way to avoid estimation of μ is to transform the vector of observations Y such that μ vanishes from the likelihood. Let U be such a transformation where $U_i = Y_i - Y_{i+1}$,

then

$$U = \begin{pmatrix} Y_1 - Y_2 \\ Y_2 - Y_3 \\ \vdots \\ Y_{N-1} - Y_N \end{pmatrix} = A'Y \sim N(0, \sigma^2 A'A) \tag{3.26}$$

where A is a $(N-1) \times N$ matrix with elements $A_{i,i} = 1$, $A_{i,i+1} = -1$ and zero elsewhere. Based on this transformation the MLE of σ^2 is exactly as given by s^2 which is unbiased for σ^2. The transformation operator A defines a set of $N-1$ linearly independent error contrasts and s^2 is called the REML estimate of σ^2, and is independent of A. The above formulation can be extended to the case of the linear regression model. Thus consider a set of N observations $Y_1, Y_2, ..., Y_N$ from a normal linear regression model $Y \sim N(X\beta, \sigma^2 I_N)$ where Y is a N-dimensional vector of observations, X is the design matrix, β is a vector of regression parameters and σ^2 the residual variance. Following the above arguments the MLE of σ^2 for the linear regression model is

$$\sigma^2_{ML} = (Y - X\hat{\beta})'(Y - X\hat{\beta})/N \tag{3.27}$$

and the REML estimate is given by

$$\sigma^2_{REML} = (Y - X\hat{\beta})'(Y - X\hat{\beta})/(N - p) \tag{3.28}$$

where p is the number of parameters in β. The REML estimate can also be obtained by transforming the data orthogonal to X that is, from Y to

$$U = A'Y \sim N(0, \sigma^2 A'A) \tag{3.29}$$

In essence, REML estimation involves applying the maximum likelihood (ML) method to linear functions of Y, say $A'Y$, for which A' is designed so that $A'Y$ contains none of the fixed effects which are apart of the model for Y. The idea behind REML estimation is to adjust the variance which will protect it against potential bias that comes from using ML estimation. Two important consequences of using REML estimation is that the variance components are estimated without being affected by the fixed effects, making them invariant to the values of the fixed effects. Secondly, REML estimation takes into account the degrees of freedom of the fixed effects implicitly, whereas in ML estimation these are not taken into account.

REML estimation is used for estimating variance components. By choosing A such that $A'Y$ contains no fixed effects will result in $A'X = 0$. Maximum likelihood estimation will be done on

$A'Y$ instead of Y. Thus $A'Y \sim N(0, \sigma^2 A'A)$ and the ML equations for $A'Y$ can be derived from those for $Y \sim N(X\beta, \sigma^2)$ by replacing Y with $A'Y$, X with $A'X = 0$ and σ with $\sigma^2 A'A$.

3.3.4 REML Estimation for the General Linear Mixed Model

Let Y_i denote the individual n_i-dimensional vector of observations that is $Y_i = (Y_{i1}, Y_{i2}, ..., Y_{in_i})'$ where it is assumed that $Y_i \sim N(X_i\beta, V_i)$. The strategy is first to combine the N subject-specific information into one augmented vector Y such that $Y \sim N(X\beta, V)$ where

$$Y = \begin{pmatrix} Y_1 \\ : \\ Y_N \end{pmatrix}, \quad X = \begin{pmatrix} X_1 \\ : \\ X_N \end{pmatrix}, \quad V(\alpha) = \begin{pmatrix} V_1 & 0 & ... & 0 \\ 0 & V_2 & ... & 0 \\ : & : & ... & : \\ 0 & 0 & ... & V_N \end{pmatrix} \tag{3.30}$$

Next the data are transformed orthogonal from X to $U = A'Y \sim N(0, A'V(\alpha)A)$ where U is a vector of error contrasts defined earlier. The MLE of α, based on U is called the REML estimate and is denoted by $\hat{\alpha}_{REML}$. The resulting estimate for β will be denoted by $\hat{\beta}_{REML}$. Alternatively $\hat{\alpha}_{REML}$ and $\hat{\beta}_{REML}$ can also be obtained from maximising

$$L_{REML}(\theta) = |\sum_{i=1}^{N} X_i' W_i(\alpha) X_i|^{-\frac{1}{2}} L_{ML}(\theta) \tag{3.31}$$

with respect to $\theta = (\alpha', \beta')$, i.e. with respect to α and β simultaneously. Note that the expression above for $L_{REML}(\theta)$ is $L_{ML}(\theta)$ but subjected to a penalty. $L_{REML}(\alpha, \hat{\beta}(\alpha))$ is the likelihood of the error contrasts U often called the REML likelihood function. It is important to note that $L_{REML}(\theta)$ is not a likelihood of the original data Y. Solution for the ML (and REML) equations is usually acquired using numerical iterative methods, such as the Newton-Rhapson or Fisher's Scoring methods. They are the best known methods for finding successively better approximations to the zero of a function - hence they are termed root-finding algorithms.

3.4 Inference

3.4.1 Inference and Testing for the Marginal Model

After fitting the correct model to the data, interest also lies in testing hypotheses about the parameters which have been estimated. For fixed effects, a hypotheses of the form that difference between levels of a factor are zero, or some constant, can be tested. The inference for the estimated parameters in the marginal model will be briefly discussed, for both the mean model (or fixed

effects) and the variance components. In particular, the Wald test, the t-test, the F-test, robust inference and the likelihood ratio (LR) test for fixed effects are revisited. For variance components, the methods that will be given attention to are the Wald and the LR tests. The information criteria (IC) for making inference will also be discussed. Recall that the estimate for β is given by

$$\hat{\beta}(\alpha) = \left(\sum_{i=1}^{N} X_i' W_i X_i \right)^{-1} \sum_{i=1}^{N} X_i W_i y_i \tag{3.32}$$

with α replaced by its ML or REML estimate. It follows that conditional on α, $\hat{\beta}(\alpha)$ is MVN with mean β and covariance

$$var(\hat{\beta}) = \left(\sum_{i=1}^{N} X_i' W_i X_i \right)^{-1} \left(\sum_{i=1}^{N} X_i' W_i' var(Y_i) W_i X_i \right) \left(X_i' W_i X_i \right)^{-1} \tag{3.33}$$

which is simplified by

$$var(\hat{\beta}) = \left(\sum_{i=1}^{N} X_i' W_i X_i \right)^{-1} \tag{3.34}$$

if the assumption var(Y_i)=V_i=W_i^{-1} holds.

3.4.1.1 Approximate Wald Test

For each fixed effect, an approximate Wald test can be obtained from approximating the distribution of $(\hat{\beta}_j - \beta_j)/SE(\hat{\beta}_j)$, $j = 1, 2, ..., p$, by a standard normal distribution. For any known matrix L, the test statistic for the hypothesis

$$H_0 : L\beta = 0 \qquad vs \qquad H_A : L\beta \neq 0 \tag{3.35}$$

is given by the Wald test statistic

$$W_s = \hat{\beta}' L' \left(L \left(\sum_{i=1}^{N} X_i' V_i^{-1} X_i \right)^{-1} L' \right)^{-1} L\hat{\beta} \tag{3.36}$$

which is asymptotically χ^2 distributed with degrees of freedom equal to rank(L) under H_0.

The Wald test is based on the variance of the fixed effects estimate $\hat{\beta}$,

$$var(\hat{\beta}) = \left(\sum_{i=1}^{N} X_i' W_i(\alpha) X_i \right)^{-1} \tag{3.37}$$

However, it should be noted that the variability introduced by replacing the variance components α by some estimate (ML or REML) is not taken into account in Wald tests. Thus the Wald test will only provide valid inferences in sufficiently large samples.

3.4.1.2 Approximate t- and F-test

The downward bias from using the Wald test can be resolved by using an F-test for the hypothesis

$$H_0 : L\beta = 0 \qquad vs \qquad H_A : L\beta \neq 0 \tag{3.38}$$

where the F-statistic is

$$F_s = \frac{\hat{\beta}' L' (L(\sum_{i=1}^{N} X_i' V_i^{-1} X_i)^{-1} L')^{-1} L\hat{\beta}}{rank(L)} \tag{3.39}$$

where the approximate null-distribution for F_s is the F distribution with numerator degrees of freedom equal to rank(L). The denominator degrees of freedom can be estimated using methods, such as the containment method, Satterthwaite approximation, or the Kenward & Roger approximation. For longitudinal data, all methods would lead to large numbers of degrees of freedom, and therefore to similar p-values for the different methods. For a univariate hypothesis, rank(L)=1 and the F-test is equivalent to a t-test.

3.4.1.3 Robust Inference

Since

$$\hat{\beta}(\alpha) = \left(\sum_{i=1}^{N} X_i' W_i X_i \right)^{-1} \sum_{i=1}^{N} X_i' W_i y_i \tag{3.40}$$

with α replaced by either its ML or REML estimate, it implies that E[$\hat{\beta}(\alpha)$]=β provided E(Y_i)=$X_i\beta$. Hence in order for $\hat{\beta}$ to be unbiased it is sufficient that the mean of the response be correctly specified. Further, conditional on α, $\hat{\beta}$ has covariance

$$var(\hat{\beta}) = \left(\sum_{i=1}^{N} X_i' W_i(\alpha) X_i \right)^{-1} = C_N \tag{3.41}$$

provided var(Y_i) is correctly modelled as $V_i = Z_i D Z_i' + \Sigma_i$. The covariance estimate C_N is called the "naive"estimate. The so-called "robust"estimate for var($\hat{\beta}$), which is denoted by C_R, does not require a correct specification of var(Y_i) in (3.33). Rather, it is obtained by replacing var(Y_i) by $(Y_i - X_i\hat{\beta})(Y_i - X_i\hat{\beta})'$), which is the empirically based estimate of var(Y_i). The robust variance estimate of var($\hat{\beta}$) is called the sandwich estimator and based on this, robust versions of the Wald, t- and F-tests can be derived.

The above analysis suggests that as long as interest is only on inference of the mean structure, little effort may be spent in modelling the covariance structure of Y_i, provided the data is sufficiently large. However this is not to say an appropriate covariance modelling is not of interest. An appropriate covariance structure is still of interest for gaining efficiency in parameter estimation. In addition, in the presence of missing data, robust inference is only valid under very restrictive assumptions about the underlying missingness process such as data be missing completely at random (MCAR).

3.4.1.4 Likelihood Ratio Test

The likelihood ratio test is commonly used to compare two models when one model is a special case of the other. In the case where two models with different mean structures (or fixed effects) but with equal covariance structures are compared, the hypothesis is

$$H_0 : \beta \epsilon \Theta_{\beta,0}; V_i = \Gamma \qquad vs \qquad H_A : \beta \epsilon \Theta_{\beta}; V_i = \Gamma \tag{3.42}$$

The second part of the statement of each hypothesis is to emphasise that the covariance structure of the data is the same in both cases. As before, let L_{ML} denote the ML function and let the ML estimates under H_0 and H_A be $\hat{\theta}_{ML,0}$ and $\hat{\theta}_{ML}$, respectively. Then the likelihood ratio test statistic is given by

$$T_{\beta,LR} = -2 \ln \lambda_N = -2 \ln \left[\frac{L_{ML}(\hat{\theta}_{ML,0})}{L_{ML}(\hat{\theta}_{ML})} \right] \tag{3.43}$$

where

$$\lambda_N = \frac{L_{ML}(\hat{\theta}_{ML,0})}{L_{ML}(\hat{\theta}_{ML})} \tag{3.44}$$

is the ratio of the likelihoods under H_0 and H_A. Thus the closer λ_N is to 1, the more probable that H_0 is true. When H_0 is not true, $T_{\beta,LR}$ will be large and positive, therefore providing evidence that H_0 is not true. The asymptotic null distribution of $T_{\beta,LR}$ is χ^2 with degrees of freedom equal to dim(Θ_β)-dim($\Theta_{\beta,0}$). Note that as stated earlier, it is assumed that dim($\Theta_{\beta,0}$)$\subseteq$dim(Θ_β).

However, it should be noted that the LR tests are not valid under REML estimation because here the response vector Y is first transformed into error contrasts $U = A'Y$, for some matrix of constants A such that $A'X = 0$. Then ML estimation is done on U as the data. The likelihood value $L_{REML}(\hat{\theta})$ is the likelihood at the maximum based on the error contrasts U. Thus the model with different mean structures lead to different REML error contrasts hence the subsequent likelihoods are not comparable.

3.4.2 Inference for the Variance Components

Most often it is the inference for the mean structure that is usually of primary interest. However, inferences for the covariance structure could also be of interest as well, for obvious reasons, among them the interpretation of the random variation in the data. It is also important to note that an over-parameterized covariance structure (e.g. the UN structure) may lead to inefficient inferences for the mean model (due to overspending on the degrees of freedom in estimating the variance-covariance components). On the other hand, a covariance model which is too restrictive will invalidate inferences for the mean structure. The best covariance model is therefore a balance between a fully unstructured model and the independence assumption.

3.4.2.1 Approximate Wald Test

Asymptotically, ML and REML estimates of α are normally distributed with correct mean and inverse Fisher information matrix as covariance. Therefore approximate standard errors and Wald tests can easily be obtained. However, there is need for caution in the context of the hierarchical model in relation to the marginal model interpretation. A null hypothesis of a zero variance component is meaningful only under the marginal model when no underlying random effects structure is believed to describe the data. The quality of the normal approximation for $\hat{\alpha}_{ML}$ and $\hat{\alpha}_{REML}$ estimates strongly depends on the true value of α. The approximation is poor once α is relatively close to the boundary of the parameter space. If α is a boundary value then the normality approximation fails completely. Under the hierarchical normal interpretation, a null hypothesis of a zero variance component implies the p-value is based on an incorrect null distribution for the Wald test statistic. The test is only correct when the null hypothesis is not a boundary value. However, even under the hierarchical model interpretation a Wald test is valid for testing a zero covariance parameter such as $d_{12} = 0$ versus $d_{12}>0$ where $d_{12} = d_{21} = \text{cov}(b_{i1}, b_{i2})$ is the covariance between the first and second individual specific parameters or random effects.

3.4.2.2 The Likelihood Ratio Test

The Likelihood Ratio (LR) test is best for comparing nested models with equal mean structures but different covariance structures. The null hypothesis of interest is similar to that of the mean structure, namely

$$H_0 : \alpha \epsilon \Theta_{\alpha,0} \qquad vs \qquad H_A : \alpha \epsilon \Theta_{\alpha} \tag{3.45}$$

where $\Theta_{\alpha,0} \subset \Theta_\alpha$. Let $\hat{\alpha}_{ML,0}$ and $\hat{\alpha}_{ML}$ be the MLEs under H_0 and H_A. Then the LR test statistic is given by

$$T_\alpha = -2\ln\lambda_N = -2\ln\left[\frac{L_{ML}(\hat{\alpha}_{ML,0})}{L_{ML}(\hat{\alpha}_{ML})}\right] \tag{3.46}$$

The asymptotic null distribution of T_α is χ^2 with degrees of freedom equal to the difference in dimensions of Θ_α and $\Theta_{\alpha,0}$. Now, as long as the comparison is under the same mean structure, a valid LR test can still be obtained under REML since the error contrasts U are the same in both cases.

It is important to note that when the likelihood ratio test is used to compare covariance models, keeping the mean structures (or fixed effects) the same for both models is necessary. When testing two fixed effects models, the covariance structure is kept the same in both models. This is done so that when, for example, two models with different sets of fixed effects are compared and a significant difference is detected, then the difference is solely as a result of the difference in fixed effects and not because of a better covariance structure for the one model and not the other. Likewise, the same applies when the aim is to compare two models with different covariance structures.

3.4.3 Marginal Testing for the Need of Random Effects

Under the hierarchical model interpretation, the asymptotic null distribution for the LR test statistic for significance of all variance components related to one or multiple random effects, can be derived. For example, consider the hypothesis of no random effects versus one random effect model

$$H_0 : D = 0 \qquad vs \qquad H_A : D = d_{11} \tag{3.47}$$

for some scalar d_{11}>0. Here the asymptotic null distribution of T_α is $\chi^2_{0:1}$, a mixture of χ^2_0 and χ^2_1 with 50:50 weights. Under H_0, T_α equals 0 in 50% of the cases. Intuitively, the extended parameter space $\boldsymbol{R}$ can be considered for d_{11}. Under H_0, $\hat{d}_{11}$ will be negative in 50% of the cases which means that under the restriction d_{11}>0, these cases lead to $\hat{d}_{11} = 0$. Hence $L_{ML}(\hat{\alpha}_{ML}, 0) = L_{ML}(\hat{\alpha}_{ML})$ in 50% of the cases. In general to test the hypothesis of q versus $q+1$ random effects, that is

$$H_0 : D = D_{q\times q} \qquad vs \qquad H_A : D = D_{(q+1)\times(q+1)} \tag{3.48}$$

the null distribution of T_α is distributed as $\chi^2_{q:q+1}$, a mixture of χ^2_q and χ^2_{q+1} with equal weights $w_q = w_{q+1} = 0.5$. However, to test the hypothesis of q versus $q+k$ random effects where under H_0, D is $q \times q$ and under H_A, D is a $(q+k)\times(q+k)$ positive definite matrix, simulations are needed to

derive the asymptotic null distribution. Practically, correcting for the boundary problem reduces the p-value. On the other hand, ignoring the boundary problem too often leads to oversimplified covariance structures which may lead to invalid inferences, even for the mean structure.

3.4.4 Information Criteria (IC)

Note that the general idea behind the LR test for comparing model A to a more extended model B, is to select model A if the increase in L under model B is small compared to increase in complexity. Thus to compare non-nested models, the model with the largest likelihood is selected, provided it is not too complicated. Under the IC method, the model with the highest penalised log-likelihood l-Pen(θ) for some penalty function Pen(.) dependent on the number of parameters, $\#\theta$, is selected. Different forms of Pen(.) lead to different criteria. Some commonly used ones are listed in the table below:

Criteria	Penalty
Akaike (AIC)	Pen($\#\theta$)=$\#\theta$
Schwarz (BIC)	Pen($\#\theta$)=($\#\ln n^*$)/2
Hannan and Quinn (HQIC)	Pen($\#\theta$)=$\#\ln(\ln n^*)$
Bozdogan (CAIC)	Pen($\#\theta$)=$\#\theta(\ln n^* + 1)/2$

However, it should be emphasised here that IC are not formal testing procedures. For the same reason mentioned earlier for comparing models with different mean structures, IC should be based on ML rather than REML, because REML values will be based on different sets of error contrasts and therefore no longer comparable.

3.4.5 Inference for the Random Effects

In this section the problem of making inference on the random effects b_i is addressed. In particular, the idea of empirical Bayes (EB) and best linear unbiased predictors will be given attention (BLUP). The concept of shrinkage estimators will be derived and the normality assumption for random effects discussed. The random intercepts and slopes model will be used as a special case.

3.4.5.1 Empirical Bayes Inference

Consider the linear mixed model

$$Y_i = X_i\beta + Z_i b_i + \varepsilon_i \tag{3.49}$$

under consideration where $b_i \sim N(0, D)$, $\varepsilon_i \sim N(0, \Sigma_i)$ and b_i and ε_i are independent. The random effects b_i reflect how the evolution for the ith subject deviates from the expected evolution $X_i\beta$. Estimation of the random effects b_i is helpful for detecting outlying profiles from the expected profile. Thus inference from random effects is only meaningful under the hierarchical model assumptions where

$$Y_i|b_i \sim N(X_i\beta + Z_ib_i, \Sigma_i) \tag{3.50}$$

and

$$b_i \sim N(0, D) \tag{3.51}$$

Since b_i behaves like random "parameters", it is most natural to consider Bayesian-like approaches where the prior distribution of the random parameters (here random effects) is $b_i \sim N(0, D)$. Thus using the Bayes rule the posterior distribution of the b_i, given the data $Y_i = y_i$, can be expressed as

$$f(b_i|y_i) = \frac{f(y_i|b_i)f(b_i)}{\int f(y_i|b_i)f(b_i)\delta b_i} \tag{3.52}$$

Since the marginal distribution of b_i and the conditional distribution of $Y_i|b_i$ is known, the posterior distribution of b_i, given Y_i is

$$b_i|y_i \sim N(DZ_i'W_i(y_i - X_i\beta), \Lambda_i) \tag{3.53}$$

for some positive definite matrix Λ_i. Thus it can be seen that the posterior mean of b_i, given y_i, as an estimate of b_i is

$$\hat{b}_i(\theta) = E(b_i|Y_i = y_i) = \int f_{b_i|y_i}(b_i|y_i)\delta b_i = DZ_i'W_i(\alpha)(y_i - X_i\hat{\beta}) \tag{3.54}$$

and the covariance of $\hat{b}_i(\theta)$ is given by

$$var(\hat{b}_i(\theta)) = DZ_i'\left(W_i - W_iX_i\left(\sum X_i'W_iX_i\right)^{-1}X_i'W_i\right)Z_iD \tag{3.55}$$

However inference on b_i should take into account the variability in b_i, therefore inference for b_i is usually based on

$$var(\hat{b}_i(\theta) - b_i) = D - var(\hat{b}_i(\theta)) \tag{3.56}$$

It follows that once the correct variance in (3.56) is found, Wald tests can be performed in order to test hypotheses about $b_i(\theta)$. Parameters in θ are replaced by their ML and REML estimates, obtained after fitting the marginal model. The estimate $\hat{b}_i = \hat{b}_i(\theta)$ is called the empirical Bayes estimate (EB) of b_i. Approximate t and F tests to account for the variability introduced by replacing θ by $\hat{\theta}$ can be constructed similarly to tests for fixed effects.

3.4.5.2 Best Linear Unbiased Prediction

Often parameters of interest are linear combinations of fixed effects in β and random effects in b_i. For example, a subject-specific slope is the sum of the average slope for subjects with same covariate values and the subject-specific random slope for that subject. In general suppose that

$$\gamma = l'_\beta \beta + l'_b b_i \tag{3.57}$$

is the quantity of interest. Then conditionally on α,

$$\hat{\gamma} = l'_\beta \hat{\beta} + l'_b \hat{b}_i \tag{3.58}$$

is the best linear unbiased predictor of γ. Note that $\hat{\gamma}$ is linear in the observations Y_i, unbiased and minimises the variance among all unbiased linear estimators. In practise, histograms and scatter plots of certain components of b_i can be used to detect subjects with exceptional or extreme evolutions over time. The predicted evolution of the ith subject is given by

$$\begin{aligned}
\hat{Y}_i &= X_i\hat{\beta} + Z_i\hat{b}_i \\
&= X_i\hat{\beta} + Z_i D Z_i' V_i^{-1}(y_i - X_i\hat{\beta}) \\
&= (I_{n_i} - Z_i D Z_i' V_i^{-1}) X_i\hat{\beta} + Z_i D Z_i' V_i^{-1} y_i \\
&= \Sigma_i V_i^{-1} X_i\hat{\beta} + (I_{n_i} - \Sigma_i V_i^{-1}) y_i
\end{aligned} \tag{3.59}$$

which is a weighted average of the population-averaged profile $X_i\hat{\beta}$ and the observed individual data y_i, with weights $\Sigma_i V_i^{-1}$ and $(I_{n_i} - \Sigma_i V_i^{-1})$, respectively. Note that $X_i\hat{\beta}$ gets more weight if the residual variability is large compared to the total variability given by $V_i = Z_i D Z_i' + \Sigma_i$. This phenomenon is the so-called "shrinkage", meaning the observed data are shrunk towards the prior marginal average profile $X_i\beta$, depending on the degree of how much within-subject variability there is. This is also reflected in the fact that for any linear combination $l'b_i$ of random effects

$$var(l'\hat{b}_i) \leq var(l'b_i) \tag{3.60}$$

A simple example is now considered to demonstrate some of the concepts raised above for purposes of clarity. Consider the random intercepts model given by

$$y_{ij} = \beta_0 + b_{0i} + \beta_1 t_{ij} + \varepsilon_{ij} \tag{3.61}$$

where y_{ij} is the jth observation from the ith individual in the study for $i = 1, 2, ..., N$ and $j = 1, 2, ..., n_i$. β_0 is the average intercept, b_{0i} is the subject-specific intercept which is a random effect assumed to be distributed as $N(0, d_0^2)$, β_1 is the common average slope for all individuals which is assumed not to suffer from between-subject variability, t_{ij} is the actual time measurement and ε_{ij} is the measurement error or residual. Following the above model derivations, it follows that the empirical Bayes estimate for the random intercept b_{0i} is given by

$$\begin{aligned}
\hat{b}_{0i} &= DZ_i' W_i(\alpha)(y_i - X_i\beta) \\
&= d_0^2 I_{n_i}' \left(\sigma^2 1_{n_i} 1_{n_i}' + \sigma^2 I_{n_i}\right)^{-1} (y_i - X_i\beta) \\
&= \frac{d_0^2}{\sigma^2} 1_{n_i}' \left(I_{n_i} - \frac{d_0^2}{\sigma^2 + n_i d_0^2} 1_{n_i} 1_{n_i}'\right) (y_i - X_i\beta) \\
&= \frac{n_i d_0^2}{\sigma^2 + n_i d_0^2} \frac{1}{n_i} \sum_{j=1}^{n_i} (y_{ij} - X_i^{[j]}\beta)
\end{aligned} \tag{3.62}$$

where $\hat{b}_{0i}$ is a weighted average of 0 (the 'prior' mean) and the average residual for i. Thus the larger n_i is, the less the shrinkage effect. Likewise, the smaller σ^2 is, relative to d_0^2, the lesser the shrinkage.

3.4.6 A Comment on the Normality Assumption for Random Effects

In practise, histograms for empirical Bayes (EB) estimates are often plotted to check the normality assumption for the random effects. But since

$$\begin{aligned}
\hat{b}_i &= DZ_i' W_i(y_i - X_i\hat{\beta}) \\
var(\hat{b}_i) &= DZ_i' \left\{ W_i - W_i X_i \left(\sum_{i=1}^{N} X_i' W_i X_i \right)^{-1} X_i' W_i \right\} Z_i D
\end{aligned} \tag{3.63}$$

one should at least standardise the EB estimates. Further, due to the shrinkage effect, the EB estimates do not fully reflect the heterogeneity in the data. Thus EB estimates obtained under the normality assumption cannot again be used to check the same normality assumption. This suggests that the best strategy to check the normality assumption would be to fit a more general model, with the classical generalised linear mixed model as a special case and then compare the two using the LR test. One other possible way to address the distributional assumption of the random effects would be to assume a finite mixture model for the random effects of the form

$$b_i \sim \sum_{k=1}^{d} p_k N(\mu_k, D) \quad with \quad \sum_{k=1}^{d} p_k = 1 \quad and \quad \sum_{k=1}^{d} p_k \mu_k = 0 \tag{3.64}$$

The implication of such an assumption is that the population consists of d sub-populations. Each sub-population contains a fraction p_k of the total population. In each sub-population a linear mixed model holds. The classical model is then just a special case with $d = 1$. The finite mixture model is popularly commonly fitted using the EM algorithm. A SAS macro is also available to run such a model.

3.4.7 Power for Fixed Effects Under the Linear Mixed Model

In this section, the question of power of the F-test for fixed effects will be looked into. Consider testing the hypothesis

$$H_0 : L\beta = \beta_0 = 0 \qquad vs \qquad H_A : L\beta \neq 0 \tag{3.65}$$

for some matrix constant L such that rank(L)>0. Then the F-statistic for the above test is given by

$$F_\beta = \frac{\hat{\beta}' L' (L(\sum_{i=1}^{N} X_i' V_i^{-1} X_i)^{-1} L')^{-1} L\hat{\beta}}{rank(L)} \tag{3.66}$$

as stated earlier. Under H_0, F_β is distributed as F with numerator degrees of freedom equal to rank(L) and denominator degrees of freedom will be estimated by one of three methods, namely the (1) Containment method (2) Satterthwaite approximation, or (3) Kenward-Roger approximation, or any other available method. In general when H_0 is not true, F_β is approximately F with the same numerator and denominator degrees of freedom, but now with a non-centrality parameter

$$\delta = \beta' L' \left[L \left(\sum_{i=1}^{N} X_i' V_i^{-1} X_i \right) L' \right]^{-1} L\beta \tag{3.67}$$

which is equal to 0 under H_0. Here δ can be used to calculate power for the test under a variety of models and under a variety of alternative hypotheses. Note that δ = rank(L)$\times F_\beta$ and with $\hat{\beta}$ replaced by its expected value β. SAS procedure `MIXED` can be used to calculate δ and the related number of degrees of freedom. Under H_A, where δ>0, the power is calculated as

$$P(\delta) = P(F_{n,d,\delta} > F_{calc}) \tag{3.68}$$

where F_{calc} is the calculated F-value such that

$$P(F_{n,d,0} > F_{calc}) = \alpha (= \text{level of significance}) \tag{3.69}$$

where n and d denote the numerator and denominator degrees of freedom, respectively. The `PROC MIXED` options `finv` and `probf` are used to calculate F_{calc} and the power respectively. Finally, it

should be noted that in longitudinal data settings, within-subject correlation increases power for inferences on within-subject effects but decreases power for inferences on between-subject effects.

3.4.8 Discussion

This section in the chapter is concluded by discussing general guidelines for model building. Recall that the probability distribution of the general linear mixed model for Y_i can be written as

$$Y_i \sim N(X_i\beta, Z_iDZ_i' + \sigma^2 I_{n_i}) \tag{3.70}$$

Thus fitting a linear mixed model requires the specification of a mean structure, as well as the covariance structure. The mean structure may in fact include the effects of measured or observed covariates, time effects and possibly their interactions. The covariance model includes random effects and possible serial correlation. Both components (mean and covariance models) affect each other through the estimation of $\theta = (\alpha', \beta')$, the covariance matrix for $\hat{\theta}$, the construction of t- and F- tests, confidence intervals, efficiency and prediction. It should be noted that recursive or the stage fitting of the model makes dealing with high dimensional parameters easier.

When most variability is due to between-subject effects, the two stage model will often lead to an acceptable marginal model. In the presence of a lot of within-subject variability, the two-stage approach is less straight forward. Also, a two stage approach may imply an unrealistic marginal model. Thus the general fitting strategy is to work with a preliminary mean structure $X_i\beta$, preliminary random effects structure Z_ib_i and residual covariance structure Σ_i. Then look for a more parsimonious model by first attempting to reduce the random effects in Z_ib_i, then reduce the mean structure $X_i\beta$.

3.5 Statistical Software

3.5.1 Fitting Linear Mixed Models Using Statistical Software

A number of statistical software available now have capability to fit linear mixed models with ease. These include SAS, GenStat, S-Plus, SPSS and many more. In this thesis, the SAS software is used to fit the various models under consideration. In SAS software, for estimation of fixed effects and variance components, `PROC MIXED` is used to primarily specify the data set and method of estimation (REML or ML) with REML being the default method. The `CLASS` statement is used to declare categorical or factor variables in the data. The `MODEL` statement is used to state the model

relating the response to the fixed effects variables. This statement also has an option of whether to call for solutions and whether to fit a model with an intercept or not. The `RANDOM` statement is used to define the random effects in the model. The statement has options to specify which variable identifies the subjects, assuming independence across subjects, type of random effects matrix D, options `g` and `gcorr` to print the matrix D and the corresponding correlation matrix, options `v` and `vcorr` to print the matrix V_i and the corresponding correlation matrix. The `REPEATED` statement is used to first identify the factor variable used for ordering the repeated measurements within a subject e.g. time, age, birth order in a family, and so on. There is also an option with the `REPEATED` statement to specify which variable identifies the individual or subject, the type of residual covariance matrix Σ_i, options `r` and `rcorr` to print Σ_i and the corresponding correlation matrix. Frequently used covariance structures available to the `RANDOM` and `REPEATED` statement are the unstructured (`UN`), variance components or simple independence (`VC` or `SIMPLE`), compound symmetry (`CS`), first-order autoregressive (`AR(1)`), several spatial structures (`SP`), and so on. A more exhaustive list of possible covariance structures can be found in the books by Verbeke and Molenberghs (2000), Diggle et al. (2002), Molenberghs and Verbeke (2005) among others.

Chapter 4

Joint Modelling

4.1 Introduction

Longitudinal studies are often analysed using mixed models, taking into account repeated measurements for each subject in the model. Since these subjects are followed up for a substantial amount of time, some of them will drop out of the study or they will experience some event which is related to the outcome or measurement of interest. This thesis aims to model the evolution of HIV markers, CD4+ count and viral load, within a group of newly HIV infected individuals. However, since some of these individuals initiate antiretroviral (ARV) treatment because their CD4+ count falls below 200 cells/$\mu\ell$ and they are progressing towards acquired immunodeficiency syndrome (AIDS) they cease being in the CAPRISA 002: Acute Infection Study and are no longer in follow-up. CD4+ count is no longer measured after they drop out of the study and a lot of information is lost as these individuals no longer contribute data in the initial cohort. Modelling the HIV markers without taking this informative drop-out into account will lead to optimistic conclusions about CD4+ count within this population.

This chapter will look at the application of the method of joint modelling as proposed by Henderson, Diggle and Dobson (2000) in order to model the evolution of each of the HIV markers separately, while taking into account informative drop-out. First the theory of survival analysis will briefly be discussed. Survival analysis is the method which will be used to model the informative drop-out, and then the theory behind joint modelling will be discussed. Bivariate joint modelling is also possible but currently beyond the scope of this thesis. This advanced type of analysis will be left as a future extension on the current project.

4.2 Survival Analysis

Survival analysis allows one to study the occurrence of and time to events, such as deaths or onset of disease. It has many applications in various different areas in the natural and social sciences, and is designed to be used on longitudinal data with the occurrence of a particular event of interest.

Survival analysis can be done prospectively, as well as retrospectively. For a prospective analysis, a cohort of individuals or subjects are observed from a well-defined point in time and followed up for a substantial amount of time, while recording the times at which the events occur. Risk factors of experiencing an event can also be accounted for by recording several covariates at baseline or even time-varying covariates during follow-up. However, not everyone who is being followed has to experience the event and those who do not experience the event are regarded as censored. Subjects can be censored without experiencing the event.

A retrospective survival analysis can be done by asking individuals to recall important dates of events of interest, such as deaths, marriages, disease, etc. However, there is an element of recall bias with such type of information which can compromise the validity of the results. The individuals might not be able to accurately remember information and particularly in collecting information on time-varying covariates. In the current study, the survival analysis data is prospective where a cohort of newly HIV infected participants are followed over time. The aim of survival analysis in joint modelling lies in its ability to model the survival data including informative drop-out as events of interest in the cohort and then simultaneously combining the survival sub-model to the longitudinal sub-model. This will give rise to the so-called joint model for the two processes.

4.2.1 Parametric Survival Models

In survival analysis, the component which is of most interested is the survival function, S, which is defined as

$$S(t) = P(T>t) \tag{4.1}$$

where t is the observed time and T is a random variable denoting the time to event. The survival function is clearly non-increasing and it is usually assumed to approach zero as time increases without bound, thus $S(t) \to 0$ as $t \to \infty$

All standard approaches to survival analysis are probabilistic (or stochastic), thus the times at which the events occur are assumed to be realisations of a random process. The probability dis-

tribution of the events can be described by three functions, namely (i) the cumulative distribution function (c.d.f.), (ii) probability density function (p.d.f.), or the more commonly used (iii) hazard function.

The c.d.f. of the variable T is denoted by $F(t)$ and it is defined as the probability that T will be less or equal to a specific time t.

$$F(t) = P(T \leq t) = 1 - S(t) \tag{4.2}$$

The p.d.f. of T is the derivative (slope) of the c.d.f. and is related to $S(t)$ as follows

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt} \tag{4.3}$$

The hazard function is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{P(T \geq t)} \tag{4.4}$$

The aim of the hazard function is to quantify the instantaneous risk that the event will happen at time t. Because time is a continuous variable, the probability that the event will happen exactly at time t is highly unlikely. Thus instead, the probability that the event will happen in an interval between t and $t + \Delta t$, conditional on the subject having survived to time t, is specified. Since one wants to get as close to time t as possible, Δt is let to approach zero, reaching its limiting value.

The survival function $S(t)$, the p.d.f. $f(t)$ and the hazard function $h(t)$ can all be used to describe the probability distribution of the events in the survival analysis and the three functions are related as follows

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

Thus from (4.3) and (4.5),

$$h(t) = -\frac{d}{dt} \log S(t) \tag{4.6}$$

which leads to

$$S(t) = \exp\left\{-\int_0^t h(u)du\right\} \tag{4.7}$$

and so

$$f(t) = h(t) \exp\left\{-\int_0^t h(u)du\right\} \tag{4.8}$$

Furthermore, parametric survival models can be constructed by choosing a specific probability distribution for the survival function. The distribution chosen will lead to a specific form or structure for $S(t)$. Some of the commonly used distributions and the corresponding formulae for $S(t)$ are given in the table below where Φ is the cumulative distribution function of the standard normal distribution.

Distribution	$S(t)$
Exponential	$e^{-\lambda t}$
Weibull	$e^{-\lambda t^{\gamma}}$
Gompertz	$e^{\frac{\lambda}{\theta(1-e^{\theta t})}}$
Log-normal	$1-\Phi\left(\frac{\ln(t)-\mu}{\sigma}\right)$
Log-logistic	$\left(1+(\frac{t}{\alpha})^{\beta}\right)^{-1}$

Table 4.1: Common survival functions

The exponential distribution is the most simple, with an assumption that the hazard is constant over time. Thus

$$h(t) = \lambda \tag{4.9}$$

which can also be written as $\log h(t) = \mu$ where $-\infty<\mu<\infty$. Substituting (4.9) into (4.7) gives $S(t) = e^{-\lambda t}$ which gives the p.d.f. as $f(t) = \lambda e^{-\lambda t}$. When the hazard rate varies over exposure time, a Weibull distribution is more appropriate.

The survival model can be extended to allow for the effects of explanatory variables. If there are covariates $x_1, x_2, ..., x_k$, then the exponential model, for example, would be written as

$$\log h(t|x) = \mu + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \tag{4.10}$$

and a Weibull which has $\log h(t) = \mu + \alpha t$, will have a model with covariates written as

$$\log h(t|x) = \mu + \alpha t + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \tag{4.11}$$

These model parameters are estimated using maximum likelihood estimation, but not a standard likelihood because now there is the possibility of censored observations in the current data. It is assumed that there are n independent individuals in the sample ($i = 1, 2, ..., n$) and each individual i has three parts of information measured. These are t_i, δ_i and x_i, where t_i is the time to event or censoring, δ_i is the indicator variable specifying whether individual i actually experienced the event or whether i was censored and x_i is the vector of covariates. Usually $\delta_i = 1$ if t_i is uncensored

(subject i experienced the event) and $\delta_i = 0$ if t_i is censored (i.e. subject i dropped out of the study). Also $x_i = (1, x_{i1}, ..., x_{ik})$ denotes the vector of covariate values.

If it is assumed that all participants experienced an event then the likelihood function would be

$$L = \prod_{i=1}^{n} f_i(t_i) \tag{4.12}$$

but in practise, the likelihood aught to account for those individuals who are censored. If it is assumed that r individuals were not censored and $n - r$ were censored then

$$L = \prod_{i=1}^{r} f_i(t_i) \prod_{i=r+1}^{n} S_i(t_i) \tag{4.13}$$

Using the indicator variable δ_i which denotes a censored or uncensored observation for each subject, the likelihood function can be written as

$$L = \prod_{i=1}^{n} [f_i(t_i)]^{\delta_i} [S_i(t_i)]^{1-\delta_i} \tag{4.14}$$

Thus if $\delta_i = 1$ (i.e. an event), it is $f_i(t_i)$ which contributes to the likelihood formulation while if $\delta_i = 0$ (i.e. censored) then it is $S_i(t_i)$ which contributes to the likelihood.

Once a probability distribution for the survival times (and consequently hazard function) has been chosen, the necessary terms can be substituted in (4.14) and the derivative of the log-likelihood w.r.t. the parameters of interest in the model can be calculated. This derivative is then equated to zero and the resulting equations solved in order to estimate the survival model.

As an example, the case of the exponential distribution gives

$$f_i(t_i) = \lambda_i e^{-\lambda_i t_i} \text{ and } S_i(t_i) = e^{-\lambda_i t_i}$$

where $\lambda_i = e^{-\beta x_i}$ and β is the vector of coefficients for the covariates. Now

$$\begin{aligned} L &= \prod_{i=1}^{n} \left(\lambda_i e^{-\lambda_i t_i}\right)^{\delta_i} \left(e^{-\lambda_i t_i}\right)^{1-\delta_i} \\ &= \prod_{i=1}^{n} \lambda_i^{\delta_i} e^{-\lambda_i t_i} \end{aligned} \tag{4.15}$$

By taking the logarithm of the likelihood,

$$\begin{aligned}\log L &= \sum_{i=1}^{n} \delta_i \log \lambda_i - \sum_{i=1}^{n} \lambda_i t_i \\ &= \sum_{i=1}^{n} \delta_i \log(e^{-\beta x_i}) - \sum_{i=1}^{n} (e^{-\beta x_i}) t_i \\ &= -\beta \sum_{i=1}^{n} \delta_i x_i - \sum_{i=1}^{n} t_i e^{-\beta x_i} \end{aligned} \tag{4.16}$$

Equation (4.16) is then differentiated w.r.t. β, set equal to zero and then solved for β. This will give the following equation

$$\sum_{i=1}^{n} \delta_i x_i = \sum_{i=1}^{n} x_i t_i e^{-\beta x_i} \tag{4.17}$$

From here iterative methods are used to estimate β, such as the Newton-Rhapson algorithm, which is the default method for `PROC LIFEREG`.

4.2.2 Semi-parametric Survival Models

Cox (1972) proposed a regression method for survival analysis which does not require a specific distribution for the survival times, which is why it is called a semi-parametric model and this approach is therefore more robust than the parametric models because of less parametric restriction. The basic Cox model can be derived in terms of the hazard function of individual i at time t given by

$$h_i(t) = \lambda_0(t) \exp(\beta_1 x_{i1} + ... + \beta_k x_{ik}) \tag{4.18}$$

where $\lambda_0(t)$ is called the baseline hazard and $\beta_1 x_{i1} + ... + \beta_k x_{ik}$ is a linear function of k fixed covariates. The baseline hazard is the hazard function for an individual when the covariates have values of 0. If the logarithm of (4.18) is taken then

$$\log h_i(t) = \alpha(t) + \beta_1 x_{i1} + ... + \beta_k x_{ik} \tag{4.19}$$

where $\alpha(t) = \log \lambda_0(t)$. If $\alpha(t) = \alpha$ then this is the same as the exponential model as mentioned in the subsection above and specifying $\alpha(t) = \alpha \log t$ gives the Weibull model. The function $\alpha(t)$ can take any form. The hazard for any individual is a fixed proportion of the hazard for any other individual and therefore this is called a proportional hazards model. For example the ratio of the hazards for individuals i and j assuming the presence of k predictor variables is

$$\frac{h_i(t)}{h_j(t)} = \exp\{\beta_1(x_{i1} - x_{j1}) + ... + \beta_k(x_{ik} - x_{jk})\} \tag{4.20}$$

Note that the baseline hazard function cancels out and the ratio of the two hazards becomes constant over time and is only dependent on the measured covariates. The proportional hazard property implies that the hazard functions are parallel.

Cox (1972) also proposed a new estimation method, the partial likelihood (PL) method of estimation. With this estimation method, the β coefficients can be estimated without having to specify the baseline hazard $\lambda_0(t)$. The partial likelihood method depends only on the order of the event times and not their exact values.

To explain the concept of partial likelihood estimation, suppose there are n independent individuals ($i = 1, 2, ..., n$) and for each individual i there is t_i, δ_i and $\mathbf{x}_i$, denoting the time to event, whether or not the individual was censored and a vector of k covariate variables, respectively. Usually a likelihood function is written as the product of all the individual likelihood functions. The partial likelihood can be written as a product of likelihoods for all the events experienced. If R is the number of events,

$$PL = \prod_{r=1}^{R} L_r \tag{4.21}$$

where L_r is the likelihood for the r^{th} event. The likelihood is calculated for each event by taking the hazard function for that event at time t and dividing it by the summation of the remaining at risk individual hazard functions if they were to have happened at time t without loss of generality assume the case of one covariate. The general equation for the partial likelihood for data with one fixed covariate is

$$PL = \prod_{i=1}^{n} \left(\frac{e^{\beta x_i}}{\sum_{j=i}^{n} Y_{ij} e^{\beta x_j}} \right)^{\delta_i} \tag{4.22}$$

where Y_{ij}=1 if $t_j \geq t_i$ and Y_{ij}=0 if t_j<t_i. Note that because of the proportional hazards assumption, the baseline hazard cancels out automatically. The partial likelihood can be maximised for β by taking the logarithm of (4.22), namely

$$\log PL = \sum_{i=1}^{n} \delta_i \left(\beta x_i - \log \left(\sum_{j=1}^{n} Y_{ij} e^{\beta x_j} \right) \right) \tag{4.23}$$

and differentiating with respect to β, then equated to zero and solved for β. Again, this would be solved iteratively for β, using a method such as the Newton-Rhapson algorithm. The above objective function in equation 4.23 can easily be extended to more than one covariate or predictor variable.

For purposes of this thesis, the parametric exponential survival model was used to model time

to ARV initiation in order to take into account informative drop-out.

4.3 Joint modelling

This section will describe how to model time to event data jointly with longitudinal data. Let $Y_i' = (Y_{i1}, Y_{i2}, ..., Y_{in_i})$ denote the vector of repeated measurements for individual i which have been measured at times $t_i' = (t_{i1}, t_{i2}, ..., t_{tn_i})$. Also, let the observed time in the study for the i^{th} individual be T_i^O =min(T_i, C_i) where T_i denotes the time the individual experienced the event of interest and C_i denote the time at which the individual was censored. Let δ_i be the censoring indicator which is equal to 1 if the individual experienced the event, and equal to 0 if the individual was censored.

Thus each individual in the cohort contributes the information $(T_i^O, Y_i, t_i, X_{1i}, x_{2i})$, where X_{1i} is the matrix of observed covariates in the longitudinal model and x_{2i} is the vector of observed covariates in the survival model.

The joint model by Henderson, Diggle and Dobson (2000) on how to join a longitudinal model to a time-to-event sub-model will be revisited. The main concept in this modelling strategy is to assume a latent bivariate Gaussian process $W_i(t) = \{W_{1i}(t), W_{2i}(t)\}$ and then assume that the longitudinal measurements and event processes are conditionally independent given $W_i(t)$ and any additional measured covariates. So an association between the longitudinal and survival models is described through the cross-correlation between $W_{1i}(t)$ and $W_{2i}(t)$. The direct link between $W_{1i}(t)$ and $W_{2i}(t)$ is called the latent association. Note that the longitudinal measurements depends on $W_{1i}(t)$, while the time-to-event process depends on $W_{2i}(t)$. For example, the longitudinal data would take on the form

$$Y_i = \mu_i(t_i) + W_{1i}(t) + \epsilon_i \tag{4.24}$$

The $\mu_i(t_i)$ component in the model above is the mean response that can be described by a linear model, for example $\mu_i(t_i) = X_{1i}\beta$, which represents both baseline and time-varying explanatory variables and their regression coefficients. The ϵ_i is a vector of the mutually independent measurement errors in the model, which is normally distributed with mean zero and variance $\sigma_\epsilon^2 I$. A basic example for the latent process $W_{1i}(t)$ is

$$W_{1i}(t) = U_{1i} + U_{2i}t \tag{4.25}$$

where (U_{1i}, U_{2i}) is a bivariate normal random vector with mean zero, and variance-covariance structure

$$G = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \tag{4.26}$$

$W_{1i}(t)$ as stated in (4.25) may include the random effects for intercept and slope over time and this allows different individuals to have different baseline measurements and different time trends or slopes for these measurements. Note that $\mu_i(t_i)$ corresponds to $X_i\beta$ in (3.5) and $W_{1i}(t)$ corresponds to $Z_i b_i$ in (3.5), where b_i is equivalent to (U_{1i}, U_{2i}) in the current setup.

The association between the measurement and time to event processes is achieved through the random term $W_{2i}(t)$. If this survival model is a Cox model then the hazard function for the i^{th} individual would be

$$\lambda_i(t) = \lambda_0(t) \exp\left(\mathbf{x}_{2i}\beta_S + W_{2i}(t)\right) \tag{4.27}$$

If the survival model is parametric, say an exponential distribution, then the survival model with the random component $W_{2i}(t)$ would take the form

$$\lambda_i(t) = \exp\left(\mathbf{x}_{2i}\beta_S + W_{2i}(t)\right) \tag{4.28}$$

It is assumed that (4.27) and (4.28) are conditionally independent to (4.24), given $W_i(t)$. In order to model an association between the longitudinal and time-to-event sub-models, $W_{2i}(t)$ is taken to be related to particular components of $W_{1i}(t)$. A general equation for $W_{2i}(t)$, assuming proportionality, is

$$W_{2i}(t) = \gamma W_{1i}(t) \tag{4.29}$$

while yet another equation for $W_{2i}(t)$ would allow the random slope and intercept to have different effects on the event process leading to a more general relation

$$W_{2i}(t) = \gamma_1 U_{1i} + \gamma_2 U_{2i} + \gamma_3 W_{1i}(t) \tag{4.30}$$

Thus this would have both the random intercept U_{1i}, slope U_{2i} and current value of $W_{1i}(t)$ in (4.25) affect the risk of an individual experiencing an event, as modelled in the time-to-event sub-model. The parameters of the longitudinal models and the survival model are estimated jointly by maximising the observed joint likelihood of the data.

As mentioned before, the longitudinal and the time-to-event (or survival) processes are condi-

tionally independent, given W_{1i}, W_{2i} and the measured covariates X. Without this conditioning, dependence between the longitudinal and time-to-event processes comes from the deterministic effects of common covariates in the two models, or it comes through the stochastic dependence between W_{1i} and W_{2i} (or the latent association). If there is neither common covariates between the longitudinal and survival sub-models, nor a latent association, then there is no use in joint modelling as no additional precision is gained from this method.

There are various relationships and different combinations that can be modelled between W_{1i} and W_{2i}. These are

$$W_{1i}(t) = U_{1i} \quad \text{and} \quad W_{2i}(t) = 0 \tag{4.31}$$

$$W_{1i}(t) = U_{1i} + U_{2i}t \quad \text{and} \quad W_{2i}(t) = 0 \tag{4.32}$$

$$W_{1i}(t) = U_{1i} \quad \text{and} \quad W_{2i}(t) = \gamma W_{1i}(t) \tag{4.33}$$

$$W_{1i}(t) = U_{1i} + U_{2i}t \quad \text{and} \quad W_{2i}(t) = \gamma_1 U1i + \gamma_2 U2i + \gamma_3 W_{1i}(t) \tag{4.34}$$

Equation (4.31) and (4.32) would assume independence between the longitudinal process and the time-to-event process, whereas (4.33) and (4.34) allows for dependence between these two processes leading to the joint model, combining the two processes.

Through this joint modelling it is assumed that the time to event, T_i^o, is correlated to the random effects b_i through the vector of covariances (Thiébaut, 2005), $B_{q\times 1}$, where

$$\begin{pmatrix} b_i \\ T_i^o \end{pmatrix} \sim N\left\{ \begin{pmatrix} 0 \\ \mu_{T^o} \end{pmatrix}, \begin{pmatrix} D & B \\ B' & \sigma^2_{T^o} \end{pmatrix} \right\} \tag{4.35}$$

4.3.1 The likelihood function

The likelihood function of the joint model is obtained by taking the product of the marginal distribution of the longitudinal data sub-model and the conditional time-to-event sub-model, given the observed values of Y.

Let θ denote the combined vector of unknown parameters, ω_{2i} denotes the complete path of W_{2i} for subject i and ω_2 the collection of these paths over all subjects. Also, let N denote the time-to-event

process. The likelihood of the joint model can be written as

$$\begin{aligned} L(\theta, Y, N) &= L_Y \times L_{N|Y} \\ &= L_Y(\theta, Y) \times E_{\omega_2|Y}\left\{L_{N|W_2}(\theta, N|\omega_2)\right\} \end{aligned} \tag{4.36}$$

The likelihood $L_Y(\theta, Y)$ is the likelihood function of the marginal multivariate normal distribution of longitudinal data Y. The conditional likelihood of the time-to-event process $L_{N|W_2}(\theta, N|\omega_2)$ contributes to the likelihood the numbers of longitudinal measurements before the individual drops out of the study. The likelihood $L_{N|W_2}(\theta, N|\omega_2)$ in (4.36) can be written as

$$L_{N|W_2}(\theta, N|\omega_2) = \prod_i \left\{ \left(\prod_t \left[\exp\left\{x_{2i}(t)\beta_2 + W_{2i}(t)\right\} \lambda_0(t)\right]^{\delta_i} \right) \times \exp\left(- \int_0^\tau \lambda_0(u) \exp\left\{x_{2i}(t)'\beta_2 + W_{2i}(t)\right\} du \right) \right\} \tag{4.37}$$

The current problem can be related to a similar approach by Thiébaut et al (2005) who writes their likelihood of the joint model, combining the longitudinal data and time-to-event process, as

$$L(\theta) = \prod_{i=1}^{N} \left(\int_{R^q} f_{Y_i|b_i}(Y_i|b_i = u) \left\{ f_{T_i^o|b_i}(T_i|b_i = u) \right\}^{\delta_i} \left\{ 1 - F_{T_i^o|b_i}(T_i|b_i = u) \right\}^{1-\delta_i} f_{b_i}(u) du \right) \tag{4.38}$$

where

$$f_{Y_i|b_i}(Y_i|b_i) = \left\{ \prod_{j=1}^{n_i} f_{Y_{ij}|b_i}(Y_{ij}|b_i) \right\} \tag{4.39}$$

In (4.38), $f_{T_i^o|b_i}$ and $F_{T_i^o|b_i}$ are the conditional probability density function and cumulative distribution function of T_i^o given b_i, respectively, and

$$T_i^o|b_i \sim MVN\left(\mu_{T_o} + \mathbf{B}'\mathbf{D}_i^{-1}b_i, \sigma_{T_o}^2 - \mathbf{B}'\mathbf{D}_i^{-1}\mathbf{B}\right)$$

Thus in equation (4.38), if there was no informative drop-out, that is $\delta_i = 0$ for all i, then the likelihood would only depend on the contribution of the response variables Y if $F_{T_i^o|b_i}(T_i|b_i) = 0$, as drop-out is impossible.

4.3.2 Left-censoring Of Viral Load

Viral load is one of the most common biomarker or measurement used as the primary endpoint in HAART studies as it gives an indication of viral suppression within an individual. Viral suppression can also occur in those recently infected with HIV as their bodies fight off the infection and is seen in those who are able to control the virus. Assays used to quantify viral load copies in the blood often have a lower detection limit which renders viral load measurements that fall

below this lower threshold undetectable. In the CAPRISA 002: Acute Infection Study cohort, the lower detection limit is 400 copies/mℓ and therefore all viral load measurements below this limit are unobservable and therefore considered as left-censored.

This is a relatively complex problem when modelling viral load data from patients on HAART as viral suppression is the aim of the treatment. In the current data, approximately 5% of all viral load measurements are undetectable. Part of the current work is to model the effect of left-censoring in the analysis and assess whether the results will benefit from incorporating methods capable of handling this problem such as that proposed by Jacqmin-Gadda et al. (2000).

There are two frequently used methods when dealing with the problem of the lower detection limit for viral load. In HAART studies the outcome is frequently modelled as a binary value so that subjects are classified as having a suppressed viral load or not. The proportion of subjects with a viral load below the detection limit is calculated. However, this method loses information, such as slopes and magnitude of peaking viral load and this method cannot be used in the current data as majority of viral load measures are in fact detectable. In addition, the aim of the current study is more to understand the natural evolution of the HIV markers, CD4+ count and viral load, over time during the acute infection stage of HIV. Another method of dealing with the issue of undetectable viral load measurements involves imputing half the lower detection limit of the assay (O'Brien et al., 1998) for those viral load measurements which are undetectable but this approach may result in biased estimates.

In this thesis, left-censoring of viral load will be handled using a likelihood approach to estimate undetectable viral load data. This approach involves replacing the conditional probability density function presented in (4.39) by

$$f_{Y_i|b_i}(Y_i|b_i) = f_{Y_i^o|b_i}(Y_i^o|b_i)P(Y_i^c<s_i|b_i) = \left\{\prod_{j=1}^{n_{io}} f_{Y_{ij}^o|b_i}(Y_{ij}^o|b_i)\right\}\left\{\prod_{j=1}^{n_{ic}} F_{Y_{ij}^c|b_i}(s_{ij}|b_i)\right\} \tag{4.40}$$

where Y_i^o represents the n_i^o-dimensional vector of observed measurements, Y_i^c is the n_i^c-dimensional vector of left-censored measurements and s_i is the n_i^c-dimensional vector of measurement thresholds.

Thus each individual contributes to the likelihood the product of the density of the observed measurements and of the conditional distribution function of the censored measurements given the

observed measures and random effects. Since normal distribution is assumed,

$$P(Y_i^c < s_i | b_i) = \Phi(s_i | b_i) \tag{4.41}$$

where Φ is the c.d.f of the standard normal distribution.

Chapter 5

Application

5.1 Introduction

The software of choice used to do the analysis and fit the various models was SAS version 9.1.3 (SAS Institute Inc., Cary, NC, USA). Specific procedures used were `PROC MIXED`, `PROC NLMIXED`, and `PROC LIFEREG`.

`PROC MIXED` is a procedure which generalises standard linear models, which would be fitted using `PROC GLM` or `PROC REG`, and fits the wider class of mixed linear models. `PROC MIXED` allows one to incorporate both fixed and random effects in modelling repeated measures problems. The `RANDOM` statement in `PROC MIXED` is used to add random effects to the model while the `REPEATED` statement is used to specify repeated measurements. `PROC MIXED` also allows a wider variety of covariance structures and carries out several analyses, including the estimation and testing of linear combinations of fixed and random effects. `PROC MIXED` assumes that the outcome variable is normally distributed and either maximum likelihood (ML) or restricted maximum likelihood (REML) estimation can be used. `PROC MIXED` will be used to fit the linear mixed models to viral load and CD4+ count. For this thesis, REML estimation will be applied and empirically correct estimators will be used for the fixed effects by using the option `EMPIRICAL` in the `PROC MIXED` statement. This method gives a consistent estimator of precision, even if the covariance structure is incorrectly specified (Verbeke and Molenberghs, 2000).

`PROC NLMIXED` can be viewed as a generalisation of the random effects models fit by the `PROC MIXED`. This procedure lets the random effects to enter the model non-linearly, in contrast to `PROC MIXED` where random effects are linear. `PROC NLMIXED` only allows ML estimation because the procedure

involves high dimensional integrals over all of the fixed effects parameters non-linearly and this integral is not available in a closed form. `PROC NLMIXED` allows one to fit not only data that is normally distributed, but also binomial, Poisson or any distribution for which a likelihood function can be user-programmed. This procedure will be used to address the problem of left-censoring of viral load, as well as the univariate joint modelling.

`PROC LIFEREG` is a parametric regression procedure for modelling the distribution of survival time. This procedure will be used to model the survival or time-to-event data, where the events are the informative drop-outs.

Because of the nature of HIV during acute infection, with the viral load reaching a peak and then decreasing, while the CD4+ count drops quickly and then recovers, the time component was divided up into intervals and a piecewise model was fitted. These time components were modelled in order to describe and quantify the evolution of the markers during acute HIV infection. The piecewise time intervals were chosen in consideration to the results from the exploratory analysis, namely the scatter plot and Loess smoothing line in Figures 3.1 and 3.2. From the models presented the slopes for CD4+ count and viral load within these intervals are calculated and the evolutionary change in these HIV markers are described. The intervals are 0 to 2, 2 to 4, 4 to 8, 8 to 12, 12+ weeks post infection. Initially the interval 0 to 4 was used, but this seemed to underestimate the viral load peak that was evident immediately after infection and it did not entirely agree with the results that were seen in the Loess smoothing line. Viral load was log-transformed, while CD4+ count had a square-root transformation in order to ensure the normality assumption in the modelling process.

5.2 Linear Mixed Models

5.2.1 CD4+ count

5.2.1.1 Marginal Model

A marginal model was fitted to CD4+ count with weeks post infection as a continuous predictor, after a square root transformation was applied to the CD4+ count measurements to ensure normality. The initial model had the following form:

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \varepsilon_{ij} \tag{5.1}$$

where y_{ij} is the j^{th} square root CD4+ count measurement for subject i ($j = 1, 2, ..., n_i$), β_0 is the intercept, t_{ij} represents the weeks post infection at the i^{th} measurement, while β_1 is the slope estimate for the change in square root CD4+ count for every one week increase. The ε_{ij} is the random error associated with the j^{th} measurement for subject i. The model was fitted using SAS `PROC MIXED` with the variable `week` in the `MODEL` statement, thus declaring it a fixed effect. No random effects were defined. As with the viral load model, various covariance structures were used to model the within-subject variance. The unstructured covariance (UN) would not allow the model to converge as there were too many repeated observations per individual. The autoregressive AR(1) structure was not suitable for the type of data as the measurements were not equally spaced. Covariance structures used was the compound symmetry (CS), the exponential spatial (SP(EXP)), power spatial (SP(POW)) and the Gaussian spatial (SP(GAU)) structures. The best covariance structure was determined using the fit statistics output by the model in Table 5.1.

Covariance Structure	CS	SP(EXP)	SP(POW)	SP(GAU)
-2 Log Likelihood	6989.8	7090.5	7090.5	7940.2
AIC	6993.8	7094.5	7094.5	7944.2
BIC	6998.1	7098.7	7098.7	7948.5
Chi-Square	1115.15	1014.51	1014.51	164.75
Pr>Chi-Square	<.0001	<.0001	<.0001	<.0001

Table 5.1: Fit statistics and the Null Model Likelihood Ratio Test for fitting a marginal model to CD4+ count

It can be seen that the covariance structure that provided the best fit to the data was the CS structure since that was the model which provided that lowest AIC of 6989.8. Note that the SP(EXP) and SP(POW) structures produce the same fit statistics and if the AIC is the lowest, then either one of these structures can be used.

Solution for Fixed Effects

Effect	Estimate	Std Err	DF	tValue	Prob>|t|
Intercept	22.9690	0.4775	61	48.10	<.0001
week	-0.02361	0.004248	1313	-5.56	<.0001

Table 5.2: Fixed effects estimates modelling CD4+ count in a marginal model with weeks post infection as a continuous linear predictor

The effect estimates show that the model intercept is equal to 22.9690 square root CD4+ cells/$\mu\ell$ when weeks post infection is fitted linearly and equals 0. The effect estimate for weeks post infection shows a highly significant negative effect with -0.02361 square root CD4+ cells/$\mu\ell$ ($p<0.0001$).

This is the slope, or rate of change in square root CD4+ count per unit increase in weeks post infection. Hence for every week post infection, CD4+ count decreases by 0.02361 square root cells/$\mu\ell$, showing that over time CD4+ count decreases after HIV infection, which is what is expected. The covariance parameter estimates are displayed in Table 5.3.

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Std Err	Z Value	Prob>|Z|
CS	pid	12.8615	2.3969	5.37	<.0001
Residual		7.9629	0.3107	25.62	<.0001

Table 5.3: Covariance parameter estimates for modelling CD4+ count in a marginal model

Since the covariance structure used is the compound symmetry, which assumes constant covariance, the estimate 12.8615 is the covariance between any two measures on the same subject, i.e. Cov(y_{ij},y_{ik})=$\rho\sigma^2$=12.8615 where y_{ij} and y_{ik} are the j^{th} and k^{th} measurements for subject i, respectively. The estimate 7.9629 is the residual variance component, the variance of y_{ij} conditional on a participant, thus Var($y_{ij}|i$)=7.9629. The unconditional variance of y_{ij} is Var(y_{ij})=Cov(y_{ij},y_{ik}) + Var($y_{ij}|i$)= 12.8615 + 7.9629 = 20.8244. The estimated covariance matrix of Y for any subject i has dimensions $n_i \times n_i$ and looks like the following

$$\begin{pmatrix} 20.8244 & 12.8615 & 12.8615 & \cdots & 12.8615 \\ 12.8615 & 20.8244 & 12.8615 & \cdots & 12.8615 \\ 12.8615 & 12.8615 & 20.8244 & \cdots & 12.8615 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 12.8615 & 12.8615 & 12.5225 & \cdots & 20.8244 \end{pmatrix}$$

Note that within-subject variation, i.e. Cov(y_{ij},y_{ik}), is statistically significant from zero, indicating that there is significant variation between CD4+ count measurements within each participant as expected.

5.2.1.2 Random Intercept Model

By allowing for a random intercept in the model, it can be determined whether the intercept for CD4+ count is subject-specific. If there is a lot of variation in CD4+ count when weeks post infection is equal to zero, then it is important to have a random intercept in the model to account for this between-subject variation. The random intercept model can be written in the following form:

$$y_{ij} = (\beta_0 + b_{0i}) + \beta_1 t_{ij} + \varepsilon_{ij} \tag{5.2}$$

where the added term b_{0i} is the subject-specific effect for the intercept for subject i. By adding the term `intercept` to the `RANDOM` statement in `PROC MIXED`, the intercept will be modelled as random and the subject-specific variation will be accounted for. Given the model 5.2.1.2, it is important to note that there will only be one random effect in the model and therefore a covariance structure for random effects does not need to be specified. If a covariance structure is not specified then the default variance components (VC) or SIMPLE structure will be applied. However, when the VC structure is used in the `RANDOM` or `REPEATED` statements then the Null Model Likelihood Ratio Test is not performed. This is because the null hypothesis lies on the boundary of the parameter space and the standard asymptotic theory does not apply in this case. Because of this, the unstructured (UN) covariance structure will be fitted under the `TYPE=` option in the `RANDOM` statement to ensure SAS performs the Null Model Likelihood Ratio test. This has no effect on the model and would be no different than using the default VC structure for modelling the variance between random effects. Different covariance structures will be used to model the within-subject variance, for the repeated measurements, under the `REPEATED` statement. The best covariance structure is selected using the fit statistics for the separate models. As before the unstructured (UN) and autoregressive (AR(1)) covariance structures were not suitable to model the within-subject variance.

Covariance Structure	CS	SP(EXP)	SP(POW)	SP(GAU)
-2 Log Likelihood	6989.8	6816.6	6816.6	6934.1
AIC	6995.8	6822.6	6822.6	6940.1
BIC	7002.2	6829.0	6829.0	6946.4
Chi-Square	1115.15	1288.36	1288.36	1170.91
Pr>Chi-Square	<.0001	<.0001	<.0001	<.0001

Table 5.4: Fit statistics for fitting a random intercept model to CD4+ count

The fit statistics for the models with various covariance structures in Table 5.4 indicate that either the spatial exponential or power covariance structure are suitable for fitting the random intercept model as they both have the same and lowest AIC of 6816.6. The SP(POW) was used to model the within-subject variation.

Solution for Fixed Effects

Effect	Estimate	Std Err	DF	tValue	Prob>|t|
Intercept	22.9756	0.4879	61	47.09	<.0001
week	-0.02345	0.004102	1313	-5.72	<.0001

Table 5.5: Fixed effects estimates modelling CD4+ count in a random intercept model with weeks post infection as a continuous linear predictor

The effect estimates for the fixed effects change slightly, with a slight increase in the standard errors. Again the rate of change in CD4+ count over weeks post infection is significantly negative with CD4+ count decreasing by 0.02345 square root cells/$\mu\ell$ per week (p<0.0001). The covariance parameter estimates are displayed in Table 5.6.

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Std Err	Z Value	Prob>|Z|
UN(1,1)	pid	13.4961	2.5772	5.24	<.0001
SP(POW)	pid	0.7417	0.02176	34.09	<.0001
Residual		8.7100	0.4220	20.64	<.0001

Table 5.6: Covariance parameter estimates for modelling CD4+ count in a random intercept model

The covariance parameter estimate for UN(1,1) is the between-subject variance in the model intercept modelled as a random effect. This is statistically significant which indicates that there is indeed variation between subject intercepts. The resulting covariance matrix of y_{ij} is different for each subject as the SP(POW) structure models the covariance as a function of the Euclidean distances between observations and intervals between observations are unequal within and between subjects. To get the subject-specific estimates for intercept the `SOLUTION` option needs to be specified in the `RANDOM` statement. This will output a random effect estimate for each subject and a sample of the first 10 subjects are displayed in Table 5.7.

Solution for Random Effects

Effect	pid	Estimate	Std Err	DF	tValue	Prob>|t|
Intercept	1	-3.8137	0.8665	1313	-4.40	<.0001
Intercept	2	1.7191	0.9878	1313	1.74	0.0820
Intercept	3	-1.7809	0.8558	1313	-2.08	0.0376
Intercept	4	-1.4447	0.8625	1313	-1.68	0.0942
Intercept	5	10.7701	0.8071	1313	13.34	<.0001
Intercept	6	0.4151	0.7867	1313	0.53	0.5978
Intercept	7	-2.4596	1.2517	1313	-1.96	0.0496
Intercept	8	-5.2773	0.8463	1313	-6.24	<.0001
Intercept	9	-4.5015	1.0889	1313	-4.13	<.0001
Intercept	10	2.5007	0.8496	1313	2.94	0.0033

Table 5.7: Subject-specific effect estimates for the first 10 subjects, modelling CD4+ count in a random intercept model

The subject-specific estimates show how much each subject differs from the population-averaged estimate (from the Solution for Fixed Effects). Thus the intercept for pid (participant) 1 is $\beta_0 + b_{01}$

= 22.9756 - 3.8137 = 19.1619 square root CD4+ cells/$\mu\ell$, while the intercept for pid 2 is $\beta_0 + b_{02}$ = 22.9756 + 1.7191 = 24.6947 square root CD4+ cells/$\mu\ell$, and so forth. The t-test displayed in the table tests whether the subject-specific estimate is statistically significant from zero, i.e. whether it is significantly different from the population estimate of the intercept. In this small sample it can be seen that the intercept estimates of eight of the ten subjects are different from the population estimate, supporting the fact that there is significant variability from individual to individual intercept in the data.

5.2.1.3 Random Intercept and Slope Model

Since it has already been shown that there exists significant variation between subjects at baseline which corresponds to the intercept, the slope will now also be fit as a random effect in order to determine whether this level of variation is also necessary to take into account. By adding the variable `week` to the `RANDOM` statement, along with intercept, a random slope (and intercept) model will be fitted. Weeks post infection will also be under the `MODEL` statement as a fixed effect in order to get a population estimate while accounting for the subject-specific variation by fitting it as a random effect at the same time. Following the same strategy, different covariance structures are fitted to the repeated measures for the within-subject variance and the best model is chosen. Since there are two random effects in the model (intercept and slope), a covariance structure needs to be specified. To keep it simple and robust, the unstructured (UN) covariance structure was applied to estimating the covariance between random effects. The UN structure is also the most liberal and since there are only two random effects it would not be computationally intense to estimate the covariance parameters. The random intercept and slope model has the following form

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_{ij} + \varepsilon_{ij} \tag{5.3}$$

where now the added term b_{2i} is the subject-specific random effect to account for between-subject variability in the rate of change of CD4+ count over weeks post infection.

Covariance Structure	CS	SP(EXP)	SP(POW)	SP(GAU)
-2 Log Likelihood	6757.6	6701.5	6701.5	6733.9
AIC	6767.6	6711.5	6711.5	6743.9
BIC	6778.2	6722.2	6722.2	6754.5
Chi-Square	1347.40	1403.44	1403.44	1371.11
Pr>Chi-Square	<.0001	<.0001	<.0001	<.0001

Table 5.8: Fit statistics for fitting a random intercept and slope model to CD4+ count

Since both the spatial exponential and power covariance structures provide the best fit for the within-subject variance, with the lowest AIC of 6711.5, either one of these structures can be used to model the repeated measurements. For the random intercept and slope model, the SP(POW) covariance structure will be used. The fixed effect estimates change slightly when adding both

Solution for Fixed Effects

Effect	Estimate	Std Err	DF	tValue	Prob>|t|
Intercept	23.1207	0.4596	61	50.31	<.0001
week	-0.03044	0.004772	61	-6.38	<.0001

Table 5.9: Fixed effects estimates modelling CD4+ count in a random intercept and slope model with weeks post infection as a continuous linear predictor

intercept and slope as subject-specific effects, with the estimate for week decreasing to -0.03044 square root CD4+ cells/$\mu\ell$. In essence, the magnitude of the slope is bigger than when the model accounts for only baseline heterogeneity.

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Std Err	Z Value	Prob>|Z|
UN(1,1)	pid	12.1265	2.3958	5.06	<.0001
UN(2,1)	pid	0.001152	0.01673	0.07	0.9451
UN(2,2)	pid	0.001040	0.000305	3.41	0.0003
SP(POW)	pid	0.5822	0.03884	14.99	<.0001
Residual		6.4949	0.2981	21.79	<.0001

Table 5.10: Covariance parameter estimates for modelling CD4+ count in a random intercept and slope model

In Table 5.10, UN(1,1) refers to the between-subject variance of the model intercept, while UN(2,2) is the between-subject variance for the slope. It can be seen that the variance of the slope between subjects is significant (p=0.0003) indicating that a random slope should be fitted to the model. Note that these variance estimates make up the D matrix of covariance of random effects given by

$$D = \begin{pmatrix} 12.1265 & 0.001152 \\ 0.001152 & 0.001040 \end{pmatrix} \tag{5.4}$$

The estimates of the random effects, namely the intercept and slope, are obtained through the solution of the random effects in SAS.

Solution for Random Effects

Effect	pid	Estimate	Std Err	DF	tValue	Prob>|t|
Intercept	1	-2.6842	0.9816	1252	-2.73	0.0063
week	1	-0.01105	0.01083	1252	-1.02	0.3079
Intercept	2	2.7147	1.1266	1252	2.41	0.0161
week	2	-0.01704	0.01960	1252	-0.87	0.3848
Intercept	3	-0.7263	0.9632	1252	-0.75	0.4509
week	3	-0.01379	0.01154	1252	-1.20	0.2323
Intercept	4	-4.9332	0.9544	1252	-5.17	<.0001
week	4	0.05584	0.01165	1252	4.79	<.0001
Intercept	5	7.4587	0.9143	1252	8.16	<.0001
week	5	0.04215	0.008782	1252	4.80	<.0001

Table 5.11: Subject-specific effect estimates for the first 5 subjects, modelling CD4+ count in a random intercept and slope model

The subject-specific estimates for the random intercept and slopes for five subjects are shown in Table 5.11, and their interpretation is similar to the random intercept model (and fixed slopes) as described in Table 5.7. For example, the intercept for pid 1 is $\beta_0 + b_{01}$ = 23.1207 - 2.6842 = 20.4365 square root CD4+ cells/$\mu\ell$ and the slope for pid 1 is $\beta_1 + b_{11}$ = -0.03044 - 0.01105 = -0.04149 square root CD4+ cells/$\mu\ell$ per week. This gives the estimated regression model for subject 1 as $\sqrt{CD4+}$=20.4365 - 0.04149 t_{ij}. Note that the correlation between the subject-specific intercept and slope is not significant (p=0.9451). Figure 5.1 is a graphical representation of the fixed effect estimates for the marginal model, where weeks post infection is modelled as a linear effect with a random intercept and slope. Needless to say, this model does not offer a precise description of the CD4+ count data, especially when it is compared to what is seen in Figure 2.2.

5.2.2 Viral load

5.2.2.1 Marginal Model

A marginal model, with no random effects, was fitted to viral load with weeks post infection as a continuous linear predictor variable. Viral load was log-transformed to ensure that the normality assumption is met. The model had the following general form:

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \varepsilon_{ij} \tag{5.5}$$

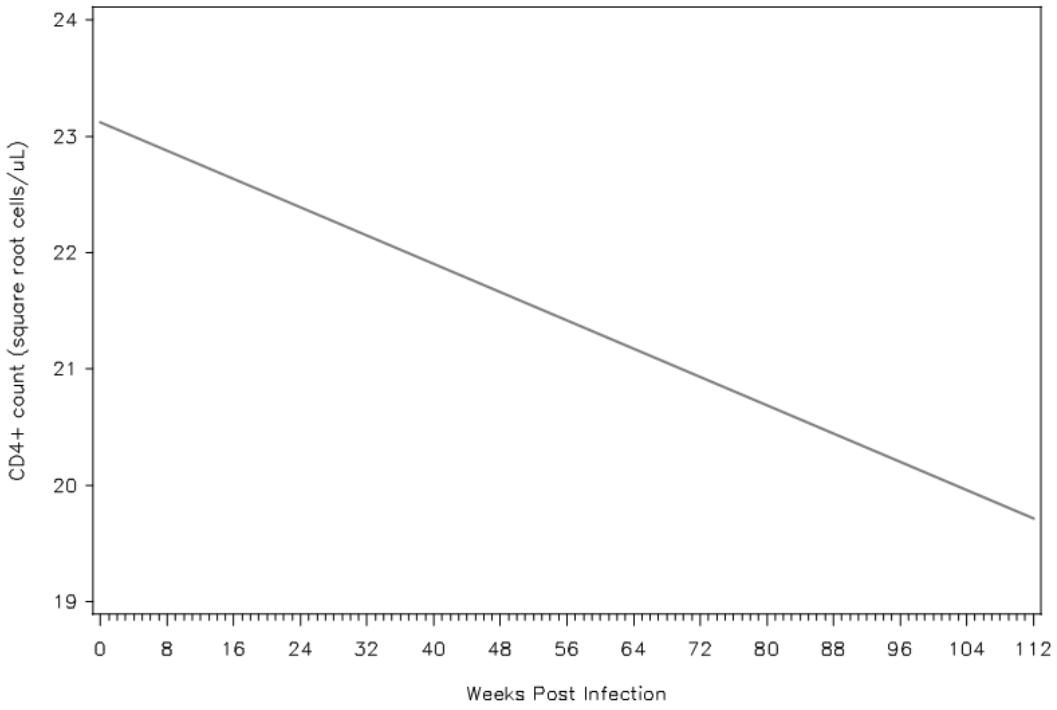

Figure 5.1: Predicted marginal model for CD4+ count with time as a linear effect

where y_{ij} is the j^{th} log viral load measurement for subject i ($j = 1, 2, ..., n_i$), β_0 is the intercept, t_{ij} represents the weeks post infection at the i^{th} measurement, while β_1 is the slope estimate for the change in log viral load for every one week increase. The ε_{ij} term is the random error associated with the j^{th} measurement for subject i. Different covariance structures were used to model the within-subject variation. The unstructured (UN) covariance structure was much too computationally intensive since there are so many repeated measurements for each individual. It is also not correct to use the autoregressive AR(1) covariance structure as this assumes that data are equally spaced and are taken at the same points in time for all individuals, which is not the case for this data. The compound symmetry (CS) and various spatial covariance structures were applied to the repeated measurements and the fit statistics used to determine which covariance structure was to be used (Table 5.12).

Since the CS structure had the lowest AIC, it was chosen as the covariance structure appropriate to model the repeated measurements. The model estimates are shown in Table 5.13. The results show that the intercept of viral load is 4.0759 log copies/mℓ while the slope over time 0.001631 log copies/mℓ per week. Both of these estimates are significantly different from zero with p-values $<$0.0001 and 0.0263, respectively. There is a definite increase in viral load over time after HIV

Covariance Structure	CS	SP(EXP)	SP(POW)	SP(GAU)
-2 Log Likelihood	4429.7	4955.3	4955.3	6519.0
AIC	4433.7	4959.3	4959.3	6523.0
BIC	4438.0	4963.6	4963.6	6527.3
Chi-Square	319.93	2477.84	2477.84	914.14
Pr>Chi-Square	<.0001	<.0001	<.0001	<.0001

Table 5.12: Fit statistics and the Null Model Likelihood Ratio Test for fitting a marginal model to viral load

Solution for Fixed Effects

Effect	Estimate	Std Err	DF	tValue	Prob>|t|
Intercept	4.0759	0.08916	61	45.71	<.0001
week	0.001631	0.000733	1380	2.22	0.0263

Table 5.13: Fixed effects estimates modelling viral load in a marginal model with weeks post infection as a continuous linear predictor

infection, which agrees with the opposite result seen in the model for CD4+ count which showed a significant decrease. However, the estimate for the intercept being 4.0759 log copies/mℓ is a bit confusing, as this is when weeks post infection is zero and it is assumed that log viral load is nil at zero weeks post infection. This happens because weeks post infection is fitted as a linear predictor so the actual value of log viral load at week 0 is not really captured. Because of this, a no-intercept model will be fitted by using the `NOINT` option in the `MODEL` statement. It is also to be noted that at this point the intercept of the viral load is not of interest as it is already assumed to be zero. Note that now the intercept falls away and the no-intercept model would have the following form:

$$y_{ij} = \beta_1 t_{ij} + \varepsilon_{ij} \tag{5.6}$$

The different covariance structures for modelling the within-subject correlation were applied and the CS structure provided the best fit. The model estimates, fitting a no-intercept model to log viral load, are provided in Table 5.14.

Solution for Fixed Effects

Effect	Estimate	Std Err	DF	tValue	Prob>|t|
week	0.001883	0.000736	1380	2.56	0.0106

Table 5.14: Fixed effects estimates modelling viral load in a no-intercept marginal model with weeks post infection as a continuous linear predictor

The slope of viral load over time is increasing at 0.001883 log copies/mℓ per week and this is statistically significant (p=0.0106). This result also agrees with the earlier CD4+ count model which showed that CD4+ count decreased over time after HIV infection. The covariance parameter estimates for the model are shown in Table 5.15. The estimate 16.9524 is the covariance between two measures on the same subject, while the estimate 1.1319 is the residual variance component, the variance of y_{ij} conditional on a participant. The unconditional variance of y_{ij} is Var(y_{ij})= 16.9524 + 1.1319 = 18.0843. Note that within-subject variation, i.e. Cov(y_{ij},y_{ik}), is statistically significant from zero, hence there is significant variation between viral load measurements within each participant.

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Std Err	Z Value	Prob>|Z|
CS	pid	16.9524	3.0622	5.54	<.0001
Residual		1.1319	0.04310	26.26	<.0001

Table 5.15: Covariance parameter estimates for modelling viral load in a marginal no-intercept model

5.2.2.2 Random Slope Model

Since a no-intercept model was used, there was no need to test for a random intercept. Thus a random slope model was fitted and it had the following form:

$$y_{ij} = (\beta_1 + b_{1i})t_{ij} + \varepsilon_{ij} \tag{5.7}$$

where the added term b_{1i} is the subject-specific slope estimate of log viral load for subject i. The variable `week` is added to the `RANDOM` statement. Note that as with the CD4+ count model, the UN covariance structure is specified even though there is only one random effect (the slope). This is because the default VC structure does not allow the Null Model Likelihood Ratio Test to be performed as described earlier, but fitting the UN structure does not impact on the results. The best covariance structure for the within-subject correlation is selected using the fit statistics for the separate models and the CS structure proved to have the better fit, even after adding a random slope (Table 5.16).

Covariance Structure	CS	SP(EXP)	SP(POW)	SP(GAU)
-2 Log Likelihood	4640.2	4955.3	4955.3	6452.6
AIC	4646.2	4959.3	4959.3	6458.6
BIC	4652.6	4963.6	4963.6	6465.0
Chi-Square	2792.98	2477.84	2477.84	980.59
Pr>Chi-Square	<.0001	<.0001	<.0001	<.0001

Table 5.16: Fit statistics for fitting a random slope no-intercept model to viral load

Solution for Fixed Effects

Effect	Estimate	Std Err	DF	tValue	Prob>|t|
week	0.002286	0.000765	61	2.99	0.0041

Table 5.17: Fixed effects estimates modelling viral load in a random slope no-intercept model with weeks post infection as a continuous linear predictor

Again, the estimates show that there is a significantly increasing log viral load after infection with a 0.002286 log copies/mℓ increase for every week post infection (p=0.0041). Note that the slope is now larger than when a random slope effect is not accounted for in the model.

The covariance estimate UN(1,1) = 0.000011 is the variance between subjects in the slope and clearly it is very small. There seems to be marginal significant variation between subjects regarding there rate of change of log viral load over time post infection (p=0.0414) when accounting for both within- and between-subject variation. Although this between-subject variation is small. Table 5.19 is a sample of the first ten subjects and their subject-specific slope estimates. Note that none of the subject-specific estimates differ significantly from the population (or fixed) estimate for the slope. Although the covariance estimate for the between-subject variation in the slope estimates is significant overall (p=0.0414), this is just a mild significant effect at the 5% significance level.

When looking at the solution to the random effects in Table 5.19 it can be seen that the subject-specific slope estimates for the first 10 subjects are not statistically significant from zero. This may suggest that viral load as an HIV marker may be a more reliable predictor of the disease as claimed by other researchers (Huang et al., 2006).

Figure 5.2 is a graphical representation of the marginal model shown in Table 5.17, and as with Figure 5.1, it can be seen that Figure 5.2 does not fit the data properly. The no-intercept model

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Std Err	Z Value	Prob>|Z|
UN(1,1)	pid	0.000011	6.316E-6	1.73	0.0414
CS	pid	16.8713	3.0473	5.54	<.0001
Residual		1.1078	0.04298	25.78	<.0001

Table 5.18: Covariance parameter estimates for modelling viral load in a random slope no-intercept model

Solution for Random Effects

Effect	pid	Estimate	Std Err	DF	tValue	Prob>|t|
week	1	-0.00158	0.002529	1381	-0.62	0.5326
week	2	-0.00135	0.003097	1381	-0.44	0.6631
week	3	0.001571	0.002629	1381	0.60	0.5501
week	4	-0.00241	0.002650	1381	-0.91	0.3640
week	5	-0.00268	0.002217	1381	-1.21	0.2272
week	6	-0.00214	0.002071	1381	-1.04	0.3008
week	7	0.000498	0.003278	1381	0.15	0.8794
week	8	-0.00247	0.002361	1381	-1.05	0.2950
week	9	0.000787	0.003223	1381	0.24	0.8070
week	10	-0.00157	0.002225	1381	-0.71	0.4805

Table 5.19: Subject-specific effect estimates for the first 10 subjects, modelling viral load in a random slope no-intercept model

forces the regression line through zero viral load copies/mℓ when week is zero, since it is assumed that subjects are HIV-negative at that point.

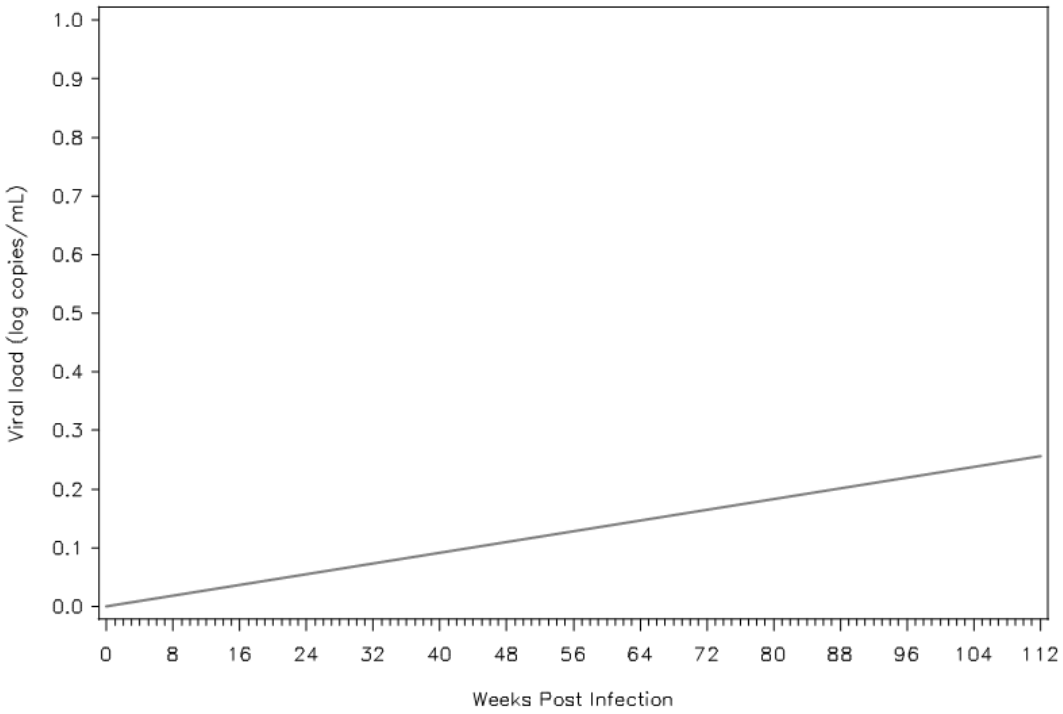

Figure 5.2: Predicted marginal model for viral load with time as a linear effect

5.2.3 Discussion

Fitting time as a linear predictor is very useful as it indicates whether weeks post infection has a significant effect on CD4+ count or viral load. For CD4+ count there was a significant within-subject variation, which agrees with the results seen in the line plot in Figure 2.4. The results from fitting the random intercept and slope model to CD4+ count showed that there is significant variation between subjects for both the intercept and the slope over time. The population estimate for weeks post infection shows a significant decreasing trend in CD4+ count after HIV infection with a loss of 0.03044 square root CD4+ cells/$\mu\ell$ per week on average ($p<0.0001$). When modelling viral load, a no-intercept model was fitted. A significant within-subject variation was found and subsequently it was shown that there was also significant between-subject variation in the slope over time. However, looking at the subject-specific slope estimates it was clear that most of these were not statistically significant from zero even though overall there was significant variation in the slopes. The population estimate for weeks post infection shows a significant increasing viral load after infection of 0.002286 log copies/mℓ increase per week ($p=0.0041$). Both models agree with the literature in that after infection viral load increases, CD4+ count declines until the person is diagnosed with AIDS and thereafter start receiving antiretroviral therapy or dies due to opportunistic infections. However, as useful as these linear models may be, they do not depict

much about the evolution of the HIV markers or the disease since time is modelled as a linear predictor. Thus a closer look at the trajectory of both markers is necessary. Thus in Section 5.3 the analysis is enhanced by applying piecewise modelling techniques in order to better understand the natural disease process for individuals infected with HIV-1 Subtype C.

5.3 Piecewise Linear Mixed Models

In the following section, piecewise linear time components will be fitted to both CD4+ count and viral load in order to explore the natural evolution of HIV infection via the HIV markers. In using piecewise effects, the rate of change (or slope) of CD4+ count or viral load can be measured for pre-specified intervals of weeks post infection.

5.3.1 CD4+ count

5.3.1.1 Marginal Model

Five piecewise linear components of weeks post infection, as described earlier, were fitted to CD4+ count. The model fitted had the following form

$$y_{ij} = \beta_0 + \beta_1 t_{ij}^{(0,2)} + \beta_2 t_{ij}^{(2,4)} + \beta_3 t_{ij}^{(4,8)} + \beta_4 t_{ij}^{(8,12)} + \beta_5 t_{ij}^{(8+)} + \varepsilon_{ij} \tag{5.8}$$

where y_{ij} is the j^{th} square root CD4+ count measurement for subject i ($j = 1, 2, ..., n_i$), β_0 is the intercept. $t_{ij}^{(0,2)}$ represents the piecewise component for 0 to 2 weeks post infection, while β_1 is the slope estimate for the change in square root CD4+ count for every week increase between 0 to 2 weeks post infection. Similarly, $t_{ij}^{(2,4)}$, $t_{ij}^{(4,8)}$, $t_{ij}^{(8,12)}$ and $t_{ij}^{(8+)}$ represent the piecewise components for 2 to 4, 4 to 8, 8 to 12 and 12 or more weeks post infection. The regression parameters β_2, β_3, β_4 and β_5 represent the corresponding piecewise slope parameters. The ε_{ij} is the random error associated with the j^{th} measurement for subject i. As with the linear model case in Section 5.2, various covariance structures were used to model the within-subject variation in the repeated measurements.

The CS covariance structure provided the better fit to the model as it had the lowest AIC. The population estimates of the piecewise linear effects are shown in Table 5.21. The results show that the intercept for CD4+ count is 31.1216 square root cells/$\mu\ell$ ($p<0.0001$). Within the first two weeks of acquiring HIV, the CD4+ count decreases by 4.4492 square root cells/$\mu\ell$ per week. This initial drop in CD4+ count is statistically significant with $p<0.0001$. In the following fortnight the CD4+ count recovers slightly by 0.2875 square root cells/$\mu\ell$ per week, however this is not a significant

Covariance Structure	CS	SP(EXP)	SP(POW)	SP(GAU)
-2 Log Likelihood	6748.4	6922.7	6922.7	7850.0
AIC	6752.4	6926.7	6926.7	7854.0
BIC	6756.7	6931.0	6931.0	7858.3
Chi-Square	1268.18	1093.88	1093.88	166.59
Pr>Chi-Square	<.0001	<.0001	<.0001	<.0001

Table 5.20: Fit statistics for fitting a piecewise linear effects marginal model to CD4+ count

Solution for Fixed Effects

Effect	Estimate	Std Err	DF	tValue	Prob>|t|
Intercept	31.1216	1.0384	61	29.97	<.0001
Slope 0 to 2	-4.4492	0.8601	1309	-5.17	<.0001
Slope 2 to 4	0.2875	0.7960	1309	0.36	0.7180
Slope 4 to 8	0.1678	0.1737	1309	0.97	0.3342
Slope 8 to 12	-0.3246	0.09275	1309	-3.50	0.0005
Slope 12+	-0.01764	0.003870	1309	-4.56	<.0001

Table 5.21: Fixed effects estimates for modelling a piecewise linear marginal model for CD4+ count

increase (p=0.7180) Between weeks 4 to 8 of HIV infection, CD4+ count continues a slight increase with 0.1678 square root cells/$\mu\ell$ per week, but this is still not significant (p=0.3342). From week 8 to 12 CD4+ count decreases at a rate of 0.3246 square root cells/$\mu\ell$ per week (p<0.0001) and after 12 weeks post infection decreases at a rate of 0.01764 square root cells/$\mu\ell$ per week (p<0.0001). These results show the initial drop in CD4+ count, its slight recovery thereafter and the eventual decline. The significant initial drop in CD4+ count cells is due to the impact of the invading disease pathogen. Immediately after this the body mounts its own immune response and this is the stage of the disease where a supposedly increase in CD4+ count is seen. However this seems only temporary as the body continually loses CD4+ cells and the infected subjects progressively move towards the AIDS stage of the disease.

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Std Err	Z Value	Prob>|Z|
CS	pid	12.8864	2.3899	5.39	<.0001
Residual		6.6170	0.2586	25.59	<.0001

Table 5.22: Covariance parameter estimates for modelling a piecewise linear marginal model for CD4+ count

The covariance parameter estimate for CS (12.8864) is statistically significant at p<0.0001 indi-

cating that there is significant within-subject variance, which also implies that observations within an individual are highly correlated, as expected in clustered correlated data. The model fitted above is a population averaged model or marginal model but also accounts for correlation between observations from the same individual.

5.3.1.2 Random Intercept Model

In order to determine whether there is significant between-subject variation in the intercept or baseline mean outcome, a random intercept is specified in the model. The model fitted has the form:

$$y_{ij} = (\beta_0 + b_{0i}) + \beta_1 t_{ij}^{(0,2)} + \beta_2 t_{ij}^{(2,4)} + \beta_3 t_{ij}^{(4,8)} + \beta_4 t_{ij}^{(8,12)} + \beta_5 t_{ij}^{(8+)} + \varepsilon_{ij} \tag{5.9}$$

where the added component b_{0i} is the subject-specific estimate for the intercept. Following the same strategy as before, various covariance structures are used to model the within-subject variance. An UN covariance structure is specified for the random intercept effect, since the default VC structure does not allow for the Null Model Likelihood Ratio test to be performed.

Covariance Structure	CS	SP(EXP)	SP(POW)	SP(GAU)
-2 Log Likelihood	6748.4	6578.8	6578.8	6692.9
AIC	6754.4	6584.8	6584.8	6698.9
BIC	6760.8	6591.2	6591.2	6705.3
Chi-Square	1268.18	1437.83	1437.83	1323.70
Pr>Chi-Square	<.0001	<.0001	<.0001	<.0001

Table 5.23: Fit statistics for fitting a piecewise linear effects marginal model to CD4+ count

The fit statistics indicate that either SP(POW) or SP(EXP) provide a better fit to the repeated measures data as both have the lowest AIC of 6584.8. The SP(POW) covariance structure was used to model the within-subject variance. The population estimates shown in Table 5.24 change slightly but the conclusions remain the same. The estimated effect for Slope 2 to 4 is now negative, but it is insignificant (p=0.8115).

The estimate for UN(1,1)=13.5069 is the between-subject variance for the intercept and this is statistically significant (p<0.0001) which shows that there is indeed variation between subjects in the intercept. At this point it should be stated that even though it is clear from the previous analysis that there would be significant variation in slopes between subjects, fitting a model with all piecewise time components as well as the intercept as random effects proved impossible as model convergence could not be achieved. However, a common random effect for all the piecewise slopes

Solution for Fixed Effects

Effect	Estimate	Std Err	DF	tValue	Prob>|t|
Intercept	31.1500	1.0284	61	30.29	<.0001
Slope 0 to 2	-4.0301	0.6358	1309	-6.34	<.0001
Slope 2 to 4	-0.1444	0.6053	1309	-0.24	0.8115
Slope 4 to 8	0.02992	0.1360	1309	0.22	0.8259
Slope 8 to 12	-0.2076	0.08543	1309	-2.43	0.0152
Slope 12+	-0.01686	0.003711	1309	-4.54	<.0001

Table 5.24: Fixed effects estimates for modelling a piecewise linear random intercept model for CD4+ count

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Std Err	Z Value	Prob>|Z|
UN(1,1)	pid	13.5069	2.5448	5.31	<.0001
SP(POW)	pid	0.7093	0.02353	30.14	<.0001
Residual		6.9886	0.3235	21.60	<.0001

Table 5.25: Covariance parameter estimates for modelling a piecewise linear random intercept model for CD4+ count

were fitted in the application of the joint modelling, in Section 5.5, where the more generalisable SAS procedure `NLMIXED` was used.

The correlation matrix of Fixed Effects in Table 5.26 shows how the different piecewise components are related to one another.

Effect	Intercept	Slope 0 to 2	Slope 2 to 4	Slope 4 to 8	Slope 8 to 12	Slope 12+
Intercept	1.0000	-0.4091	-0.2401	-0.2242	0.0107	-0.2306
Slope 0 to 2	-0.4091	1.0000	-0.6650	0.0250	-0.0757	0.0817
Slope 2 to 4	-0.2401	-0.6650	1.0000	-0.1850	0.0811	0.0198
Slope 4 to 8	-0.2242	0.0250	-0.1850	1.0000	-0.3653	-0.1179
Slope 8 to 12	0.0107	-0.0757	0.0811	-0.3653	1.0000	0.1422
Slope 12+	-0.2306	0.0817	0.0198	-0.1179	0.1422	1.0000

Table 5.26: Correlation Matrix of Fixed Effects fitting a random intercept model to CD4+ count with piecewise linear effects

The correlation matrix shows that adjacent piecewise slope estimates are more correlated than piecewise components further apart. One such correlation is between Slope 0 to 2 and Slope 2 to 4 with ρ=-0.6650. In the first two weeks of HIV infection the CD4+ count falls significantly

and then recovers between weeks 2 and 4. The rate of CD4+ count loss in the first fortnight is negatively correlated to the rate of CD4+ count recovery in the second fortnight. Slope 2 to 4 and subsequent Slope 4 to 8 have a weak negative correlation (ρ=-0.1850), while Slope 4 to 8 and subsequent Slope 8 to 12 have a relatively stronger negative correlation (ρ=-0.3653). There is another weaker correlation between Slope 8 to 12 and Slope 12+ (ρ=0.1422). Figure 5.3 graphically represents the fixed effect estimates for the model. When comparing this against Figure 2.2, it can be seen that the piecewise model gives a more clear picture on what is happening during the first stage of acute HIV infection.

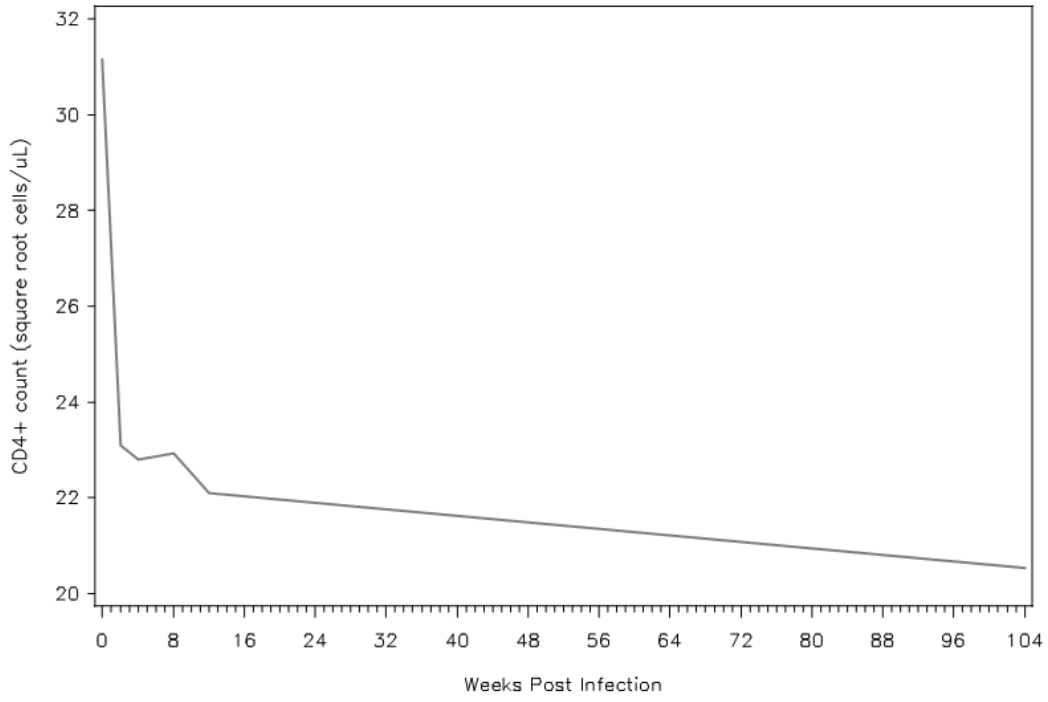

Figure 5.3: Modelling CD4+ count with time as piecewise effects

5.3.2 Viral Load

5.3.2.1 Marginal Model

The same five piecewise linear components of weeks post infection is fitted to viral load. This is done to make both the piecewise models comparable in their results regarding the evolution of the HIV markers. The model has the following form:

$$y_{ij} = \beta_0 + \beta_1 t_{ij}^{(0,2)} + \beta_2 t_{ij}^{(2,4)} + \beta_3 t_{ij}^{(4,8)} + \beta_4 t_{ij}^{(8,12)} + \beta_5 t_{ij}^{(8+)} + \varepsilon_{ij} \tag{5.10}$$

where y_{ij} is the j^{th} log viral load measurement for subject i ($j = 1, 2, ..., n_i$), β_0 is the intercept. $t_{ij}^{(0,2)}$ represents the piecewise component for 0 to 2 weeks post infection, while β_1 is the slope estimate for the change in log viral load for every week increase between 0 to 2 weeks post infection. Similarly, $t_{ij}^{(2,4)}$, $t_{ij}^{(4,8)}$, $t_{ij}^{(8,12)}$ and $t_{ij}^{(8+)}$ represent the piecewise components for 2 to 4, 4 to 8, 8 to 12 and 12 or more weeks post infection. The regression coefficients β_2, β_3, β_4 and β_5 represent the corresponding piecewise slope parameters. The ε_{ij} is the random error associated with the j^{th} measurement for subject i.

Covariance Structure	CS	SP(EXP)	SP(POW)	SP(GAU)
-2 Log Likelihood	2639.4	2812.0	2812.0	3519.5
AIC	2643.4	2816.0	2816.0	3523.5
BIC	2647.6	2820.3	2820.3	3527.7
Chi-Square	1085.95	913.31	913.31	205.85
Pr>Chi-Square	<.0001	<.0001	<.0001	<.0001

Table 5.27: Fit statistics for fitting a piecewise linear effects marginal model to viral load

The CS covariance structure provided the better fit to the model as it had the lowest AIC of 2643.4. The population estimates of the piecewise linear effects are shown in Table 5.28.

Solution for Fixed Effects

Effect	Estimate	Std Err	DF	tValue	Prob>|t|
Intercept	0.03470	0.02771	61	1.25	0.2153
Slope 0 to 2	2.6525	0.1145	1376	23.17	<.0001
Slope 2 to 4	-0.4022	0.1535	1376	-2.62	0.0089
Slope 4 to 8	0.03518	0.04935	1376	0.71	0.4760
Slope 8 to 12	-0.08449	0.02993	1376	-2.82	0.0048
Slope 12+	-0.00113	0.000741	1376	-1.52	0.1289

Table 5.28: Fixed effects estimates for modelling a piecewise linear marginal model for viral load

It is interesting to note that the fixed effect estimates show that the intercept for viral load is 0.03470 and that this is not significantly different from zero (p=0.2153). In the marginal viral load model where weeks post infection was fitted as a linear continuous predictor in Section 5.2.2, the intercept estimate was 4.0759 log copies/mℓ. This estimate did not make sense as an assumption is already made that log viral load is equal to zero when weeks post infection is zero. This happened because weeks post infection was fitted as a linear predictor and a straight line was fitted through log viral load. The intercept was where this line crossed the y-axis when weeks post infection was zero. Thus in a revised model, a no-intercept model was used. In the piecewise linear effects

model it is noted that the intercept is not significantly different from zero (p=0.2150) and hence the intercept is therefore omitted in the case of the piecewise model by using the `NOINT` option in the `MODEL` statement. The piecewise components of weeks post infection is adjusting for the initial peak of viral load and it's subsequent decline. Fit statistics show that the CS structure is still the best to model the within-subject variation.

Solution for Fixed Effects

Effect	Estimate	Std Err	DF	tValue	Prob>|t|
Slope 0 to 2	2.6594	0.1151	1376	23.11	<.0001
Slope 2 to 4	-0.4023	0.1536	1376	-2.62	0.0089
Slope 4 to 8	0.03520	0.04935	1376	0.71	0.4758
Slope 8 to 12	-0.08450	0.02993	1376	-2.82	0.0048
Slope 12+	-0.00113	0.000741	1376	-1.52	0.1289

Table 5.29: Fixed effects estimates for modelling a piecewise linear marginal no-intercept model for viral load

The fixed effect estimates reveal that viral load increases significantly in the first two weeks of infection at 2.6594 log copies/mℓ per week (p<0.0001). In the following fortnight it decreases at a slower rate of 0.4023 log copies/mℓ per week and this was statistically significant (p=0.0089). From 4 to 8 weeks post infection, there is a slight increase in viral load but this slope is not significantly different from zero (p=0.4758). Between 8 and 12 weeks post infection, viral load decreases significantly (-0.08450 log copies/mℓ per week with p-value=0.0048). Interestingly, viral load decreases after 12 weeks post infection, albeit very slightly at 0.00113 log copies/mℓ decrease per week but this is insignificant at p=0.1289.

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Std Err	Z Value	Prob>|Z|
CS	pid	0.4723	0.08758	5.39	<.0001
Residual		0.3057	0.01165	26.24	<.0001

Table 5.30: Covariance parameter estimates for modelling a piecewise linear marginal no-intercept model for viral load

The covariance parameter estimate for CS (0.4723) is statistically significant at p<0.0001, indicating that there is significant within-subject variance which also allows to account for correlation between observations from the same subject. As with modelling piecewise linear effects for CD4+ count, it is clear from previous analysis that there would be significant variation in slopes between subjects. However, fitting a model with all piecewise time components as random effects

proved impossible as model convergence could not be achieved. As noted previously, this could be achieved using a more generalisable model with SAS `PROC NLMIXED` where a common random effect was modelled for all the piecewise slopes.

Effect	Slope 0 to 2	Slope 2 to 4	Slope 4 to 8	Slope 8 to 12	Slope 12+
Slope 0 to 2	1.0000	-0.7882	0.1941	-0.0912	-0.1967
Slope 2 to 4	-0.7882	1.0000	-0.6296	0.1549	0.1567
Slope 4 to 8	0.1941	-0.6296	1.0000	-0.5486	-0.1191
Slope 8 to 12	-0.0912	0.1549	-0.5486	1.0000	-0.0513
Slope 12+	-0.1967	0.1567	-0.1191	-0.0513	1.0000

Table 5.31: Correlation Matrix of Fixed Effects fitting a marginal model to viral load with piecewise linear effects

The correlation matrix in Table 5.31 shows that adjacent piecewise slope estimates are more correlated than piecewise components further apart. The same was seen in the correlation matrix for the CD4+ count model above. The correlation between Slope 0 to 2 and Slope 2 to 4 is ρ=-0.7882. In the first two weeks of HIV infection the viral load increases rapidly and then falls between weeks 2 and 4. The rate of viral load increase in the first fortnight is negatively correlated to the rate of change in the second fortnight. Slope 2 to 4 and subsequent Slope 4 to 8 also have a strong negative correlation (ρ=-0.6296). Slope 4 to 8 and subsequent Slope 8 to 12 also have a strong negative correlation (ρ=-0.5486). There is a very weak correlation between Slope 8 to 12 and Slope 12+ (ρ=-0.0513). When representing the fixed effects estimates graphically in Figure 5.4, it can be seen that once again the piecewise model is more accurate in capturing the true evolution of HIV during the first weeks of infection when as depicted by the scatter plot and Loess smoothing line on the viral load measurements in Figure 2.3.

5.4 The Effects Of Left-censoring On Viral Load

The left-censoring of viral load was applied using a method proposed by Jacqmin-Gadda (2000) using the SAS procedure `PROC NLMIXED`. Results from the left-censored univariate model for viral load are shown in the table below, alongside the results from the model without left-censoring. It can be seen how accounting for left-censoring affects the parameter estimates. The difference in degrees of freedom (DF) for the two models is as a result of the way the repeated measures and subject effects are taken into account using the specific SAS procedures. When fitting `PROC MIXED` for viral load without left-censoring, the subject effect is taken into account through the `REPEATED` statement. When fitting a left-censored model for viral load, `PROC NLMIXED` is used which does

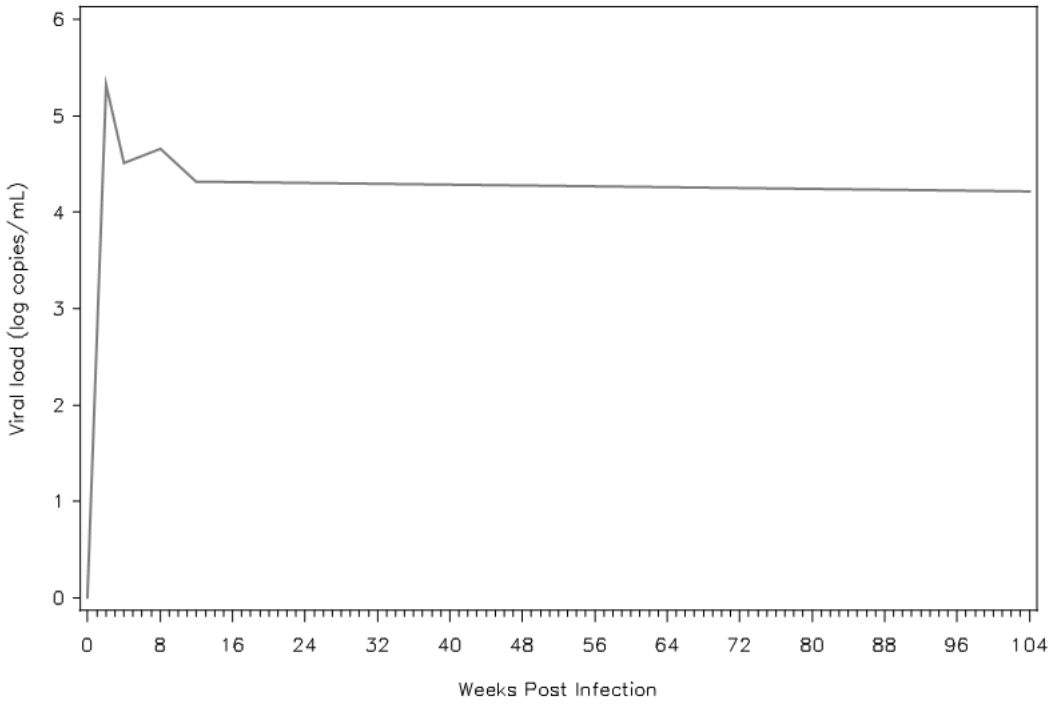

Figure 5.4: Modelling viral with time as piecewise effects

not accommodate for a `REPEATED` statement, however it can take into account the random subject effect by creating a participant variable and accounting for it with the `RANDOM` statement within the `NLMIXED` procedure. Also of importance is that maximum likelihood (ML) estimation was used in both the `MIXED` and `NLMIXED` procedures because `PROC NLMIXED` does not cater for restricted maximum likelihood (REML) estimation and the models need to be comparable.

The actual parameter estimates (Table 5.32), their standard errors and conclusions based on the estimates do not change much after accounting for left-censoring. However, it does seem as if the standard errors are slightly higher in the model adjusting for left-censoring. This is not surprising considering that the cohort under study was recently infected. Since the body's immune system has not had enough time to suppress the viral load during this early stage of HIV, not many viral load measurements are below the limit of detection (approximately only 5% of all measurements fell below the 400 copies/mℓ limit).

Table 5.33 gives the fit statistics for the two models of having viral load with and without left-censoring and shows that the model without left-censoring seems to provide the better fit to the data. When constructing the likelihood ratio test between the two models, the χ^2 test statistic is (2733.5-2601.5)=132 and this value is significant on the χ^2-distribution with two degrees of freedom

Solution for Fixed Effects

Without left-censoring						With left-censoring				
Effect	Estimate	Std Err	DF	tValue	Prob>|t |	Estimate	Std Err	DF	tValue	Prob>|t|
Slope 0 to 2	2.6594	0.0617	1376	43.13	<.0001	2.6585	0.0641	61	41.46	<.0001
Slope 2 to 4	-0.4023	0.0736	1376	-5.46	<.0001	-0.3931	0.0766	61	-5.13	<.0001
Slope 4 to 8	0.0352	0.0312	1376	1.13	0.2640	0.0286	0.0325	61	0.88	0.3834
Slope 8 to 12	-0.0845	0.0170	1376	-4.96	<.0001	-0.0871	0.0178	61	-4.90	<.0001
Slope 12+	-0.0011	0.0003	1376	-3.31	0.0016	-0.0011	0.0004	61	-3.21	0.0021

Table 5.32: Univariate results for fixed effects modelling viral load, with and without left-censoring.

Fit Statistics

Fit Statistic	Without left-censoring	With left-censoring
-2 Log Likelihood	2601.5	2733.5
AIC (smaller is better)	2615.5	2747.5
BIC (smaller is better)	2630.3	2762.4

Table 5.33: Fit statistics for modelling viral load, with and without left-censoring.

($p<0.0001$). This result works in favour of the model which does not apply left-censoring and this could be due to the fact that only 5% of viral load measurements are undetectable. Thus it can be concluded that left-censoring of viral load is not necessary for this particular cohort.

5.5 Joint Modelling

5.5.1 CD4+ count

5.5.1.1 Time as a Linear Effect

In Section 5.2.1 it was shown that when fitting weeks post infection as a linear continuous predictor, there was indeed a significant random intercept and random slope effect. Thus a random intercept and slope model will be adopted for the longitudinal measurement model. To fit the joint model for the measurement and time to event processes, an exponential survival model will be used to model time to informative drop-out (in the current application this is the time to ARV initiation). Using the method by Henderson, Diggle and Dobson (2000) described in Section 4.3, the longitudinal and survival sub-models are joined using a Gaussian latent process $W_i(t) = \{W_{1i}(t), W_{2i}(t)\}$. The association between the longitudinal and survival sub-models are described through the cross-correlation between $W_{1i}(t)$ and $W_{2i}(t)$ and the relationship between these processes are specified. The following sections will look at different latent associations between the longitudinal model for

CD4+ count and the survival model for informative drop-out.

Model 1: $W_1(t) = U_0 + U_1 t$ **and** $W_2(t) = 0$

A random intercept and slope was fitted to the longitudinal model, with $W_1(t) = U_0 + U_1 t$. There was no joint effect specified as $W_2(t) = 0$. Note that now the longitudinal and survival models are not joined through the Gaussian latent process and this would be the same as fitting the two models separately.

Fit Statistics

-2 Log Likelihood	6960.8
AIC (smaller is better)	6974.8
BIC (smaller is better)	6989.7

Table 5.34: Fit statistics for a joint CD4+ count and informative drop-out model with $W_1(t) = U_0 + U_1 t$ and $W_2(t) = 0$

Parameter Estimates

Parameter	Estimate	Std Err	DF	t Value	Pr >|t |	Lower	Upper
Surv Intercept	8.6220	0.3014	60	28.61	<.0001	8.0192	9.2248
Long Intercept	23.1499	0.4531	60	51.09	<.0001	22.2435	24.0563
Long Week	-0.0318	0.00518	60	-6.14	<.0001	-0.04216	-0.02145
Residual	6.0689	0.2437	60	24.90	<.0001	5.5814	6.5565
v11	11.9850	2.2690	60	5.28	<.0001	7.4463	16.5236
v12	-0.0016	0.0174	60	-0.09	0.9257	-0.03640	0.03314
v22	0.00128	0.00033	60	3.88	0.0003	0.000620	0.001938

Table 5.35: Parameter estimates for a joint CD4+ count and informative drop-out model with $W_1(t) = U_0 + U_1 t$ and $W_2(t) = 0$

The fit statistics and parameter estimates are shown in Tables 5.34 and 5.35 respectively. Both the longitudinal and survival sub-models have been modelled using the same `PROC NLMIXED`. The intercept for the exponential survival model is 8.6220 and the intercept for the longitudinal CD4+ count model is 23.1499 square root cells/$\mu\ell$. The CD4+ count decreases at a rate of 0.0318 square root cells/$\mu\ell$ per week and as seen before, this slope is statistically significant (p<0.0001). The residual of 6.0689 in Table 5.35 is the residual variance component, the variance of CD4+ count conditional on participant. The estimates v11, v12 and v22 are the covariance estimates for the random intercept and slope.

Model 2: $W_1(t) = U_0 + U_1 t$ **and** $W_2(t) = r_0 U_0$

Now a joint effect is brought into the Gaussian process linking the longitudinal and survival models through the random intercept. Here the term $r_0 U_0$ gets added to the survival model in the `NLMIXED` procedure.

Fit Statistics

-2 Log Likelihood	6954.9
AIC (smaller is better)	6970.9
BIC (smaller is better)	6987.9

Table 5.36: Fit statistics for a joint CD4+ count and informative drop-out model with $W_1(t) = U_0 + U_1 t$ and $W_2(t) = r_0 U_0$

Parameter Estimates

Parameter	Estimate	Std Err	DF	t Value	Pr >|t |	Lower	Upper
Surv Intercept	8.8813	0.4038	60	21.99	<.0001	8.0735	9.6891
Long Intercept	23.1414	0.4538	60	50.99	<.0001	22.2337	24.0491
Long Week	-0.0314	0.00503	60	-6.25	<.0001	-0.04148	-0.02136
Residual	6.0844	0.2449	60	24.84	<.0001	5.5945	6.5743
r0	0.2481	0.1128	60	2.20	0.0317	0.02252	0.4737
v11	12.0262	2.2774	60	5.28	<.0001	7.4707	16.5817
v12	-0.00090	0.01693	60	-0.05	0.9580	-0.03477	0.03297
v22	0.00119	0.00031	60	3.86	0.0003	0.000575	0.001812

Table 5.37: Parameter estimates for a joint CD4+ count and informative drop-out model with $W_1(t) = U_0 + U_1 t$ and $W_2(t) = r_0 U_0$

The fit statistics and parameter estimates are shown in Tables 5.36 and 5.37 respectively. The parameter estimates for this model do not change much when compared to the previous model parameter estimates in Table 5.35. Interestingly the new addition, the joint effect estimate r_0, which accounts for the longitudinal and survival model to be linked through the random intercept, is now significant with p=0.0317. This suggests that there is indeed an importance in joining the longitudinal CD4+ count model to the survival model in order to take into account informative drop-out. This also gives a more valid representation of the true process. Note that all three fit statistics are in support of the joint model.

Model 3: $W_1(t) = U_0 + U_1 t$ **and** $W_2(t) = r_0 U_0 + r_1 U_1$

Taking the joint modelling a step further, the survival model is linked to the longitudinal model via both a random intercept and slope. The terms $r_0 U_0 + r_1 U_1$ are added to the survival model and the following results are obtained. The parameter estimates in Table 5.39 do not change much from the previous models, although the intercept for the exponential survival model has increased. However the standard errors in the joint model under Model 3 are generally higher as compared to those under Model 2. Both estimates r_0 and r_1 which measure the joint effect through the random intercept and slope are statistically significant (p=0.0307 and p=0.0006, respectively) indicating that there is a significant joint effect in both the intercept and slope. It is also noted that all fit statistics are in support of the joint sub-models via a random intercept and slope.

Fit Statistics

-2 Log Likelihood	6934.4
AIC (smaller is better)	6952.4
BIC (smaller is better)	6971.5

Table 5.38: Fit statistics for a joint CD4+ count and informative drop-out model with $W_1(t) = U_0 + U_1 t$ and $W_2(t) = r_0 U_0 + r_1 U_1$

Parameter Estimates

Parameter	Estimate	Std Err	DF	t Value	Pr >|t |	Lower	Upper
Surv Intercept	9.6204	0.6162	60	15.61	<.0001	8.3877	10.8530
Long Intercept	23.1486	0.4528	60	51.12	<.0001	22.2428	24.0543
Long Week	-0.03184	0.00513	60	-6.21	<.0001	-0.04210	-0.02158
Residual	6.0624	0.2430	60	24.94	<.0001	5.5762	6.5485
r0	0.3173	0.1434	60	2.21	0.0307	0.03053	0.6041
r1	39.1503	10.8542	60	3.61	0.0006	17.4387	60.8618
v11	11.9747	2.2667	60	5.28	<.0001	7.4408	16.5087
v12	-0.0012	0.0172	60	-0.07	0.9449	-0.03567	0.03328
v22	0.00127	0.00032	60	3.93	0.0002	0.000622	0.001910

Table 5.39: Parameter estimates for a joint CD4+ count and informative drop-out model with $W_1(t) = U_0 + U_1 t$ and $W_2(t) = r_0 U_0 + r_1 U_1$

Summary

Joint Model	Model 1			Model 2			Model 3		
Parameter	Estimate	Std Err	Pr >|t |	Estimate	Std Err	Pr >|t |	Estimate	Std Err	Pr >|t |
Surv Intercept	8.6220	0.3014	<.0001	8.8813	0.4038	<.0001	9.6204	0.6162	<.0001
Long Intercept	23.1499	0.4531	<.0001	23.1414	0.4538	<.0001	23.1486	0.4528	<.0001
Long Week	-0.0318	0.00518	<.0001	-0.0314	0.00503	<.0001	-0.03184	0.00513	<.0001
Residual	6.0689	0.2437	<.0001	6.0844	0.2449	<.0001	6.0624	0.2430	<.0001
v11	11.9850	2.2690	<.0001	12.0262	2.2774	<.0001	11.9747	2.2667	<.0001
v12	-0.0016	0.0174	0.9257	-0.00090	0.01693	0.9580	-0.0012	0.0172	0.9449
v22	0.00128	0.00033	0.0003	0.00119	0.00031	0.0003	0.00127	0.00032	0.0002
r0	-	-	-	0.2481	0.1128	0.0317	0.3173	0.1434	0.0307
r1	-	-	-	-	-	-	39.1503	10.8542	0.0006
-2 Log Likelihood	6960.8			6954.9			6934.4		
AIC	6974.8			6970.9			6952.4		
BIC	6989.7			6987.9			6971.5		

Table 5.40: Summary of joint modelling results for a linear CD4+ count model

Model 1: $W_1(t) = U_0 + U_1 t$; $W_2(t) = 0$

Model 2: $W_1(t) = U_0 + U_1 t$; $W_2(t) = r_0 U_0$

Model 3: $W_1(t) = U_0 + U_1 t$; $W_2(t) = r_0 U_0 + r_1 U_1$

The differences between the three specifications of the latent Gaussian processes can be seen when comparing the three models and their effect estimates side-by-side. The model that provided the best fit was Model 3 with $W_1(t) = U_0 + U_1 t$ and $W_2(t) = r_0 U_0 + r_1 U_1$, as it had the lowest AIC and BIC fit statistics. It is interesting to note that the survival intercept of 9.6204 is higher compared to Model 1 and 2. The other estimates in the model do not change significantly between the different models. There is a significant random intercept and slope effect in Model 3 with the between-subject variation for the intercept and slope being statistically significant with p<0.0001 and p=0.0002 respectively. Both the joint effect estimates r_0 and r_1 are significant indicating that there is a significant joint effect through the random intercept and slope (p=0.0307 and p=0.0006, respectively). The graphical representation of the joint models are shown in Figure 5.5.

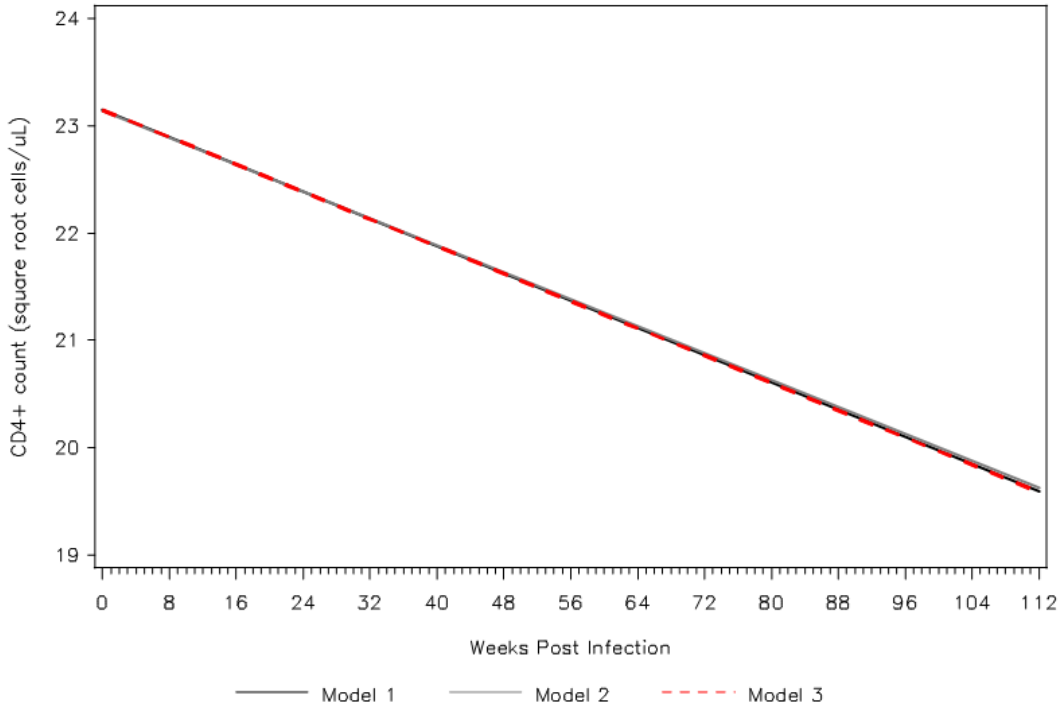

Figure 5.5: Different joint models for modelling CD4+ count with time as a linear effect

5.5.1.2 Piecewise Linear Effects

The benefits of fitting a piecewise linear effects model to CD4+ count was seen in Section 5.3.1. Indeed, the evolution of this HIV marker could be assessed more closely during the acute infection stage of the infection. This section will look at the effects of joining the survival sub-model to the longitudinal model with piecewise effects, as a better approach of representing the true underlying process compared to just a linear combination of the random intercept and slope.

Model 1: $W_1(t) = U_0$ and $W_2(t) = 0$, piecewise

This specification of the latent Gaussian process allows for a random intercept in the longitudinal CD4+ count model, with no link to the survival model for informative drop-out. Tables 5.41 and 5.42 show the fit statistics and parameter estimates for this joint model. The longitudinal and survival sub-models are not linked through a latent Gaussian process. This model is equivalent to fitting a random intercept mixed piecewise model to the CD4+ count data and a survival model separately, where v11 is the estimate for the between-subject variation in the intercept for CD4+ count.

Fit Statistics

-2 Log Likelihood	6943.3
AIC (smaller is better)	6961.3
BIC (smaller is better)	6980.4

Table 5.41: Fit statistics for a joint piecewise CD4+ count and informative drop-out model with $W_1(t) = U_0$ and $W_2(t) = 0$

Parameter Estimates

Parameter	Estimate	Std Err	DF	t Value	Pr >|t |	Lower	Upper
Surv Intercept	8.6220	0.3014	61	28.61	<.0001	8.0194	9.2246
Long Intercept	31.1290	0.6878	61	45.26	<.0001	29.7538	32.5043
Slope 0 to 2	-4.4587	0.5990	61	-7.44	<.0001	-5.6565	-3.2610
Slope 2 to 4	0.2950	0.6010	61	0.49	0.6253	-0.9067	1.4967
Slope 4 to 8	0.1674	0.1499	61	1.12	0.2684	-0.1323	0.4672
Slope 8 to 12	-0.3245	0.07985	61	-4.06	0.0001	-0.4842	-0.1648
Slope 12+	-0.01764	0.001593	61	-11.08	<.0001	-0.02083	-0.01446
Residual	6.5919	0.2572	61	25.63	<.0001	6.0777	7.1062
v11	12.6672	2.3299	61	5.44	<.0001	8.0082	17.3261

Table 5.42: Parameter estimates for a joint piecewise CD4+ count and informative drop-out model with $W_1(t) = U_0$ and $W_2(t) = 0$

The parameter estimates for the measurement process in Table 5.42 gives quite similar conclusions as those in Table 5.24 except for Slope 2 to 4 weeks post infection where the estimates are of opposite signs, but both are statistically insignificant (p=0.6253 and p=0.6053, respectively).

Model 2: $W_1(t) = U_0$ **and** $W_2(t) = r_0 U_0$**, piecewise**

Now a joint effect is specified through the random intercept accounting for the informative drop-out which occurs later on during follow-up.

Fit Statistics

-2 Log Likelihood	6925.6
AIC (smaller is better)	6945.6
BIC (smaller is better)	6966.9

Table 5.43: Fit statistics for a joint piecewise CD4+ count and informative drop-out model with $W_1(t) = U_0$ and $W_2(t) = r_0 U_0$

Parameter Estimates

Parameter	Estimate	Std Err	DF	t Value	Pr >|t |	Lower	Upper
Surv Intercept	9.5782	0.6288	61	15.23	<.0001	8.3209	10.8355
Long Intercept	31.1147	0.6881	61	45.22	<.0001	29.7388	32.4906
Slope 0 to 2	-4.4250	0.5989	61	-7.39	<.0001	-5.6226	-3.2274
Slope 2 to 4	0.2757	0.6008	61	0.46	0.6479	-0.9257	1.4772
Slope 4 to 8	0.1636	0.1498	61	1.09	0.2793	-0.1361	0.4632
Slope 8 to 12	-0.3245	0.07981	61	-4.07	0.0001	-0.4841	-0.1649
Slope 12+	-0.01777	0.001591	61	-11.17	<.0001	-0.02095	-0.01459
Residual	6.5910	0.2571	61	25.64	<.0001	6.0770	7.1051
r0	0.5090	0.1455	61	3.50	0.0009	0.2180	0.8000
v11	12.7010	2.3364	61	5.44	<.0001	8.0291	17.3729

Table 5.44: Parameter estimates for a joint piecewise CD4+ count and informative drop-out model with $W_1(t) = U_0$ and $W_2(t) = r_0 U_0$

It can be seen from the parameter estimates in Table 5.44 that r0, the joint effect through the random intercept, is statistically significant (p=0.0009) indicating that there is indeed an important joint effect via the random intercept. The standard error of the survival model fixed effect parameter has increased because of the extra variability accounted for in this model. Otherwise the parameter estimates of the longitudinal measurement process have not changed much and standard errors are still in the same order of magnitude.

Model 3: $W_1(t) = U_0 + U_1 t$ **and** $W_2(t) = r_0 U_0 + r_1 U_1$**, piecewise**

This following specification of the latent Gaussian process indicates that the longitudinal and survival sub-models will be joined through both a random intercept and slope. The fit statistics and parameter estimates follow in Tables 5.45 and 5.46

Fit Statistics

-2 Log Likelihood	6632.6
AIC (smaller is better)	6658.6
BIC (smaller is better)	6686.2

Table 5.45: Fit statistics for a joint piecewise CD4+ count and informative drop-out model with $W_1(t) = U_0 + U_1 t$ and $W_2(t) = r_0 U_0 + r_1 U_1$

Parameter Estimates

Parameter	Estimate	Std Err	DF	t Value	Pr >|t |	Lower	Upper
Surv Intercept	9.7517	0.6571	60	14.84	<.0001	8.4374	11.0660
Long Intercept	31.1179	0.6307	60	49.34	<.0001	29.8563	32.3796
Slope 0 to 2	-4.2731	0.5201	60	-8.22	<.0001	-5.3136	-3.2327
Slope 2 to 4	0.0970	0.5200	60	0.19	0.8527	-0.9432	1.1372
Slope 4 to 8	0.1445	0.1307	60	1.11	0.2733	-0.1169	0.4058
Slope 8 to 12	-0.2566	0.06950	60	-3.69	0.0005	-0.3957	-0.1176
Slope 12+	-0.0253	0.005080	60	-4.99	<.0001	-0.03550	-0.01517
Residual	4.7788	0.1915	60	24.95	<.0001	4.3957	5.1618
r0	0.3478	0.1431	60	2.43	0.0181	0.06158	0.6340
r1	42.8867	11.5358	60	3.72	0.0004	19.8117	65.9618
v11	12.1906	2.2813	60	5.34	<.0001	7.6274	16.7539
v12	-0.00333	0.01698	60	-0.20	0.8451	-0.03729	0.03063
v22	0.001246	0.000304	60	4.10	0.0001	0.000638	0.001853

Table 5.46: Parameter estimates for a joint piecewise CD4+ count and informative drop-out model with $W_1(t) = U_0 + U_1 t$ and $W_2(t) = r_0 U_0 + r_1 U_1$

As can be seen from the parameter estimates in Table 5.46 that the joint effect estimates which combine the longitudinal and survival sub-models through the random intercept and slope are significant with p=0.0181 and p=0.0004, respectively. Overall, most standard errors of the fixed effects in Table 5.46 are higher than those in Table 5.44 except parameter r0 which has reduced slightly from 0.1455 to 0.1431. It should be noted that the random slope and intercept effects are not significantly correlated (p=0.8451).

Summary

Joint Model	Model 1			Model 2			Model 3		
Parameter	Estimate	Std Err	Pr >|t |	Estimate	Std Err	Pr >|t |	Estimate	Std Err	Pr >|t |
Surv Intercept	8.6220	0.3014	<.0001	9.5782	0.6288	<.0001	9.7517	0.6571	<.0001
Long Intercept	31.1290	0.6878	<.0001	31.1147	0.6881	<.0001	31.1179	0.6307	<.0001
Slope 0 to 2	-4.4587	0.5990	<.0001	-4.4250	0.5989	<.0001	-4.2731	0.5201	<.0001
Slope 2 to 4	0.2950	0.6010	0.6253	0.2757	0.6008	0.6479	0.0970	0.5200	0.8527
Slope 4 to 8	0.1674	0.14992	0.2684	0.1636	0.1498	0.2793	0.1445	0.1307	0.2733
Slope 8 to 12	-0.3245	0.07985	0.0001	-0.3245	0.07981	0.0001	-0.2566	0.06950	0.0005
Slope 12+	-0.01764	0.001593	<.0001	-0.01777	0.001591	<.0001	-0.0253	0.005080	<.0001
Residual	6.5919	0.25723	<.0001	6.5910	0.2571	<.0001	4.7788	0.1915	<.0001
v11	12.6672	2.3299	<.0001	12.7010	2.3364	<.0001	12.1906	2.2813	<.0001
v12	-	-	-	-	-	-	-0.00333	0.01698	0.8451
v22	-	-	-	-	-	-	0.001246	0.000304	0.0001
r0	-	-	-	0.5090	0.1455	0.0009	0.3478	0.1431	0.0181
r1	-	-	-	-	-	-	42.8867	11.5358	0.0004
-2 Log Likelihood	6943.3			6925.6			6632.6		
AIC	6961.3			6945.6			6658.6		
BIC	6980.4			6966.9			6686.2		

Table 5.47: Summary of joint modelling results for a piecewise CD4+ count model

Model 1: $W_1(t) = U_0$; $W_2(t) = 0$

Model 2: $W_1(t) = U_0$; $W_2(t) = r_0 U_0$

Model 3: $W_1(t) = U_0 + U_1 t$; $W_2(t) = r_0 U_0 + r_1 U_1$

Table 5.47 depicts the different methods of joint modelling used on the piecewise model for CD4+ count and it can be seen that Model 3 has the best fit since it has the lowest AIC and BIC fit statistics. Again this model has $W_1(t) = U_0 + U_1 t$ and $W_2(t) = r_0 U_0 + r_1 U_1$ where the longitudinal and survival sub-models are joined through a random intercept and common random slope.

Both Model 2 and Model 3 have larger survival intercepts and both models fit better than Model 1, which is in fact not applying joint modelling at all. Model 3 has a lower CD4+ count decline slope in the first two weeks of infection compared to Model 1, with a decrease of 4.27 square root cells/$\mu\ell$l per week. The CD4+ count recovery between 2 and 4 weeks post infection is also predicted to be lower than previously modelled in Model 1 and 2 at 0.097 square root cells/$\mu\ell$ per week. However this rebound in CD4+ count is not statistically significant in all three models. Of interest as well is that the slope after 12 weeks post infection is also steeper compared to the other models with a decrease of -0.025 square root cells/$\mu\ell$ per week.

When looking at this data graphically in Figure 5.6, it can be seen that Model 3 is predicting CD4+ count to be lower after 12 weeks post infection and participants are doing worse than what was previously thought. The effect of adjusting for informative drop-out can be seen more clearly in Figure 5.6, as CD4+ count is the main HIV marker used to determine when someone should be initiated on ARVs. Not accounting for these participants dropping out of the study because their CD4+ count is low gives over-optimistic results, when in fact the CD4+ count should actually be lower, as can be seen in Figure 5.6. Thus a model which includes individual heterogeneity as well as informative drop-out has a better predictive feature of the disease than one with none of these.

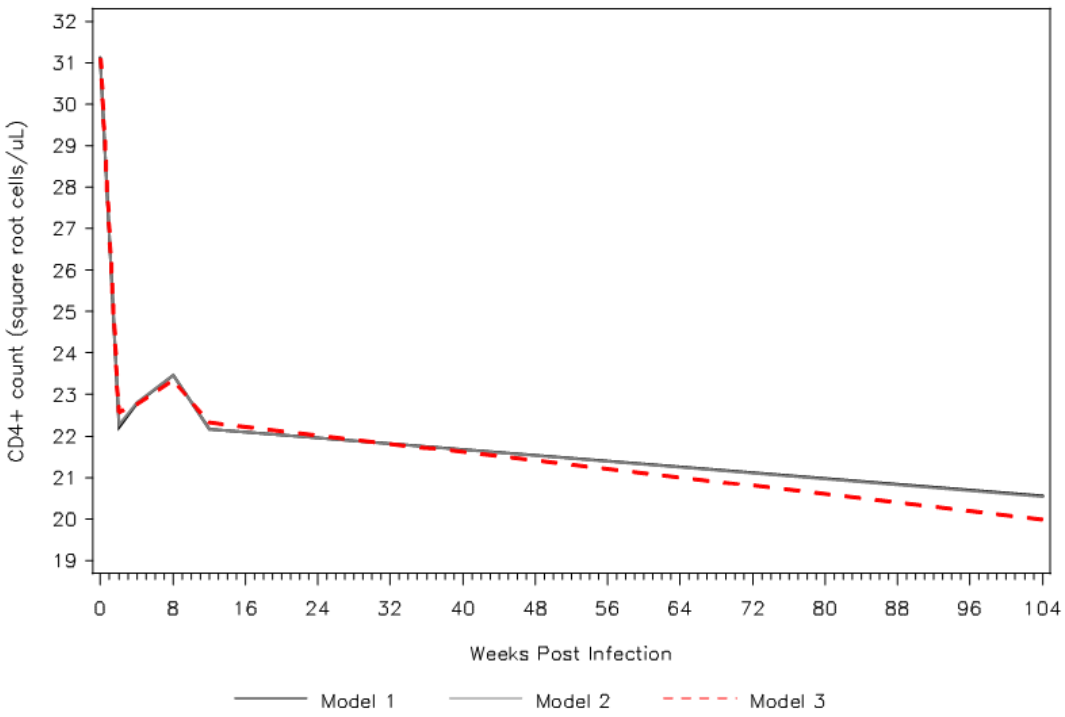

Figure 5.6: Different joint models for modelling CD4+ count with time as piecewise effect

5.5.2 Viral load

5.5.2.1 Time as a Linear Effect

In Section 5.2.2 it was discussed that when modelling viral load over weeks post infection, where week 0 represents the HIV negative state, there is no need to have an intercept for the model since viral load is zero when you are HIV uninfected. It was also highlighted that the necessary model, when modelling time as a linear component, was a no-intercept random slope model.

Model 1: $W_1(t) = U_1 t$ **and** $W_2(t) = 0$

This specification of the latent Gaussian process is equivalent to fitting a random slope model. Note that there is no joint link between the longitudinal and survival sub-models. The fit statistics and parameter estimates for this model is shown in Tables 5.48 and 5.49. As seen in the analysis, the intercept for the exponential survival model is 8.623. The slope of viral load over weeks post infection is 0.05427 log copies/mℓ per week (p<0.0001). The residual of 8.9074 is the residual variance component and the estimate v22 is the variance of the random slope.

Fit Statistics

-2 Log Likelihood	7569.6
AIC (smaller is better)	7577.6
BIC (smaller is better)	7586.1

Table 5.48: Fit statistics for a joint viral load and informative drop-out model with $W_1(t) = U_1 t$ and $W_2(t) = 0$

Parameter Estimates

Parameter	Estimate	Std Err	DF	t Value	Pr >|t |	Lower	Upper
Surv Intercept	8.6230	0.3015	61	28.60	<.0001	8.0201	9.2259
Long Week	0.05427	0.003604	61	15.06	<.0001	0.04706	0.06147
Residual	8.9074	0.3475	61	25.64	<.0001	8.2126	9.6022
v22	0.000489	0.000172	61	2.85	0.0059	0.000146	0.000833

Table 5.49: Parameter estimates for a joint viral load and informative drop-out model with $W_1(t) = U_1 t$ and $W_2(t) = 0$

Model 2: $W_1(t) = U_0 + U_1 t$ **and** $W_2(t) = r_0 U_0$

Now a joint link is specified through a random intercept. Even though there is no need for an intercept for the viral load model, the joint estimate is introduced to see whether there is a significant joint effect between the average viral load in the longitudinal model and the informative drop-out in the survival model.

Fit Statistics

-2 Log Likelihood	4824.4
AIC (smaller is better)	4838.4
BIC (smaller is better)	4853.3

Table 5.50: Fit statistics for a joint viral load and informative drop-out model with $W_1(t) = U_0 + U_1 t$ and $W_2(t) = r_0 U_0$

From the fit statistics in Table 5.50 an improvement can already be seen as the AIC of 4838.4 is much smaller than that in Table 5.48 (AIC=7577.6). The parameter estimates in Table 5.51 show that the joint effect estimate r0 (-2.0932) is statistically significant with p=0.0047. This supports the fact that there is a significant joint effect between the longitudinal viral load model and the survival model. Note that the survival intercept has increased approximately two-fold, compared to the results shown in 5.49 where no joint effect is present. Also note that the slope estimate for viral load over weeks post infection has decreased, being only 0.008294 log copies/mℓ per week and

Parameter Estimates

Parameter	Estimate	Std Err	DF	t Value	Pr >|t |	Lower	Upper
Surv Intercept	18.0034	3.4130	60	5.27	<.0001	11.1764	24.8303
Long Week	0.008294	0.004992	60	1.66	0.1019	-0.00169	0.01828
Residual	1.1097	0.04300	60	25.81	<.0001	1.0237	1.1957
r0	-2.0932	0.7126	60	-2.94	0.0047	-3.5185	-0.6679
v11	16.9043	3.0551	60	5.53	<.0001	10.7931	23.0154
v12	-0.02492	0.02093	60	-1.19	0.2386	-0.06679	0.01696
v22	0.000047	0.000064	60	0.74	0.4599	-0.00008	0.000175

Table 5.51: Parameter estimates for a joint viral load and informative drop-out model with $W_1(t) = U_0 + U_1 t$ and $W_2(t) = r_0 U_0$

this is no longer significant (p=0.1019). This may imply that much of the systematic component of viral load evolution post-infection is accounted for by the drop-out model linked via the random intercept term $r_0 U_0$.

Model 3: $W_1(t) = U_1 t$ **and** $W_2(t) = r_1 U_1$

This particular specification of the latent Gaussian process links the longitudinal and survival sub-models through the random slope only. The fit statistics and parameter estimates are shown in Tables 5.52 and 5.53.

Fit Statistics

-2 Log Likelihood	7553.0
AIC (smaller is better)	7563.0
BIC (smaller is better)	7573.6

Table 5.52: Fit statistics for a joint viral load and informative drop-out model with $W_1(t) = U_1 t$ and $W_2(t) = r_1 U_1$

This model shows that there is a significant joint effect r1 via the random slope with the effect estimate being -48.1210 (p=0.0007).

Parameter Estimates

Parameter	Estimate	Std Err	DF	t Value	Pr >|t |	Lower	Upper
Surv Intercept	8.9318	0.4242	61	21.06	<.0001	8.0836	9.7800
Long Week	0.05475	0.003643	61	15.03	<.0001	0.04747	0.06204
Residual	8.8689	0.3430	61	25.86	<.0001	8.1830	9.5548
r1	-48.1210	13.4430	61	-3.58	0.0007	-75.0019	-21.2401
v22	0.000536	0.000174	61	3.07	0.0032	0.000187	0.000885

Table 5.53: Parameter estimates for a joint viral load and informative drop-out model with $W_1(t) = U_1 t$ and $W_2(t) = r_1 U_1$

Model 4: $W_1(t) = U_0 + U_1 t$ **and** $W_2(t) = r_0 U_0 + r_1 U_1$

The following processes link the longitudinal and survival sub-models through both through a random intercept and slope. However this model failed to converge properly and could not be used. A summary of the fitted models for viral load outcome are given in Table 5.54.

Summary

Joint Model	Model 1			Model 2			Model 3		
Parameter	Estimate	Std Err	Pr >|t |	Estimate	Std Err	Pr >|t |	Estimate	Std Err	Pr >|t |
Surv Intercept	8.6230	0.3015	<.0001	18.0034	3.4130	<.0001	8.9318	0.4242	<.0001
Long Week	0.05427	0.003604	<.0001	0.008294	0.004992	0.1019	0.05475	0.003643	<.0001
Residual	8.9074	0.3475	<.0001	1.1097	0.04300	<.0001	8.8689	0.3430	<.0001
v11	-	-	-	16.9043	3.0551	<.0001	-	-	-
v12	-	-	-	-0.02492	0.02093	0.2386	-	-	-
v22	0.000489	0.000172	0.0059	0.000047	0.000064	0.4599	0.000536	0.000174	0.0032
r0	-	-	-	-2.0932	0.7126	0.0047	-	-	-
r1	-	-	-	-	-	-	-48.1210	13.44308	0.0007
-2 Log Likelihood	7569.6			4824.4			7553.0		
AIC	7577.6			4838.4			7563.0		
BIC	7586.1			4853.3			7573.6		

Table 5.54: Summary of joint modelling results for a linear viral load model

Model 1: $W_1(t) = U_1 t$; $W_2(t) = 0$

Model 2: $W_1(t) = U_0 + U_1 t$; $W_2(t) = r_0 U_0$

Model 3: $W_1(t) = U_1 t$; $W_2(t) = r_1 U_1$

It would appear from Table 5.54 that Model 2 provides the best to the data. However, with an inflated survival intercept and an extremely low slope estimate for the log viral load change over time, this model does not seem to fit the data well. Figure 5.7 shows the estimates for the longitudinal data, and as discussed previously, the linear model in general is not a good model to fit to the acute infection data.

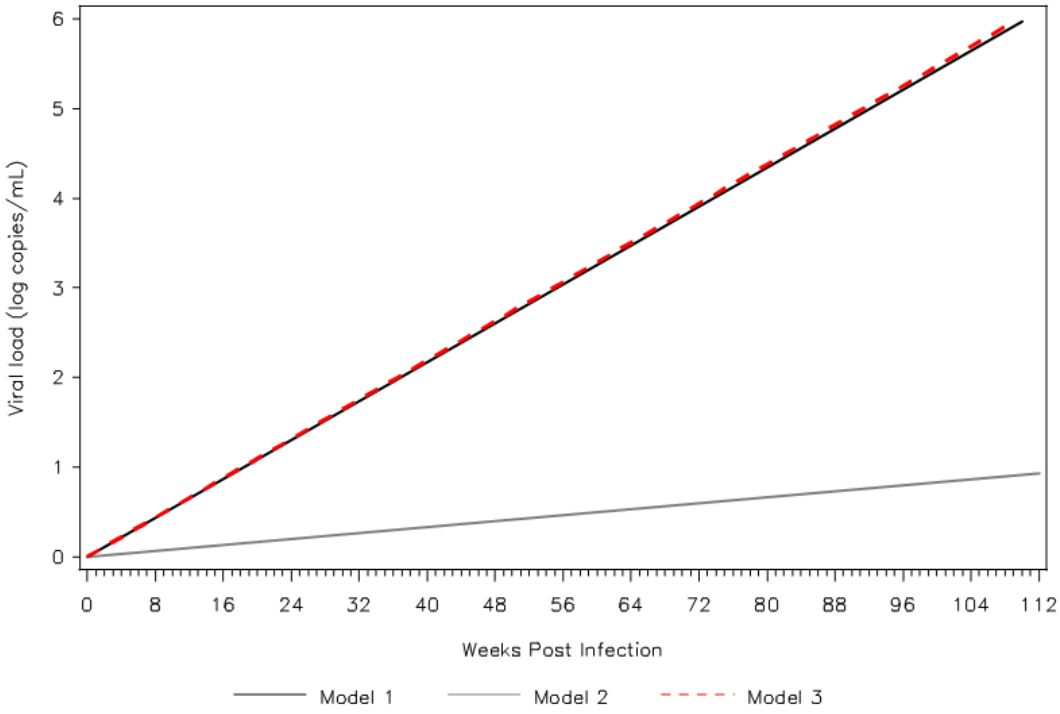

Figure 5.7: Different joint models for modelling viral load with time as a linear effect

5.5.2.2 Piecewise Linear Effects

As was seen in 5.3.2, a piecewise model can show the evolution of viral load as a disease marker more clearly including the peak in the initial few weeks post infection. Now the effect of joint modelling, taking into account informative drop-out, will be assessed using the piecewise linear mixed model to model the longitudinal viral load measurements.

Model 1: $W_1(t) = U_1 t$ **and** $W_2(t) = 0$**, piecewise**

This specification of the latent Gaussian process fits a common random effect to the piecewise slopes and there is no joint effect between the longitudinal and survival sub-models. The fit statistics and parameter estimates are given in Tables 5.55 and 5.56 respectively.

Fit Statistics

-2 Log Likelihood	3280.6
AIC (smaller is better)	3296.6
BIC (smaller is better)	3313.6

Table 5.55: Fit statistics for a joint piecewise viral load and informative drop-out model with $W_1(t) = U_1 t$ and $W_2(t) = 0$

Parameter Estimates

Parameter	Estimate	Std Err	DF	t Value	Pr >|t |	Lower	Upper
Surv Intercept	8.6230	0.3015	61	28.60	<.0001	8.0201	9.2259
Slope 0 to 2	2.7072	0.06384	61	42.41	<.0001	2.5796	2.8349
Slope 2 to 4	-0.4562	0.08573	61	-5.32	<.0001	-0.6277	-0.2848
Slope 4 to 8	0.05552	0.03593	61	1.55	0.1275	-0.01632	0.1274
Slope 8 to 12	-0.09631	0.02001	61	-4.81	<.0001	-0.1363	-0.05630
Slope 12+	-0.00111	0.001844	61	-0.60	0.5490	-0.00480	0.002576
Residual	0.4223	0.01614	61	26.17	<.0001	0.3901	0.4546
v22	0.000184	0.000038	61	4.79	<.0001	0.000107	0.000261

Table 5.56: Parameter estimates for a joint piecewise viral load and informative drop-out model with $W_1(t) = U_1 t$ and $W_2(t) = 0$

The results in Table 5.56 show that there is indeed a significant common random effect between the slope estimates, since the between-subject variance component v22=0.000184 is statistically significant (p<0.0001).

Model 2: $W_1(t) = U_0$ and $W_2(t) = r_0 U_0$, piecewise

A joint effect is added connecting the longitudinal piecewise viral load model to the survival model through a random intercept. The U_0 term is added to the viral load model as a random effect, while $r_0 U_0$ is added to the survival model.

Fit Statistics

-2 Log Likelihood	2800.9
AIC (smaller is better)	2818.9
BIC (smaller is better)	2838.0

Table 5.57: Fit statistics for a joint piecewise viral load and informative drop-out model with $W_1(t) = U_0$ and $W_2(t) = r_0 U_0$

Parameter Estimates

Parameter	Estimate	Std Err	DF	t Value	Pr >|t |	Lower	Upper
Surv Intercept	9.3926	0.5402	61	17.39	<.0001	8.3125	10.4727
Slope 0 to 2	2.6590	0.06165	61	43.13	<.0001	2.5357	2.7823
Slope 2 to 4	-0.4038	0.07362	61	-5.49	<.0001	-0.5511	-0.2566
Slope 4 to 8	0.03626	0.03122	61	1.16	0.2500	-0.02617	0.09868
Slope 8 to 12	-0.08462	0.01703	61	-4.97	<.0001	-0.1187	-0.05056
Slope 12+	-0.00110	0.000340	61	-3.24	0.0020	-0.00178	-0.00042
Residual	0.3046	0.01159	61	26.28	<.0001	0.2814	0.3278
r0	-2.0413	0.6637	61	-3.08	0.0031	-3.3684	-0.7142
v11	0.4704	0.08712	61	5.40	<.0001	0.2962	0.6446

Table 5.58: Parameter estimates for a joint piecewise viral load and informative drop-out model with $W_1(t) = U_0$ and $W_2(t) = r_0 U_0$

The results in Table 5.58 show that the common joint effect captured through r_0 is statistically significant from zero (p=0.0031) indicating that there is indeed a significant joint effect between the longitudinal and survival sub-models. It is also interesting to note that the piecewise slopes are more precisely estimated (smaller standard errors) under the joint model compared to that when $W_2(t) = 0$.

Model 3: $W_1(t) = U_1 t$ and $W_2(t) = r_1 U_1$, piecewise

Now, instead of the random intercept, a random slope is added to the longitudinal model which is common to all piecewise effects. The survival model is then linked to the longitudinal model through this random slope.

Fit Statistics

-2 Log Likelihood	3263.9
AIC (smaller is better)	3263.9
BIC (smaller is better)	3301.0

Table 5.59: Fit statistics for a joint piecewise viral load and informative drop-out model with $W_1(t) = U_1 t$ and $W_2(t) = r_1 U_1$

The results in Table 5.60 show that the common joint effect r_1 is statistically significant from zero (p=0.0001) indicating that there is indeed a significant joint effect between the longitudinal and survival sub-models via the common random slope.

Parameter Estimates

Parameter	Estimate	Std Err	DF	t Value	Pr >|t |	Lower	Upper
Surv Intercept	9.2621	0.4766	61	19.43	<.0001	8.3091	10.2151
Slope 0 to 2	2.7072	0.06383	61	42.41	<.0001	2.5796	2.8348
Slope 2 to 4	-0.4566	0.08571	61	-5.33	<.0001	-0.6280	-0.2852
Slope 4 to 8	0.05574	0.03592	61	1.55	0.1259	-0.01609	0.1276
Slope 8 to 12	-0.09672	0.02001	61	-4.83	<.0001	-0.1367	-0.05672
Slope 12+	-0.00101	0.001842	61	-0.55	0.5837	-0.00470	0.002668
Residual	0.4222	0.01612	61	26.19	<.0001	0.3900	0.4545
r1	-97.9038	24.0116	61	-4.08	0.0001	-145.92	-49.8897
v22	0.000184	0.000038	61	4.83	<.0001	0.000108	0.000261

Table 5.60: Parameter estimates for a joint piecewise viral load and informative drop-out model with $W_1(t) = U_1 t$ and $W_2(t) = r_1 U_1$

Model 4: $W_1(t) = U_0 + U_1 t$ and $W_2(t) = r_0 U_0$, piecewise

This specification places a common random slope effect for all the piecewise slopes and the longitudinal and survival models are linked through a random intercept effect.

Fit Statistics

-2 Log Likelihood	2607.7
AIC (smaller is better)	2629.7
BIC (smaller is better)	2653.1

Table 5.61: Fit statistics for a joint piecewise viral load and informative drop-out model with $W_1(t) = U_0 + U_1 t$ and $W_2(t) = r_0 U_0$

In Table 5.62, the joint effect $r0$ is statistically significant from zero (p=0.0083) thus there is a significant joint effect between the longitudinal and survival sub-models via the common random intercept. Also important to note is that the between-subject variance in the common random slope between the piecewise effects, $v22$ is significantly different from zero (p=0.0002).

Parameter Estimates

Parameter	Estimate	Std Err	DF	t Value	Pr >|t |	Lower	Upper
Surv Intercept	9.3446	0.5521	60	16.92	<.0001	8.2401	10.4490
Slope 0 to 2	2.6752	0.05591	60	47.85	<.0001	2.5634	2.7871
Slope 2 to 4	-0.4014	0.06640	60	-6.05	<.0001	-0.5343	-0.2686
Slope 4 to 8	0.03334	0.02842	60	1.17	0.2454	-0.02351	0.09019
Slope 8 to 12	-0.08725	0.01550	60	-5.63	<.0001	-0.1183	-0.05625
Slope 12+	-0.00177	0.000944	60	-1.87	0.0657	-0.00366	0.000119
Residual	0.2436	0.009515	60	25.60	<.0001	0.2246	0.2627
r0	-2.1732	0.7959	60	-2.73	0.0083	-3.7651	-0.5812
v11	0.4173	0.07955	60	5.25	<.0001	0.2582	0.5765
v12	0.000188	0.000614	60	0.31	0.7611	-0.00104	0.001416
v22	0.000040	0.000010	60	3.99	0.0002	0.000020	0.000061

Table 5.62: Parameter estimates for a joint piecewise viral load and informative drop-out model with $W_1(t) = U_0 + U_1 t$ and $W_2(t) = r_0 U_0$

Model 5: $W_1(t) = U_0 + U_1 t$ **and** $W_2(t) = r_0 U_0 + r_1 U_1$

As with modelling time as a linear slope, this specification of joining the longitudinal and survival sub-models through both a random intercept and slope failed to converge properly and was not used.

Summary

Joint Model	Model 1			Model 2			Model 3			Model 4		
Parameter	Estimate	Std Err	Pr >|t |	Estimate	Std Err	Pr >|t |	Estimate	Std Err	Pr >|t |	Estimate	Std Err	Pr >|t |
Surv Intercept	8.6230	0.3015	<.0001	9.3926	0.5402	<.0001	9.2621	0.4766	<.0001	9.3446	0.5521	<.0001
Slope 0 to 2	2.7072	0.06384	<.0001	2.6590	0.06165	<.0001	2.7072	0.06383	<.0001	2.6752	0.05591	<.0001
Slope 2 to 4	-0.4562	0.085732	<.0001	-0.4038	0.07362	<.0001	-0.4566	0.08571	<.0001	-0.4014	0.06640	<.0001
Slope 4 to 8	0.05552	0.03593	0.1275	0.03626	0.03122	0.2500	0.05574	0.03592	0.1259	0.03334	0.02842	0.2454
Slope 8 to 12	-0.09631	0.02001	<.0001	-0.08462	0.01703	<.0001	-0.09672	0.02001	<.0001	-0.08725	0.01550	<.0001
Slope 12+	-0.00111	0.001844	0.5490	-0.00110	0.000340	0.0020	-0.00101	0.001842	0.5837	-0.00177	0.000944	0.0657
Residual	0.4223	0.01614	<.0001	0.3046	0.01159	<.0001	0.4222	0.01612	<.0001	0.2436	0.009515	<.0001
v11	-	-	-	0.4704	0.08712	<.0001	-	-	-	0.4173	0.07955	<.0001
v12	-	-	-	-	-	-	-	-	-	0.000188	0.000614	0.7611
v22	0.000184	0.000038	<.0001	-	-	-	0.000184	0.000038	<.0001	0.000040	0.000010	0.0002
r0	-	-	-	-2.0413	0.6637	0.0031	-	-	-	-2.1732	0.7959	0.0083
r1	-	-	-	-	-	-	-97.9038	24.0116	0.0001	-	-	-
-2 Log Likelihood	3280.6			2800.9			3263.9			2607.7		
AIC	3296.6			2818.9			3263.9			2629.7		
BIC	3313.6			2838.0			3301.0			2653.1		

Table 5.63: Summary of joint modelling results for a piecewise viral load model

Model 1: $W_1(t) = U_1 t$; $W_2(t) = 0$

Model 2: $W_1(t) = U_0$; $W_2(t) = r_0 U_0$

Model 3: $W_1(t) = U_1 t$; $W_2(t) = r_1 U_1$

Model 4: $W_1(t) = U_0 + U_1 t$; $W_2(t) = r_0 U_0$

Model 4 provides the best fit for a joint model combining the longitudinal piecewise model for viral load to the exponential model modelling the informative drop-out. Thus the best specified association between the two models was $W_1(t) = U_0 + U_1 t$ and $W_2(t) = r_0 U_0$, where they are linked via a random intercept. A common random slope was also applied to the piecewise random effects and this also provided a better fit, compared to Model 2 which also links the models through the random intercept.

In Section 5.3.2, which looked at modelling the piecewise linear mixed model to viral load, SAS procedure `PROC MIXED` was used to fit the models. To specify the slopes as random effects would have required to put these variables after the `RANDOM`, however this assumes that each of the five slopes have their own random slope and have their own subject-specific estimates and the model would try to calculate a variance-covariance matrix for all five slopes. This proved impossible given the data used and thus could not be done. However, the `NLMIXED` procedure in SAS allows one to fit a generalised model and in this case, a common random slope estimate was specified for all five slopes, taking into account the between-subject variation that occurs over time.

In all the joint models Model 2, 3 and 4, the intercept for the survival model is higher in all the three models compared to Model 1, as well as the univariate marginal model described in Section 5.3.2. The effect estimates do not change much in magnitude, however the standard errors in Model 4 are smaller than the other models. The joint effect estimates, r_0 and r_1 are statistically significant in all models where used. The different joint models are represented graphically in Figures 5.8 and 5.9. Figure 5.8 looks at the first two years of HIV infection, while Figure 5.9 looks more closely what happens between 4 and 5 log copies/mℓ. Thus Model 4, having both random intercept and slope effects for the longitudinal measurements process and linking it to the informative drop-out process via the random intercept effect best depicts the viral load evolution better than the other three candidate models.

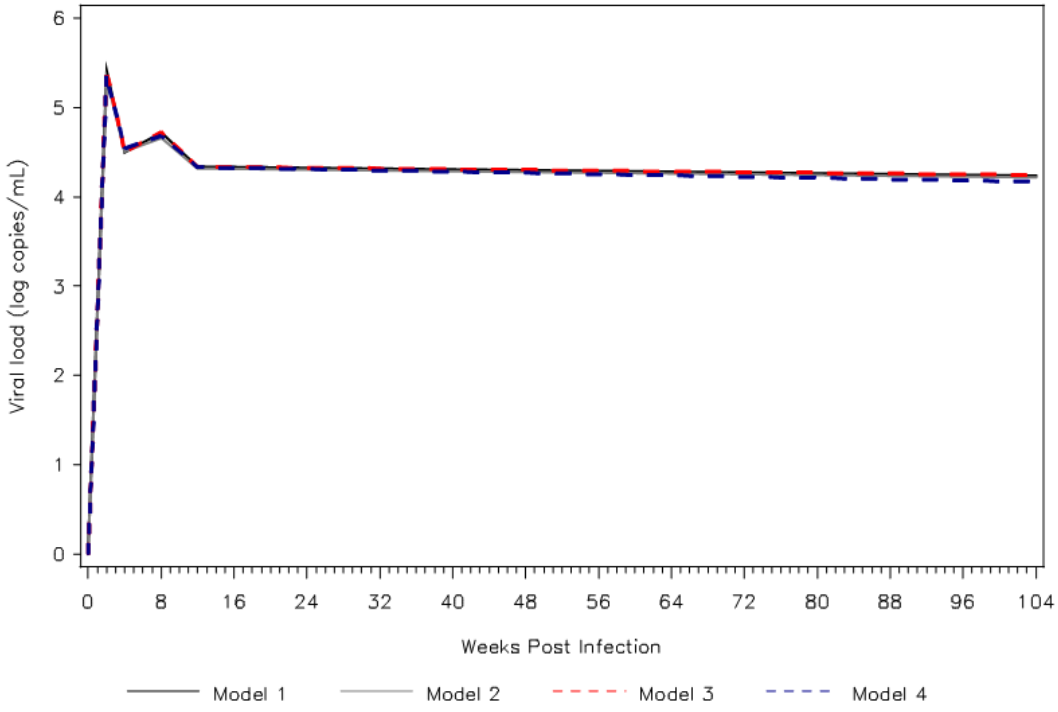

Figure 5.8: Different joint models for modelling viral load with time as piecewise effects

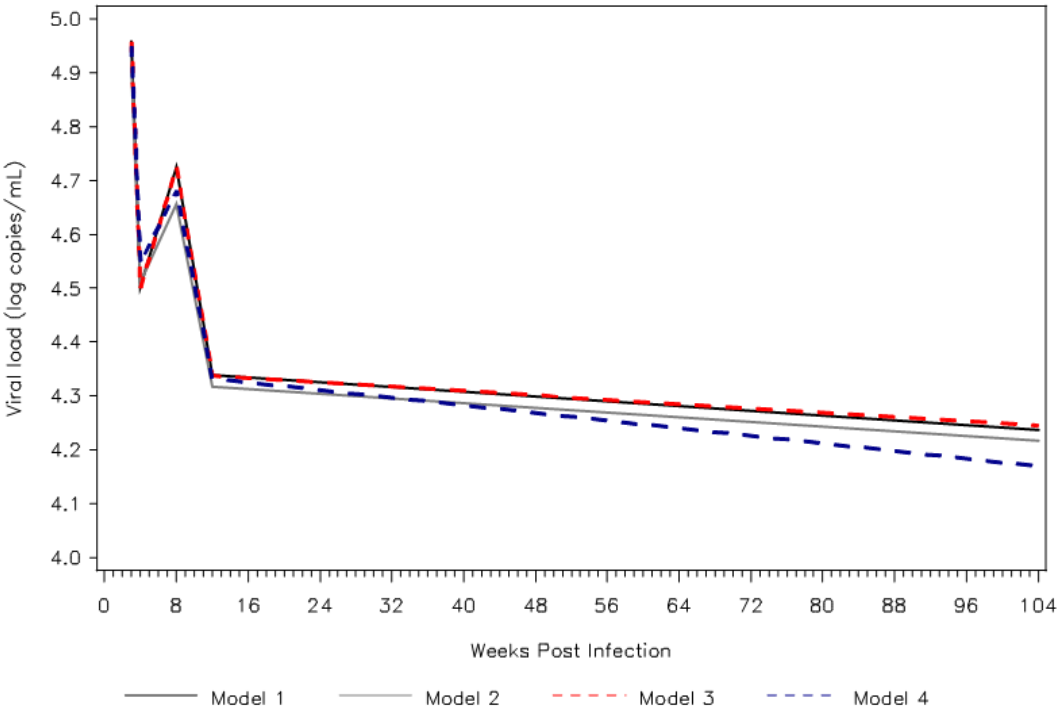

Figure 5.9: Different joint models for modelling viral load with time as piecewise effects, between 4 and 5 log copies/mℓ

Chapter 6

Conclusion and Future Work

The problem of understanding the pathogenesis of HIV/AIDS is still a very active research area, particularly now that several treatment and mitigating strategies are in consideration. Statistical methods to model and quantify the evolution of two key HIV markers during the acute infection stage of the disease were presented in this work. Because of the variability in trajectories over time in both CD4+ count and viral load during the first few weeks post infection, a piecewise linear mixed effects model was applied and shown to provide a better fit to the data in describing the biomarkers.

Fitting a simple linear regression would not be adequate and even fitting a quadratic or cubic term for weeks post infection would not be meaningful. The piecewise linear effects approach allowed for the quantification of the rate of change in the surrogate HIV markers in the initial two weeks following HIV infection, showing the peak in viral load and the decline in CD4+ count. Thereafter, in the subsequent fortnight it could be shown, that viral load decreases as the body starts to control the virus. However, CD4+ count does not recover after the initial drop and remains low.

This method of modelling the HIV markers would be useful when analysing the effect of pre-exposure to antiretroviral treatment before acquiring HIV, as would be the case in those individuals using microbicides or pre-exposure prophylaxis (PrEP). Although these methods are designed to protect against HIV, their effectiveness is still being evaluated in clinical trials. If one such subject on the active arm did indeed acquire HIV, then determining the effect of pre-exposure to antiretroviral treatment on viral load replication would be important. For instance, if it is found that those who have been pre-exposed to antiretroviral treatment have a lower peak viral load during acute infection and a sustained lower viral load, then this would be a useful finding. Similarly the effects

of intervention on CD4+ count could also be studied. Understanding and quantifying the effect of microbicides or PrEP on viral load and CD4+ count could help to better the understanding of these preventative methods. The piecewise linear effects model would also be useful when studying genetic factors which affect how someone responds to HIV infection as well as disease progression over time. For example, analysis done on the same cohort by Sewram et al. (2009) and Naicker et al. (2009) can use this model to quantify the effect of the different genotypes on the evolution of viral load and CD4+ count.

As was shown in Section 5.4, accounting for left-censoring was not necessary and did not provide a better fit to the model for this cohort. This was due to the fact that only a small proportion of viral load measurements fell below the lower limit of detection. However, the maximum-likelihood approach for left-censoring has been shown to produce less biased estimates compared to crude methods such as using the censoring limit (Jacqmin-Gadda et al., 2000). If one is modelling viral load in a cohort where a moderate proportion of the measurements fall below a certain detection limit, as would be the case in individuals on HAART, then the methods for left-censoring should be explored.

In this thesis, joint modelling was used to combine the longitudinal measurement process and time to HAART initiation based on CD4+ levels. Joint modelling was shown to provide a better fit to the data. Comparing the piecewise linear mixed effects model which accounted for informative drop-out with one that did not, showed that CD4+ count decline was indeed underestimated in the latter model, especially in the interval past 12 weeks post infection. Viral load was not affected as much when adjusting for informative drop-out, despite the correlation with CD4+ count. We note however that CD4+ count is a direct determinant of whether someone drops out and gets initiated on HAART. Joint modelling seems to be a very important method to take account of incomplete data resulting from informative drop-out. Overall the current study has helped to understand the acute infection stage of HIV better for individuals infected with HIV-1 subtype C in a South African cohort of women.

The measurement longitudinal data process due to CD4+ count and viral load were modelled univariately, where only one dependent variable was considered at a time. However, when there are two response variables observed jointly with each other then bivariate linear mixed modelling may be the most appropriate approach because it can take account of possible dependence between these response variables. Multivariate longitudinal data analysis is an area attracting renewed focus in biostatistics. Such an approach will be the subject of future studies.

References

ABDOOL KARIM, S. S. & ABDOOL KARIM, Q. (2005) HIV/AIDS in South Africa, Cambridge University Press.

ALLISON, P. D. (1995) Survival Analysis Using SAS: A Practical Guide, SAS Institute Inc., Cary, NC, USA.

ANDERSON, R. M. & MAY, R. M. (1991) Infectious diseases of humans: dynamics and control, Oxford University Press.

AVERTING HIV AND AIDS (2009) History and science of HIV & AIDS, accessed 24 January 2009 (Available at http://www.avert.org/history-science.htm)

BAETEN, J. M., CHOHAN, B., LAVREYS, L., CHOHAN, V., MCCLELLAND, R. S., CERTAIN, L., MANDALIYA, K., JAOKO, W. & OVERBAUGH, J. (2007) HIV-1 subtype D infection is associated with faster disease progression than subtype A in spite of similar plasma HIV-1 loads. J Infect Dis, 195, 1177-80.

BARRE-SINOUSSI, F., CHERMANN, J. C., REY, F., NUGEYRE, M. T., CHAMARET, S., GRUEST, J., DAUGUET, C., AXLER-BLIN, C., VEZINET-BRUN, F., ROUZIOUX, C., ROZENBAUM, W. & MONTAGNIER, L. (1983) Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). Science, 220, 868-71.

BLANCOU, P., VARTANIAN, J. P., CHRISTOPHERSON, C., CHENCINER, N., BASILICO, C., KWOK, S. & WAIN-HOBSON, S. (2001) Polio vaccine samples not linked to AIDS. Nature, 410, 1045-6.

CARROLL, R. J., RUPPERT, D., STEFANSKI, L. A. & CRAINICEANU, C. M. (2006) Measurement Error in Nonlinear Models: A Modern Perspective, Chapman & Hall.

CDC, Centers for Disease Control and Prevention (1981). Pneumocystis pneumonia–Los Angeles. MMWR Morb Mortal Wkly Rep 30, 250-2. COHEN, J. (2000) The Hunt for the Origin of AIDS. The Atlantic.

COX, D. R. (1972) Regression Models and Life-Tables. Journal of the Royal Statistical Society Series B-Statistical Methodology, 34, 187-200.

DEFRANCO, A. L., LOCKSLEY, R. M. & ROBERTSON, M. (2007) Immunity: The Immune Response to Infectious and Inflammatory Disease, USA, Oxford University Press.

DIGGLE, P., HEAGERTY, P., LIANG, K.-Y. & ZEGER, S. (2002) Analysis of longitudinal data, USA, Oxford University Press.

DU BOIS, R. M., BRANTHWAITE, M. A., MIKHAIL, J. R. & BATTEN, J. C. (1981) Primary Pneumocystis carinii and cytomegalovirus infections. Lancet, 2, 1339.

EUROPEAN STUDY GROUP ON HETEROSEXUAL TRANSMISSION OF HIV (1992) Comparison of female to male and male to female transmission of HIV in 563 stable couples. BMJ, 304, 809-13.

FERRON, J., DAILEY, R. & YI, Q. (2002) Effects of misspecifying the first-level error structure in two-level models of change. Multivariate Behavioral Research, 37, 379-403.

FLEMING, D. T. & WASSERHEIT, J. N. (1999) From epidemiological synergy to public health policy and practice: the contribution of other sexually transmitted diseases to sexual transmission of HIV infection. Sex Transm Infect, 75, 3-17.

GAO, F., BAILES, E., ROBERTSON, D. L., CHEN, Y., RODENBURG, C. M., MICHAEL, S. F., CUMMINS, L. B., ARTHUR, L. O., PEETERS, M., SHAW, G. M., SHARP, P. M. & HAHN, B. H. (1999) Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. Nature, 397, 436-41.

GOUJARD, C., BONAREK, M., MEYER, L., BONNET, F., CHAIX, M. L., DEVEAU, C., SINET, M., GALIMAND, J., DELFRAISSY, J. F., VENET, A., ROUZIOUX, C. & MORLAT, P. (2006) CD4 cell count and HIV DNA level are independent predictors of disease progression after primary HIV type 1 infection in untreated patients. Clin Infect Dis, 42, 709-15.

GUO, X. & CARLIN, B. P. (2004) Separate and joint modeling of longitudinal and event time data using standard computer packages. American Statistician, 58, 16-24.

HENDERSON, R., DIGGLE, P. & DOBSON, A. (2000) Joint modelling of longitudinal measurements and event time data. Biostatistics, 1, 465-80.

HOOPER, E. (1999) The river: a journey to the source of HIV and AIDS, Boston, Little Brown and Company.

HUANG, Y. X., LIU, D. C. & WU, H. L. (2006) Hierarchical Bayesian methods for estimation of parameters in a longitudinal HIV dynamic system. Biometrics, 62, 413-423.

HYMES, K. B., CHEUNG, T., GREENE, J. B., PROSE, N. S., MARCUS, A., BALLARD, H., WILLIAM, D. C. & LAUBENSTEIN, L. J. (1981) Kaposi's sarcoma in homosexual men - a report of eight cases. Lancet, 2, 598-600.

JACQMIN-GADDA, H., THIEBAUT, R., CHENE, G. & COMMENGES, D. (2000) Analysis of left-censored longitudinal data with application to viral load in HIV infection. Biostatistics, 1, 355-68.

KANKI, P. J., HAMEL, D. J., SANKALE, J. L., HSIEH, C., THIOR, I., BARIN, F., WOODCOCK, S. A., GUEYE-NDIAYE, A., ZHANG, E., MONTANO, M., SIBY, T., MARLINK, R., I, N. D., ESSEX, M. E. & S, M. B. (1999) Human Immunodeficiency Virus type 1 subtypes differ in disease progression. J Infect Dis, 179, 68-73.

KAPLAN, E. H. & BROOKMEYER, R. (1999) Snapshot Estimators of Recent HIV Incidence Rates. Operations Research, 47, 29-37.

KENWARD, M. G. & MOLENBERGHS, G. (1999) Parametric models for incomplete continuous and categorical longitudinal data. Stat Methods Med Res, 8, 51-83.

KESELMAN, H. J., ALGINA, J., KOWALCHUK, R. K. & WOLFINGER, R. D. (1998) A comparison of two approaches for selecting covariance structures in the analysis of repeated measurements. Communications in Statistics-Simulation and Computation, 27, 591-604.

KINCAID, C. (2005) Guidelines for selecting the covariance structure in mixed model analysis. SUGI 30 Proceedings, Paper 198-30.

KUIKEN, C. L., FOLEY, B., HAHN, B., KORBER, B., MCCUTCHAN, F., MARX, P. A., MELLORS, J. W., MULLINS, J. I. & SODROSKI, J. A. W., S. (1999) Human retroviruses and AIDS 1999: a compilation and analysis of nucleic acid and amino acid sequences. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, USA.

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.

LITTELL, R. C., HENRY, P. R. & AMMERMAN, C. B. (1998) Statistical analysis of repeated measures data using SAS procedures. J Anim Sci, 76, 1216-31.

LYLES, R. H., MUNOZ, A., YAMASHITA, T. E., BAZMI, H., DETELS, R., RINALDO, C. R., MARGOLICK, J. B., PHAIR, J. P. & MELLORS, J. W. (2000) Natural history of human immunodeficiency virus type 1 viremia after seroconversion and proximal to AIDS in a large cohort of homosexual men. Multicenter AIDS Cohort Study. J Infect Dis, 181, 872-80.

MARX, J. L. (1984) Strong new candidate for AIDS agent. Science, 224, 475-7.

MCCULLOCH, C. E., SEARLE, S. R. & NEUHAUS, J. M. (2008) Generalized, Linear, and Mixed Models, Hoboken, New Jersey, John Wiley & Sons, Inc.

MOLENBERGHS, G. & KENWARD, M. (2007) Missing Data in Clinical Studies, Wiley.

MOLENBERGHS, G. & VERBEKE, G. (2005) Models for Discrete Longitudinal Data, Springer

NAICKER, D. D., WERNER, L., KORMUTH, E., PASSMORE, J. A., MLISANA, K., KARIM, S. A. & NDUNG'U, T. (2009) Interleukin-10 promoter polymorphisms influence HIV-1 susceptibility and primary HIV-1 pathogenesis. J Infect Dis, 200, 448-52.

O'BRIEN, T. R., ROSENBERG, P. S., YELLIN, F. & GOEDERT, J. J. (1998) Longitudinal HIV-1 RNA levels in a cohort of homosexual men. Journal of Acquired Immune Deficiency Syndromes, 18, 155-161.

OSMANOV, S., PATTOU, C., WALKER, N., SCHWARDLANDER, B. & ESPARZA, J. (2002) Estimated global distribution and regional spread of HIV-1 genetic subtypes in the year 2000. J Acquir Immune Defic Syndr, 29, 184-90.

PADIAN, N. S., SHIBOSKI, S. C. & JEWELL, N. P. (1991) Female-to-male transmission of human immunodeficiency virus. JAMA, 266, 1664-7.

PANEL ON ANTIRETROVIRAL GUIDELINES FOR ADULTS AND ADOLESCENTS (2008) Guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents. Washington, DC: Department of Health and Human Services, accessed 24 January 2009 (Available at http://aidsinfo.nih.gov/contentfiles/AdultandAdolescentGL.pdf)

PANTAZIS, N. & TOULOUMI, G. (2005) Bivariate modelling of longitudinal measurements of two human immunodeficiency type 1 disease progression markers in the presence of informative drop-outs. Journal of the Royal Statistical Society Series C-Applied Statistics, 54, 405-423.

SEWRAM, S., SINGH, R., KORMUTH, E., WERNER, L., MLISANA, K., KARIM, S. S. & NDUNG'U, T. (2009) Human TRIM5alpha expression levels and reduced susceptibility to HIV-1 infection. J Infect Dis, 199, 1657-63.

THIEBAUT, R. & JACQMIN-GADDA, H. (2004) Mixed models for longitudinal left-censored repeated measures. Comput Methods Programs Biomed, 74, 255-60.

THIEBAUT, R., JACQMIN-GADDA, H., BABIKER, A. & COMMENGES, D. (2005) Joint modelling of bivariate longitudinal data with informative dropout and left-censoring, with application to the evolution of CD4+ cell count and HIV RNA viral load in response to treatment of

HIV infection. Stat Med, 24, 65-82.

UNAIDS, Joint United Nations Programme on HIV/AIDS (2008). Report on the global AIDS epidemic 2008.

VAN HARMELEN, J. H., VAN DER RYST, E., LOUBSER, A. S., YORK, D., MADURAI, S., LYONS, S., WOOD, R. & WILLIAMSON, C. (1999) A predominantly HIV type 1 subtype C-restricted epidemic in South African urban populations. AIDS Res Hum Retroviruses, 15, 395-8.

VAN LOGGERENBERG, F., MLISANA, K., WILLIAMSON, C., AULD, S. C., MORRIS, L., GRAY, C. M., ABDOOL-KARIM, Q., GROBLER, A., BARNABAS, N., IRIOGBE, I. & ABDOOL-KARIM, S. S. (2008) Establishing a Cohort at High Risk of HIV Infection in South Africa: Challenges and Experiences of the CAPRISA 002 Acute Infection Study. PLoS One, 3(4):e1954.

VERBEKE, G. & MOLENBERGHS, G. (2000) Linear Mixed Models for Longitudinal Data, New York, Springer-Verlag.

WORLD HEALTH ORGANIZATION (WHO) (1990) Interim proposal for a WHO Staging System for HIV infection and Disease. Wkly Epidemiol Rec, 65, 221-4.

Appendix: SAS Code

Linear Mixed Models

```
 /************************************ CD4 COUNT ***************************/

/*FITTING MARGINAL LINEAR MIXED MODEL ON CD4 COUNT WITH WEEKS POST INFECTION AS A LINEAR
EFFECT (MODEL IN TABLE 5.2)*/
proc mixed data=mscdata method=reml covtest empirical;
   class pid week2;
   model sqrtcd4=week/s;
   repeated week2/subject=pid type=sp(pow)(week);
   title 'CD4 = week, marginal model SP(POW)';
run;

/*FITTING A RANDOM INTERCEPT LINEAR MIXED MODEL ON CD4 COUNT WITH WEEKS POST INFECTION
AS A LINEAR EFFECT (MODEL IN TABLE 5.5)*/
proc mixed data=mscdata method=reml covtest empirical;
   class pid week2;
   model sqrtcd4=week/s;
   random intercept/s subject=pid type=un;
   repeated week2/subject=pid type=sp(pow)(week);
   title 'CD4 = week, repeated SP(POW), random intercept';
run;

/*FITTING A RANDOM INTERCEPT AND SLOPE LINEAR MIXED MODEL ON CD4 COUNT WITH WEEKS POST
INFECTION AS A LINEAR EFFECT (MODEL IN TABLE 5.9)*/
proc mixed data=mscdata method=reml empirical covtest;
   class pid week2;
   model sqrtcd4=week/s;
   random intercept week/s subject=pid type=un;
   repeated week2/subject=pid type=sp(pow)(week);
   title 'CD4 = week, repeated SP(POW), random intercept and slope';
run;

/************************************ VIRAL LOAD ***************************/

/*FITTING MARGINAL LINEAR MIXED NO-INTERCEPT MODEL ON VIRAL LOAD WITH WEEKS POST INFECTION
AS A LINEAR EFFECT (MODEL IN TABLE 5.14)*/
proc mixed data=mscdata method=reml empirical covtest;
   class pid week2;
```

```
   model logvl=week/noint s;
   repeated week2/subject=pid type=cs;
   title 'LogVL = week, marginal model CS';
run;

/*FITTING MARGINAL LINEAR MIXED NO-INTERCEPT RANDOM SLOPE MODEL ON VIRAL LOAD WITH
WEEKS POST INFECTION AS A LINEAR EFFECT (AND RANDOM EFFECT) (MODEL IN TABLE 5.17)*/
proc mixed data=mscdata method=reml empirical covtest;
   class pid week2;
   model logvl=week/noint s;
   random week/s subject=pid type=un;
   repeated week2/subject=pid type=cs;
   title 'logvl = week, repeated CS, random intercept';
run;
```

Piecewise Linear Mixed Models

```
 /************************************ CD4 COUNT ***************************/

/*FITTING MARGINAL LINEAR MIXED MODEL ON CD4 COUNT WITH WEEKS POST INFECTION AS PIECEWISE
LINEAR EFFECTS (MODEL IN TABLE 5.21)*/
proc mixed data=mscdata method=reml empirical covtest;
   class pid week2;
   model sqrtcd4=slope0_2 slope2_4 slope4_8 slope8_12 slope12_/corrb s;
   repeated week2/subject=pid type=cs;
   title 'CD4 piecewise repeated CS, marginal';
run;

/*FITTING MARGINAL LINEAR MIXED RANDOM INTERCEPT MODEL ON CD4 COUNT WITH WEEKS POST
INFECTION AS PIECEWISE LINEAR EFFECTS (MODEL IN TABLE 5.24)*/
proc mixed data=mscdata method=reml empirical covtest;
   class pid week2;
   model sqrtcd4= slope0_2 slope2_4 slope4_8 slope8_12 slope12_/corrb s;
   random intercept/s subject=pid type=un;
   repeated week2/subject=pid type=sp(pow)(week);
   title 'CD4 piecewise repeated SP(POW), random intercept';
run;

/************************************ VIRAL LOAD ***************************/

/*FITTING MARGINAL LINEAR MIXED NO-INTERCEPT MODEL ON VIRAL LOAD WITH WEEKS POST INFECTION
AS PIECEWISE LINEAR EFFECTS (MODEL IN TABLE 5.29)*/
proc mixed data=mscdata method=reml empirical covtest;
   class pid week2;
   model logvl= slope0_2 slope2_4 slope4_8 slope8_12 slope12_/noint corrb s;
   repeated week2/subject=pid type=cs;
   title 'LogVL piecewise repeated CS, marginal';
run;
```

Left-censoring Of Viral Load

```
 /********NLMIXED MODEL - VIRAL LOAD WITH LEFT-CENSORING**********/

/*FITTING MARGINAL LINEAR MIXED NO-INTERCEPT MODEL ON VIRAL LOAD WITH WEEKS POST INFECTION
AS PIECEWISE LINEAR EFFECTS TAKING INTO ACCOUNT LEFT-CENSORING (MODEL IN TABLE 5.32)*/
proc nlmixed data=mscdata;
   parms b0_2=2.66 b2_4=-0.40 b4_8=0.03 b8_12=-0.08 b12_=-0.00123 sigsqe=0.29;
   pi=2*arsin(1);
   mu=b0_2*slope0_2 + b2_4*slope2_4 + b4_8*slope4_8 + b8_12*slope8_12 + b12_*slope12_
+ participant;
   if vlOBS eq 1 then ll=(1/(sqrt(2*pi*sigsqe)))*exp(-(logvl-mu)**2/(2*sigsqe));
   if vlobs eq 0 then ll=probnorm((logvl-mu)/sqrt(sigsqe));
   L=log(ll);
   model logvl ~ general(L);
   random participant ~ normal([0],[pvar]) subject=pid out=temp;
   title 'VL piecewise model with LEFT CENSORING';
run;
```

Joint Modelling

```
 /************************************ CD4 COUNT *************************/

/*FITTING JOINT MODEL ON CD4 COUNT WITH WEEKS POST INFECTION FITTED AS A LINEAR EFFECT
(MODEL IN TABLE 5.35)*/

title 'Joint Model - CD4 count - W1(t) = U0 + U1t / W2(t)=0';
proc nlmixed data=mscdata;
   parms bs0=8.75 bl0=23.0 bl1=-0.02 residual=7.8;

   /*Survival Exponential model on time to ARV initiation*/
   if (lastvisit) then do;
      linpsurv = bs0;
      alpha = exp(-linpsurv);
      G_t = exp(-alpha*days);
      g = alpha * G_t;
      llsurv = (ARV=1)*log(g) + (ARV=0)*log(G_t);
   end; else llsurv = 0;

   /*Longitudinal model for CD4 count*/
   mu = (bl0 + u0) + (bl1 + u1)*week;
   pi = 2*arsin(1);
   ll = (1/(sqrt(2*pi*residual)))*exp(-(sqrtcd4-mu)**2/(2*residual));
   lllong = log(ll);

   model lastvisit ~ general(lllong + llsurv);
   random u0 u1 ~ normal([0, 0],[v11,v12,v22]) subject=pid;
run;

/*FITTING JOINT MODEL ON CD4 COUNT WITH WEEKS POST INFECTION FITTED AS A LINEAR EFFECT
(MODEL IN TABLE 5.37)*/
```

```
title 'Joint Model - CD4 count - W1(t) = U0 + U1t / W2(t) = r0U0';
proc nlmixed data=mscdata;
   parms bs0=8.75 bl0=23.0 bl1=-0.02 residual=7.8;

   /*Survival Exponential model on time to ARV initiation*/
   if (lastvisit) then do;
      linpsurv = bs0 + r0*U0;
      alpha = exp(-linpsurv);
      G_t = exp(-alpha*days);
      g = alpha*G_t;
      llsurv = (ARV=1)*log(g) + (ARV=0)*log(G_t);
   end; else llsurv = 0;

   /*Longitudinal model for CD4 count*/
   mu = (bl0 + u0) + (bl1 + u1)*week;
   pi = 2*arsin(1);
   ll = (1/(sqrt(2*pi*residual)))*exp(-(sqrtcd4-mu)**2/(2*residual));
   lllong = log(ll);

   model lastvisit ~ general(lllong + llsurv);
   random u0 u1 ~ normal([0, 0],[v11,v12,v22]) subject=pid;
run;

/*FITTING JOINT MODEL ON CD4 COUNT WITH WEEKS POST INFECTION FITTED AS A LINEAR EFFECT
(MODEL IN TABLE 5.39)*/

title 'Joint Model - CD4 count - W1(t) = U0 + U1t / W2(t) = r0U0 + r1U1';
proc nlmixed data=mscdata;
   parms bs0=8.75 bl0=23.0 bl1=-0.02 residual=7.8;

   /*Survival Exponential model on time to ARV initiation*/
   if (lastvisit) then do;
      linpsurv = bs0 + r0*u0 + r1*u1;
      alpha = exp(-linpsurv);
      G_t = exp(-alpha*days);
      g = alpha*G_t;
      llsurv = (ARV=1)*log(g) + (ARV=0)*log(G_t);
   end; else llsurv=0;

   /*Longitudinal model for CD4 count*/
   mu = (bl0 + u0) + (bl1 + u1)*week;
   pi = 2*arsin(1);
   ll = (1/(sqrt(2*pi*residual)))*exp(-(sqrtcd4-mu)**2/(2*residual));
   lllong = log(ll);

   model lastvisit ~ general(lllong + llsurv);
   random u0 u1 ~ normal([0, 0],[v11,v12,v22]) subject=pid;
run;

/*FITTING JOINT MODEL ON CD4 COUNT WITH WEEKS POST INFECTION FITTED AS PIECEWISE LINEAR
EFFECTS (MODEL IN TABLE 5.42)*/
```

```
title 'Joint Model - Piecewise CD4 count - W1(t) = U0/ W2(t) = 0';
proc nlmixed data=mscdata;
   parms bs0=8.75 bl0=31.16 bl02=-4.03 bl24=-0.129 bl48=0.03242 bl812=-0.2339 bl12=-0.01504
         residual=7.8;

   /*Survival Exponential model on time to ARV initiation*/
   if (lastvisit) then do;
      linpsurv = bs0;
      alpha = exp(-linpsurv);
      G_t = exp(-alpha*days);
      g = alpha*G_t;
      llsurv = (ARV=1)*log(g) + (ARV=0)*log(G_t);
   end; else llsurv=0;

   /*Longitudinal model for CD4 count*/
   mu = (bl0 + u0) + (bl02)*slope0_2 + (bl24)*slope2_4 + (bl48)*slope4_8 + (bl812)*slope8_12
         + (bl12)*slope12_;
   pi = 2*arsin(1);
   ll = (1/(sqrt(2*pi*residual)))*exp(-(sqrtcd4-mu)**2/(2*residual));
   lllong = log(ll);

   model lastvisit ~ general(lllong + llsurv);
   random u0 ~ normal([0],[v11]) subject=pid;
run;

/*FITTING JOINT MODEL ON CD4 COUNT WITH WEEKS POST INFECTION FITTED AS PIECEWISE LINEAR
EFFECTS (MODEL IN TABLE 5.44)*/

title 'Joint Model - Piecewise CD4 count - W1(t) = U0/ W2(t) = r0U0';
proc nlmixed data=mscdata;
   parms bs0=8.75 bl0=31.16 bl02=-4.03 bl24=-0.129 bl48=0.03242 bl812=-0.2339 bl12=-0.01504
         residual=7.8;

   /*Survival Exponential model on time to ARV initiation*/
   if (lastvisit) then do;
      linpsurv = bs0 + r0*U0;
      alpha = exp(-linpsurv);
      G_t = exp(-alpha*days);
      g = alpha*G_t;
      llsurv = (ARV=1)*log(g) + (ARV=0)*log(G_t);
   end; else llsurv=0;

   /*Longitudinal model for CD4 count*/
   mu = (bl0 + u0) + (bl02)*slope0_2 + (bl24)*slope2_4 + (bl48)*slope4_8 + (bl812)*slope8_12
         + (bl12)*slope12_;
   pi = 2*arsin(1);
   ll = (1/(sqrt(2*pi*residual)))*exp(-(sqrtcd4-mu)**2/(2*residual));
   lllong = log(ll);

   model lastvisit ~ general(lllong + llsurv);
   random u0 ~ normal([0],[v11]) subject=pid;
```

```
run;

/*FITTING JOINT MODEL ON CD4 COUNT WITH WEEKS POST INFECTION FITTED AS PIECEWISE LINEAR
EFFECTS (MODEL IN TABLE 5.46)*/

title 'Joint Model - Piecewise CD4 count - W1(t) = U0 + U1t/ W2(t) = r0U0 + r1U1';
proc nlmixed data=mscdata;
   parms bs0=8.75 bl0=31.16 bl02=-4.03 bl24=-0.129 bl48=0.03242 bl812=-0.2339 bl12=-0.01504
         residual=7.8;

   /*Survival Exponential model on time to ARV initiation*/
   if (lastvisit) then do;
      linpsurv = bs0 + r0*u0 + r1*u1;
      alpha = exp(-linpsurv);
      G_t = exp(-alpha*days);
      g = alpha*G_t;
      llsurv = (ARV=1)*log(g) + (ARV=0)*log(G_t);
   end; else llsurv=0;

   /*Longitudinal model for CD4 count*/
   mu = (bl0 + u0)+ (bl02 + u1)*slope0_2 + (bl24 + u1)*slope2_4 + (bl48 + u1)*slope4_8
         + (bl812 + u1)*slope8_12 + (bl12 + u1)*slope12_;
   pi = 2*arsin(1);
   ll = (1/(sqrt(2*pi*residual)))*exp(-(sqrtcd4-mu)**2/(2*residual));
   lllong = log(ll);

   model lastvisit ~ general(lllong + llsurv);
   random u0 u1 ~ normal([0,0],[v11,v12,v22]) subject=pid;
run;

/************************************ VIRAL LOAD *************************/

/*FITTING JOINT MODEL ON VIRAL LOAD WITH WEEKS POST INFECTION FITTED AS A LINEAR EFFECT
(MODEL IN TABLE 5.49)*/

title 'Joint Model - Viral Load - W1(t) = U1t / W2(t)=0';
proc nlmixed data=mscdata;
   parms bs0=8.75 bl1=0.0542 residual=8.9074;

   /*Survival Exponential model on time to ARV initiation*/
   if (lastvisit) then do;
      linpsurv = bs0;
      alpha = exp(-linpsurv);
      G_t = exp(-alpha*days);
      g = alpha * G_t;
      llsurv = (ARV=1)*log(g) + (ARV=0)*log(G_t);
   end; else llsurv = 0;

   /*Longitudinal model for Viral Load*/
   mu = (bl1 + u1)*week;
   pi = 2*arsin(1);
   ll = (1/(sqrt(2*pi*residual)))*exp(-(logvl-mu)**2/(2*residual));
```

```
   lllong = log(ll);

   model lastvisit ~ general(lllong + llsurv);
   random u1 ~ normal([0],[v22]) subject=pid;
run;

/*FITTING JOINT MODEL ON VIRAL LOAD WITH WEEKS POST INFECTION FITTED AS A LINEAR EFFECT
(MODEL IN TABLE 5.51)*/

title 'Joint Model - Viral Load - W1(t) = U0+U1t / W2(t) = r0U0';
proc nlmixed data=mscdata;
   parms bs0=8.75 bl1=0.0542 residual=8.9074;

   /*Survival Exponential model on time to ARV initiation*/
   if (lastvisit) then do;
      linpsurv = bs0 + r0*u0;
      alpha = exp(-linpsurv);
      G_t = exp(-alpha*days);
      g = alpha * G_t;
      llsurv = (ARV=1)*log(g) + (ARV=0)*log(G_t);
   end; else llsurv = 0;

   /*Longitudinal model for Viral Load*/
   mu = u0 + (bl1 + u1)*week;
   pi = 2*arsin(1);
   ll = (1/(sqrt(2*pi*residual)))*exp(-(logvl-mu)**2/(2*residual));
   lllong = log(ll);

   model lastvisit ~ general(lllong + llsurv);
   random u0 u1 ~ normal([0,0],[v11,v12,v22]) subject=pid;
run;

/*FITTING JOINT MODEL ON VIRAL LOAD WITH WEEKS POST INFECTION FITTED AS A LINEAR EFFECT
(MODEL IN TABLE 5.53)*/

title 'Joint Model - Viral Load - W1(t) = U1t / W2(t) = r1U1';
proc nlmixed data=mscdata;
   parms bs0=8.75 bl1=0.0542 residual=8.9074;

   /*Survival Exponential model on time to ARV initiation*/
   if (lastvisit) then do;
      linpsurv = bs0 + r1*u1;
      alpha = exp(-linpsurv);
      G_t = exp(-alpha*days);
      g = alpha * G_t;
      llsurv = (ARV=1)*log(g) + (ARV=0)*log(G_t);
   end; else llsurv = 0;

   /*Longitudinal model for Viral Load*/
   mu = (bl1 + u1)*week;
   pi = 2*arsin(1);
   ll = (1/(sqrt(2*pi*residual)))*exp(-(logvl-mu)**2/(2*residual));
```

```
   lllong = log(ll);

   model lastvisit ~ general(lllong + llsurv);
   random u1 ~ normal([0],[v22]) subject=pid;
run;

/*FITTING JOINT MODEL ON VIRAL LOAD WITH WEEKS POST INFECTION FITTED AS PIECEWISE LINEAR
EFFECTS (MODEL IN TABLE 5.56)*/

title 'Joint Model - Piecewise Viral Load - W1(t) = U1t/ W2(t) = 0';
proc nlmixed data=mscdata;
   parms bl02=2.70 bl24=-0.456 bl48=0.055 bl812=-0.096 bl12=-0.0011 residual=0.42;

   /*Survival Exponential model on time to ARV initiation*/
   if (lastvisit) then do;
      linpsurv = bs0;
      alpha = exp(-linpsurv);
      G_t = exp(-alpha*days);
      g = alpha*G_t;
      llsurv = (ARV=1)*log(g) + (ARV=0)*log(G_t);
   end; else llsurv=0;

   /*Longitudinal model for Viral Load*/
   mu = (bl02 + u1)*slope0_2 + (bl24 + u1)*slope2_4 + (bl48 + u1)*slope4_8 + (bl812
+ u1)*slope8_12
         + (bl12 + u1)*slope12_;
   pi = 2*arsin(1);
   ll = (1/(sqrt(2*pi*residual)))*exp(-(logvl-mu)**2/(2*residual));
   lllong = log(ll);

   model lastvisit ~ general(lllong + llsurv);
   random u1 ~ normal([0],[v22]) subject=pid;
run;

/*FITTING JOINT MODEL ON VIRAL LOAD WITH WEEKS POST INFECTION FITTED AS PIECEWISE LINEAR
EFFECTS (MODEL IN TABLE 5.58)*/

title 'Joint Model - Piecewise Viral Load - W1(t) = U0/ W2(t) = r0U0';
proc nlmixed data=mscdata;
   parms bl02=2.70 bl24=-0.456 bl48=0.055 bl812=-0.096 bl12=-0.0011 residual=0.42;

   /*Survival Exponential model on time to ARV initiation*/
   if (lastvisit) then do;
      linpsurv = bs0 + r0*u0;
      alpha = exp(-linpsurv);
      G_t = exp(-alpha*days);
      g = alpha*G_t;
      llsurv = (ARV=1)*log(g) + (ARV=0)*log(G_t);
   end; else llsurv=0;

   /*Longitudinal model for Viral Load*/
   mu = u0 + (bl02)*slope0_2 + (bl24)*slope2_4 + (bl48)*slope4_8 + (bl812)*slope8_12
```

```
+ (bl12)*slope12_;
   pi = 2*arsin(1);
   ll = (1/(sqrt(2*pi*residual)))*exp(-(logvl-mu)**2/(2*residual));
   lllong = log(ll);

   model lastvisit ~ general(lllong + llsurv);
   random u0 ~ normal([0],[v11]) subject=pid;
run;

/*FITTING JOINT MODEL ON VIRAL LOAD WITH WEEKS POST INFECTION FITTED AS PIECEWISE LINEAR
EFFECTS (MODEL IN TABLE 5.60)*/

title 'Joint Model - Piecewise Viral Load - W1(t) = U1t/ W2(t) = r1U1';
proc nlmixed data=mscdata;
   parms bl02=2.70 bl24=-0.456 bl48=0.055 bl812=-0.096 bl12=-0.0011 residual=0.42;

   /*Survival Exponential model on time to ARV initiation*/
   if (lastvisit) then do;
      linpsurv = bs0 + r1*u1;
      alpha = exp(-linpsurv);
      G_t = exp(-alpha*days);
      g = alpha*G_t;
      llsurv = (ARV=1)*log(g) + (ARV=0)*log(G_t);
   end; else llsurv=0;

   /*Longitudinal model for Viral Load*/
   mu = (bl02 + u1)*slope0_2 + (bl24 + u1)*slope2_4 + (bl48 + u1)*slope4_8 + (bl812
+ u1)*slope8_12
         + (bl12 + u1)*slope12_;
   pi = 2*arsin(1);
   ll = (1/(sqrt(2*pi*residual)))*exp(-(logvl-mu)**2/(2*residual));
   lllong = log(ll);

   model lastvisit ~ general(lllong + llsurv);
   random u1 ~ normal([0],[v22]) subject=pid;
run;

/*FITTING JOINT MODEL ON VIRAL LOAD WITH WEEKS POST INFECTION FITTED AS PIECEWISE LINEAR
EFFECTS (MODEL IN TABLE 5.62)*/

title 'Joint Model - Piecewise Viral Load - W1(t) = U0+U1t/ W2(t) = r0U0';
proc nlmixed data=mscdata;
   parms bl02=2.70 bl24=-0.456 bl48=0.055 bl812=-0.096 bl12=-0.0011 residual=0.42;

   /*Survival Exponential model on time to ARV initiation*/
   if (lastvisit) then do;
      linpsurv = bs0 + r0*u0;
      alpha = exp(-linpsurv);
      G_t = exp(-alpha*days);
      g = alpha*G_t;
      llsurv = (ARV=1)*log(g) + (ARV=0)*log(G_t);
   end; else llsurv=0;
```

```
   /*Longitudinal model for Viral Load*/
   mu = u0 + (bl02 + u1)*slope0_2 + (bl24 + u1)*slope2_4 + (bl48 + u1)*slope4_8 + (bl812
+ u1)*slope8_12
         + (bl12 + u1)*slope12_;
   pi = 2*arsin(1);
   ll = (1/(sqrt(2*pi*residual)))*exp(-(logvl-mu)**2/(2*residual));
   lllong = log(ll);

   model lastvisit ~ general(lllong + llsurv);
   random u0 u1 ~ normal([0,0],[v11,v12,v22]) subject=pid;
run;
```