Statistical and Mathematical Modelling of HIV and AIDS, Effect of Reverse Transcriptase Inhibitors and Causal Inference for HIV Mortality

Olina Ngwenya

University of Kwazulu-Natal

Supervised by: Prof H.G. Mwambi

University of Kwazulu-Natal, South Africa

25 November 2010

Submitted in partial fulfillment of the academic requirements for the degree of Master of Science in the School of Statistics and Actuarial Science,

University of Kwazulu-Natal

Abstract

The HIV and AIDS epidemic has remained one of the leading causes of death in the world and has been destructive in Africa with Sub-Saharan Africa remaining the epidemiological locus of the epidemic. HIV and AIDS hinders development by erasing decades of health, economic and social progress, reducing life expectancy by years and deepening poverty [57].The most urgent public-health problem globally is to devise effective strategies to minimize the destruction caused by the HIV and AIDS epidemic. Due to the problems caused by HIV and AIDS, well defined endpoints to evaluate treatment benefits are needed. The surrogate and true endpoints for a disease need to be specified. The purpose of a surrogate endpoint is to draw conclusions about the effect of intervention on true endpoint without having to observe the true endpoint. It is of great importance to understand the surrogate validation methods. At present the question remains as to whether $CD4$ count and viral load are good surrogate markers for death in HIV or there are some better surrogate markers. This dissertation was undertaken to obtain some clarity on this question by adopting a mathematical model for HIV at immune system level and the impact of treatment in the form of reverse transcriptase inhibitors (RTIs). For an understanding of HIV, the dissertation begins with the description of the human immune system, HIV virion structure, HIV disease progression and HIV drugs. Then a review of an existing mathematical model follows, analyses and simulations of this model are done. These gave an insight into the dynamics of the $CD4$ count, viral load and HIV therapy. Thereafter surrogate marker validation methods followed. Finally generalized estimating equations (GEEs) approach was used to analyse real data for HIV positive individuals, from the Centre for the AIDS Programme of Research in South Africa (CAPRISA). Numerical simulations for the HIV dynamic model with treatment suggest that the higher the treatment efficacy, the lower the infected cells are left in the body.

The infected cells are suppressed to a lower threshold value but they do not completely disappear, as long as the treatment is not 100% efficacious. Further numerical simulations suggest that it is advantageous to have a low proportion of infectious virions (ω) at an individual level because the individual would produce few infectious virions to infect healthy cells. Statistical analysis model using GEEs suggest that $CD4$ count< 200 and viral load are highly associated with death, meaning that they are good surrogate markers for death. An interesting finding from the analysis of this particular data from CAPRISA was that low $CD4$ count and high viral loads as surrogates for HIV survival act independently/additively. The interaction effect was found to be insignificant. Individual characteristics or factors that were found to be significantly associated with HIV related death are weight, $CD4$ count< 200 and viral load.

Declaration

This dissertation represents the original work of the author. The work done by others or by myself previously has been acknowledged and referenced accordingly. This research has not been previously submitted, in any form, to any institution.

....

Olina Ngwenya

....

Prof. H.G. Mwambi

MSc Supervisor

Acknowledgement

I would like to thank my supervisor, Prof. Henry G. Mwambi, for his tremendous support, help and guidance throughout my masters. My sincere thanks go to the University of Kwazulu Natal staff for teaching me some Statistics courses which were very useful for my project. I would also like to thank CAPRISA for providing me with data to do my analysis. Many thanks to Nonhlanhla Yende and Kogie Naidoo both from CAPRISA for all the explanations about the data. Many thanks to all the AIMS lecturers who taught me and the staff of the Department of Mathematics, NUST, Bulawayo, Zimbabwe for adding value to my life. I would not forget to appreciate my friends Dikokole Maqutu and Zodwa Makukula for helping me with the programming which was involved in my masters thesis. My sincere thanks goes to AIMS and SACEMA for funding my studies. Last but not least, I would like to thank my son Sijabuliso and my fiancee Phendu for giving me the gift of time. GOD BLESS YOU ALL

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The HIV and AIDS epidemic has remained one of the leading causes of death in the world and has been destructive in Africa with Sub-Saharan Africa being the most affected. According to the figures published by the Joint United Nations Programme on HIV and AIDS (UNAIDS) and the World Health Organisation (WHO), 22 million [$20.5 - 23.6$ million] people living with HIV are currently in Sub-Saharan Africa [57]. They account for two thirds of all people living with HIV, also 60% of all women with HIV are in Sub-Saharan Africa. It was also stated that 1.9 million [$1.3 - 1.7$ million] people died of HIV-related illnesses in the region in 2007 and 2.7 million new infections occurred worldwide during the same year [57]. The most urgent public-health problem globally is to devise effective strategies to minimize the devastation caused by the HIV and AIDS epidemic particularly in resource poor nations.

The most used and advocated for prevention efforts to; be faithful, abstain, and condom use (male condom), have since spread in Sub-Saharan and beyond. While the idea of abstaining from sex has had some success among young, unmarried people, but the lack of adherence to fidelity by many married men or failure to remain faithful to their often-monogamous wives have made the use of condoms the viable option. Since women are both biologically and socio-culturally more vulnerable to HIV infection, efforts to identify HIV-prevention methods that women could control emerged, hence the advent of the female condom [57]. Also it should be noted that evidence exists that shows that women also engage in infidelity practices [34]. Despite the knowledge of

successful HIV prevention by use of both male and female condoms, HIV continues to spread at an alarming rate in developing countries.

Despite hopeful signs of abatement in Kenya, Uganda and Zimbabwe, the HIV epidemic continues to spread. There was a decline in these countries because of the changes in sexual behaviour through the use of condoms and reduction in the number of sexual partners. In some countries such as Botswana and South Africa the HIV infections appear to be stabilizing at high levels whilst infection rates continue to rise in Mozambique [57]. South Africa has 0.7% of the world's population but it has 17% of the world's HIV and AIDS cases (5.5 million people) and the greatest HIV and AIDS burden compared to any other country [10]. HIV and AIDS hinders development, affects health, economic and social progress [57].

Due to the problems caused by HIV and AIDS, we need to develop both mathematical and statistical models to give us some insights about the dynamics of the disease progression. We also need well defined endpoints in clinical trials and observational studies so as to evaluate treatment benefits. It would be cost effective if we are able to specify the surrogate and true endpoints for a disease before the commencement of a study. The purpose of a surrogate endpoint is to draw conclusions about the effect of intervention on true endpoint without having to observe the true endpoint. In some cases the true endpoint is irreversible such as death therefore to save lives a surrogate endpoint is much better. Therefore, it is of great importance to understand the theory on surrogate validation methods.

The aim of this dissertation is to focus on the HIV and AIDS problem as a biological process and analysis of data generated from it. The development of a mathematical model of HIV and AIDS and simulating the treatment effect a prior could give us some insights about the different treatment efficacies being used. The development of an appropriate statistical model for the probability of death against some variables could give us insights about some well known surrogate markers. Thus, the goal of this research is to review an existing mathematical model and simulate the treatment of HIV and AIDS in the form of reverse transcriptase inhibitors with different efficacies. By doing this, one could be able to give some qualitative comments and suggestions about different treatment efficacies. The other goal of this research is to develop a statistical model for the probability of death against some independent variables (including

surrogate markers) and some suggestions can be made from the results as to whether the well known surrogate markers are good surrogates or not.

1.1 Immune System

The body's immune system comprises of a complex system of blood proteins and white blood cells, which work together to repel attacks by invading organisms. White blood cells are formed in the bone marrow and they form three different regimes, namely, phagocytes (including macrophages), and two types of lymphocytes, namely T cells and B cells. Phagocytes constantly patrol the whole body (the bloodstream, tissue and lymphatic system) so that they may detect an enemy and immediately try to engulf and destroy it. While phagocytes cannot destroy organic invaders such as viruses, bacteria, protozoa and fungi, they are usually effective in destroying chemical poisons and environmental pollutants such as dust, smoke and asbestos particles.

The macrophages' function is to mobilise the specific defence system, which consists of the lymphocytes (i.e T cells and B cells). The macrophages surrounds the organic invader (e.g. virus) and captures a specific particle, called an antigen, from the invading organic invader. The macrophages then display this antigen on its own cell surface as a 'flag of war'. This flag (the antigen) plays an important role in the immune system's response, because it alerts the T cells to attack the invaders [54]. T cells are pre-programmed in the thymus to recognise the antigen (the 'flag' carried by the macrophages) by its shape.

The $CD4$ cells are the type of T cells that recognise the antigen and they are also called T helper cells, or $CD4$ lymphocytes. $CD4$ cells are the most crucial in the immune system defense response because they protect the body from invasion by certain bacteria, viruses, fungi, and parasites. However once or in event that the number of $CD4$ has become radically depleted, opportunistic infections can set in and can subsequently overwhelm the body.

In the event of invasion by foreign organism e.g. a virus, the $CD4$ combines forces with the macrophages, and they activate the remaining components of the defence system. The $CD4$ cells begin to multiply, and they activate more phagocytes and send chemical messages to the B

cells and killer T cells, which are sensitive to the invading virus, to multiply. B cells are located in the lymph nodes, and they then multiply and divide into two groups, plasma B cells and memory B cells. The plasma B cells are responsible for manufacturing antibodies. These antibodies render invading organisms harmless by neutralising them or by clinging on their surfaces thus preventing them from performing their function [54].

While the immune system prepares its forces, the organic invader penetrates some of the body cells and this is the only place where they can multiply. When they are ready, the killer T cells with the aid of certain $CD4$ cells destroy these infected cells by chemically piercing their membranes so that the contents spill out and this interrupts the multiplication of the organic invader. Antibodies then neutralise the viruses by attaching themselves to the viruses' surface, thereby preventing them from attacking other cells. The progress of the invading organisms is therefore slowed and this makes them easy victims for the phagocytes or macrophages, which then come to 'digest' them. In addition chemical reactions are produced by antibodies, which can kill infected cells.

When all the invaders have been destroyed, a member of the T cell family takes control, this is the "suppressor T" or the peace maker. Suppressor T cells release a substance which stops B cells from manufacturing antibodies. The killer T cells are also ordered to stop attacking and the $CD4$ cells are ordered to stop their work. Memory T and B cells remain in the blood and lymphatic system, they 'remember' the specific invader (antigen), and they are ready to act defensively should the same virus once again invade the body. Thus in order for a modeler to be able to develop a realistic mathematical or statistical model for the interaction of the immune system and the HIV, understanding of the inherent biological processes described above are an important pre-requisite.

For an understanding of HIV and AIDS and surrogate marker validation the thesis has two distinct parts, but all equally important parts. The first part of the thesis deals with the construction of a mathematical model for HIV and AIDS with treatment. The model is linked to real data from published literature and its behaviour qualitatively investigated through simulations. The second part of the thesis deals with the analysis of the statistical modelling of surrogate markers for HIV related survival based on real data from a longitudinal study. Chapters 1 and 2 give an

introduction about HIV AND AIDS. In Chapter 3 a review of an existing mathematical model is given and Chapter 4 gives a statistical model which was used in the estimation of parameters and subsequently these parameters are then used for simulations in Chapter 5. The second part, Chapter 6 , deals with the statistical modelling of HIV related mortality with two key surrogate markers namely $CD4\ T$ cell and viral load counts as part of the predictor variables.

Chapter 2

Immune System and HIV and AIDS

2.1 HIV and Immune System

Infection with HIV results in Acquired immune deficiency syndrome (AIDS). AIDS is characterised by a failing immune system and susceptibility to opportunistic infections caused by fungi, bacteria, parasites and other viruses. The infected individual develops an inability to fight off the constant onslaught of opportunistic infections due to the collapsing state of the immune system, which finally results in death if no treatment is initiated.

HIV belongs to a class of viruses known as retroviruses [20]. Retroviruses have their genome in the form of RNA which is then translated into DNA during its lifecycle. This is the reverse of what usually happens in most biological processes, hence the prefix 'retro'. Most biological processes proceed from DNA to RNA. HIV is also a lentivirus [11]. Lentiviruses are characterized by long incubation periods and long duration of illness. As a result, HIV positive individuals can remain asymptomatic for many years and not know that they are infected, while spreading the disease to many others. HIV infection is spread by the transfer of infected bodily fluids such as blood, semen, vaginal fluid, or breast milk [5], which contain HIV present as free virus or as infected immune cells.

Usually, the routes of transmission are unprotected sexual intercourse, sharing of unsterilised hypodermic needles, usually amongst intravenous drug users, blood transfusions and from mother

to child during breast feeding and child birth [59].

2.1.1 HIV Virion Structure

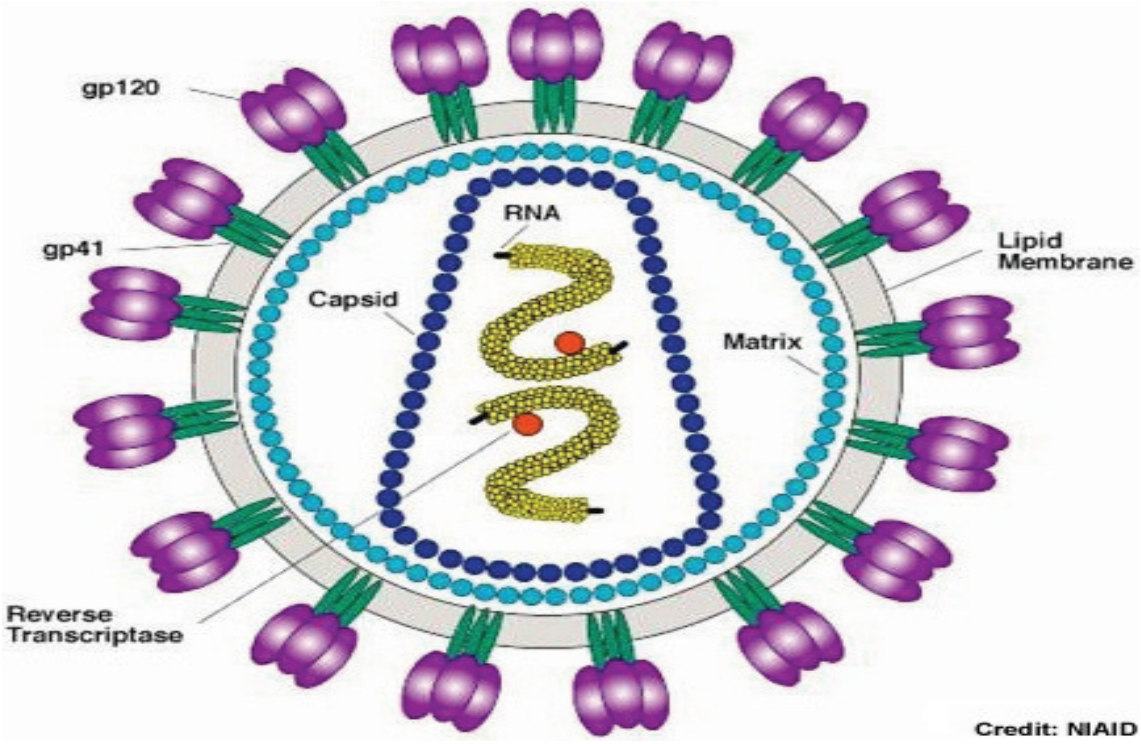

Figure 2.1: The Virus [USA.Fed.Gov.].

A pictorial representation of an HIV virion is shown on Figure 2.1. It is almost spherical and has a diameter of about 120 nanometers [30].

2.1.2 HIV Replication Cycle

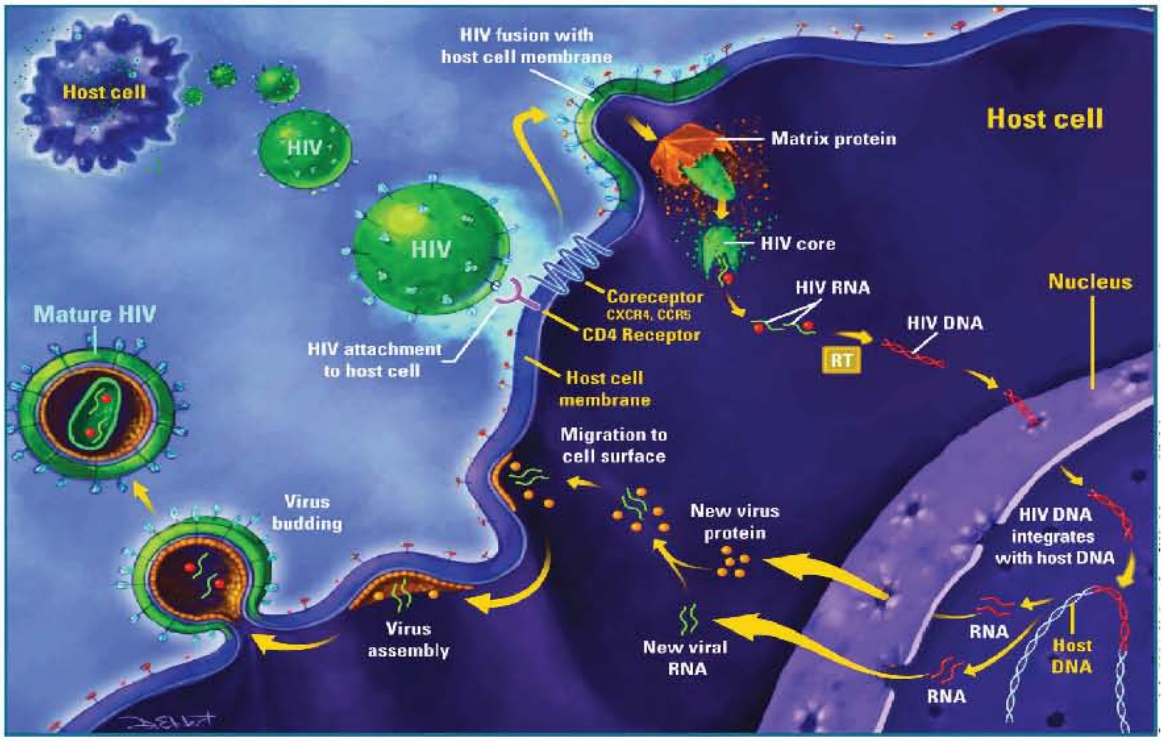

Figure 2.2: The life cycle of virus in the $CD4^{+}T$ cells [University of Washington, 2004.]

The HIV replication cycle can be broken down into five steps: 1) Fusion; 2) Reverse Transcription; 3) Integration; 4) Cleaving; and 5) Budding. These steps are explained below:

When HIV invades the body, the macrophages attempt to do their usual job by capturing particles (antigen) from HIV [54]. When the macrophages attempt to make contact with the $CD4$ cells to warn them about the invasion, the viruses attack the $CD4$ cells directly. This is a unique response that makes HIV so dangerous to its victims. The virus and the $CD4$ cell now join membranes. The virus then sheds it's outer layer and enters the $CD4$ cell with its own genetic material (viral RNA).

The HIV's viral RNA must be changed (or "reverse transcribed") to DNA, in order to use the cell to manufacture more viruses. The HIV itself carries with it an enzyme called reverse transcriptase

which it then uses to transform its viral RNA into double-strand viral DNA.

The viral DNA then fuses with the host cell's own DNA or genetic material in the nucleus of the cell, and makes numerous copies or replicas of viral RNA and viral protein. The new viral RNA and the viral protein are enabled by the protease enzyme to merge and bud from the cell membrane as fully functional HI viruses, perfect replicas of the original HI virus that entered the cell in the first place. The hijacked cells are killed, as the new HI viruses bud from the cell. They then move out into the blood stream or surrounding tissue to infect more cells and repeat the whole process.

Due to the fact that the HIV attacks the $CD4$ cells, these cells are unable to do what they normally do if confronted by an alien virus i.e. coordinate the body's defence against HIV. Instead, they are captured and forcibly turned into small factories to manufacture the very agents of destruction against which they are supposed to defend the body (HI viruses) from.

Antibodies formed are completely powerless against HIV because HI viruses hide inside the $CD4$ cells while they subvert the cell for their purpose. Therefore the body is left defenceless because the antibodies can not attack and kill the $CD4$ cells.

2.1.3 HIV Disease Progression

An HIV-infected person is classified as having AIDS, when the $CD4^{+}T$ cell count, which is normally around $1000mm^{-3}$, drops to $200mm^{-3}$ or below [46]. Just because the $CD4^{+}T$ cells play a central role in immune regulation, their depletion has widespread deleterious effects on the functioning of the immune system as a whole and leads to immunodefiency that characterizes AIDS. T cells are normally replenished in the body, and the infection may affect the source of new T cells or the homeostatic processes that control T cell production and numbers in the body [46]. Although HIV can kill cells that it infects, only a small fraction of $CD4^{+}T$ cells $(10^{-5}$ to $10^{-4})$ are productively infected at any one time. Thus in addition to direct killing of T cells, HIV may have many indirect effects [2]. The disease ($AIDS$) takes 10 years to develop on average. There are four typical stages during disease progression that an HIV-positive individual experiences. These stages are classified as: Stage one-Primary infection, Stage two-Asymptomatic

infection, Stage three-Symptomatic infection and Stage four-AIDS. These stages were classified by the World Health Organization (WHO) in 1990 [61]. Acute infection (primary infection), is a

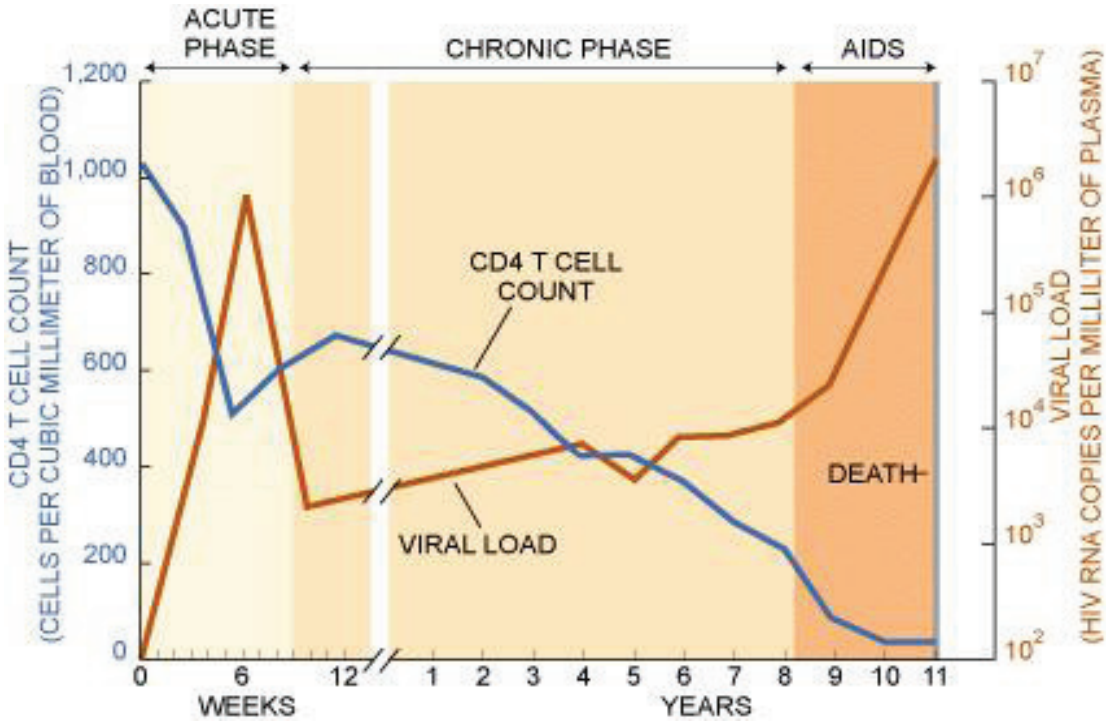

Figure 2.3: Progression of HIV in a typical patient [Wikimedia Commons]

period of rapid viral replication that immediately follows an individual's exposure to HIV leading to an abundance of virus in the peripheral blood with levels of HIV commonly approaching seven million viruses per milliliter [47]. The numbers of circulating $CD4^+T$ cells drops and the $CD8^+T$ cells are activated to kill $HIV-$infected cells. The $CD8^+T$ cell response is thought to be important in controlling virus levels, which peak and then decline, as the $CD4^+T$ cell counts rebound to around 800 cells per cubic millimeter [64]. Though a good $CD8^+T$ cell response does not eliminate the virus, it has been linked to slower disease progression and better prognosis [45]. During the acute infection period (usually $2-4$ weeks post-exposure) most individuals develop influenza or mononucleosis-like illness called $HIV-$infection [63]. The most common symptoms during this stage include fever, lymphadenopathy, pharyngitis, rash, myalgia, malaise,

mouth and esophagal sores. These symptoms usually last a week. The patient is much more infectious during this period, therefore recognizing the syndrome at this stage through voluntary testing say is very important to aid in implementing control and intervention strategies [13].

The number of viral particles in the blood stream are reduced by a strong immune defense response. This marks the start of the infection's clinical latency stage (chronic phase), and this stage may vary between 2 weeks and 20 years. HIV is active within the lymphoid, where large amounts of virus become trapped in the follicular dendritic cells (FDC) network [6]. Viral particles accumulate both in infected cells, and as free virus, and the surrounding tissues that are rich in $CD4^{+}T$ cells may also become infected. Patients are still infectious at this stage.

Cell-mediated immunity is lost when $CD4^{+}T$ cell count decline below a critical level of 200 cells per cubic millimeter, and infections with a variety of opportunistic microbes appear. The first symptoms often include moderate and unexplained weight loss, recurring respiratory tract infections (such as sinusitis, bronchitis, otitis media, pharyngitis), prostatitis, skin rashes, and oral ulcerations.

It should nonetheless be pointed out that the stages described above are for the average disease progression process but in reality there is evidence of individual to individual heterogeneity which has to be accounted for in the delivery of individual patient care.

2.1.4 HIV and Drugs

The fact that HIV replicates rapidly producing on average 10^{10} viral particles per day, led to the realization that HIV was evolving so rapidly that treatment with a single drug was bound to fail. This realization helped in speeding the recommended form of treatment from monotherapy to combination therapy employing three or more drugs, and has had a major impact in extending people's lives [46]. Even with the combination therapy virus eradication does not seem like an easily attainable goal. In addition, mathematical modelling, has shown that patients who continue taking antiretroviral drugs for a period of at least $2-3$ years after the virus is no longer detectable in the blood, are better than those who do not. This raises the issue of lower undetectable limits in measuring viral loads which requires special statistical methods but not part of the current

work.

Dynamical modelling of the disease has proved to be very important because it has uncovered important features of HIV pathogenesis and impacted the way in which $AIDS$ patients are treated with potent antiretroviral drugs. Recent developments in the analysis of such models has necessitated the use of dynamic models as well as statistical methods to aid in the estimation of model parameters as inputs in simulation studies. The current thesis demonstrates exactly how the two approaches complement each other in understanding the HIV and AIDS problem. An important and re-emerging area of active research is that of combining both mathematical models for infectious disease and statistical modelling of the processes. The advantage of this approach is that the modeller is able to describe a process and also estimate parameters associated with the process through observed data [25, 44].

Chapter 3

A Mathematical Model for HIV Dynamics Including Treatment

3.1 Model formulation

In AIDS research, the cause of progressive depletion of $CD4^+T$ cells in HIV-infected people is one of the most fundamental and controversial issues. HIV infects and kills $CD4^+T$ cells [25]. An immediately intuitive assumption is that HIV-mediated destruction of $CD4^+T$ cells directly reduces the number of these cells and that the high turnover rates of T cells and the slow progression to AIDS reflect a long, but eventually lost struggle of the immune system to replace killed cells in its effort to maintain T-cells homeostasis [41]. HIV mainly infects activated $CD4^+T$ cells. Activated cells normally follow different dynamics than cells that belong to the resting populations (or quiescent cells) whose number are controlled by homeostatic mechanisms [24]. T cells undergo several rapid rounds of division, upon activation, and then they stop dividing and most die. Some of the activated cells escape this activation-induced cell death (AICD) and enter the population of resting memory cells [24].

Many models for HIV dynamics have involved target cells (mainly uninfected $CD4$), infected cells producing viruses, and circulating viruses. The activated state is worth distinguishing, because the activation of $CD4$ has been recognized to have a central role in HIV pathogenesis [25].

Actually, activated cells make a better target than quiescent (inactive) cells and viral replication is rapid and efficient in activated cells [54]. It is important to note that on average non-infectious virions are predominant compared to infectious ones, therefore it is useful to distinguish between infectious and non-infectious virions [11]. Antiretroviral therapy, in the $ALBIANRS070$ trial [43], included reverse transcriptase inhibitors only. This type of antiretroviral drugs limits cell infection by inhibiting reverse transcription of HIV RNA and thus can be modelled by limiting the new production of activated infected T cells denoted by T^* through the parameter η [25]. We can represent the system of HIV dynamics adopted in the current thesis in graphical form as shown in Figure 3.1:

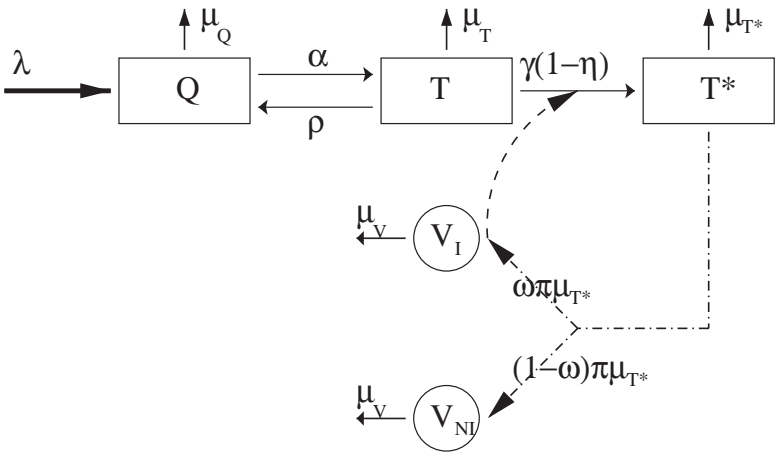

Figure 3.1: Graphical representation of the system for HIV dynamics

Thus this model can be written as a system of five ordinary differential equations(ODE) given by

$$\left.\begin{aligned}
\frac{dQ}{dt} &= \lambda + \rho T - \alpha Q - \mu_Q Q \\
\frac{dT}{dt} &= \alpha Q - (1-\eta)\gamma T V_I - \rho T - \mu_T T \\
\frac{dT^*}{dt} &= (1-\eta)\gamma T V_I - \mu_{T^*} T^* \\
\frac{dV_I}{dt} &= \omega \mu_{T^*} \pi T^* - \mu_v V_I \\
\frac{dV_{NI}}{dt} &= (1-\omega)\mu_{T^*} \pi T^* - \mu_v V_{NI}
\end{aligned}\right\} \tag{3.1}$$

The state variables of the model are Q, T, T^*, V_I and V_{NI} where Q represents quiescent non-infected $CD4$ cells, T represents activated noninfected $CD4$ cells and T^* represents activated infected $CD4$ cells. V_I and V_{NI} are infectious and non-infectious virions, respectively. This is the HIV-model, that is studied in the current work. It is different from other models because it starts with quiescent cells whilst other models commonly start with target $CD4$ cells. This feature makes it more realistic and plausible as a caricature of reality. The meaning of each parameter is given in the Table 3.1 below.

Table 3.1: Description of model parameters

Parameter	Description
α	Activation rate of Q cells (day^{-1}) ,
λ	Rate of Q cells production $(\mu_l^{-1} day^{-1})$,
μ_{T^*}	Death rate of T^* cells (day^{-1}),
π	Number of virions produced per T^* cell ,
μ_T	Death rate of T cells (day^{-1}),
η	Efficacy of treatment (proportion),
γ	Infection rate of T cells per virion,
μ_Q	Death rate of Q cells ,
μ_v	Clearance of free virions ,
ρ	Rate of reversion to the quiescent state
ω	Proportion of infectious virions

3.1.1 Assumptions

- An assumption was made that before initiation of antiretroviral treatment the values of the state variables are those of steady state of the ODE system with $\eta = 0$. The implication by this assumption is that the treatment is initiated far from the initial infection,
- Populations of cell particles are homogeneously mixed,
- Interaction of infectious virions and uninfected $CD4$ cells is mass action type.

3.1.2 Equilibrium State

The initial condition (where $t = 0$ refers to treatment initiation) represented by the equilibrium state of the system is the point $(Q(0), T(0), T^*(0), V_I(0), V_{NI}(0))$ where,

$$\left.\begin{aligned} Q(0) &= \tfrac{1}{\alpha+\mu_Q}(\lambda + \tfrac{\rho\mu_v}{\omega\gamma\pi}) \\ T(0) &= \tfrac{\mu_v}{\omega\gamma\pi} \\ T^*(0) &= \tfrac{1}{\mu_{T^*}}(\tfrac{\alpha}{\alpha+\mu_Q}(\lambda + \tfrac{\rho\mu_v}{\omega\gamma\pi}) - \tfrac{(\rho+\mu_T)\mu_v}{\omega\gamma\pi}) \\ V_I(0) &= \tfrac{\alpha\omega\pi}{\mu_v(\alpha+\mu_Q)}(\lambda + \tfrac{\rho\mu_v}{\omega\gamma\pi}) - \tfrac{\rho+\mu_T}{\gamma} \\ V_{NI}(0) &= \tfrac{(1-\omega)\pi}{\mu_v}(\tfrac{\alpha}{\alpha+\mu_Q}(\lambda + \tfrac{\rho\mu_v}{\omega\gamma\pi}) - \tfrac{\mu_v(\rho+\mu_T)}{\omega\gamma\pi}) \end{aligned}\right\}$$

The above expressions are found by setting the system of equations (3.1) to zero with $\eta = 0$ (no treatment) and solving for Q, T, T^*, V_I and V_{NI} simultaneously. Note that this is the disease endemic equilibrium as opposed to the disease free equilibrium given by $(Q(0), T(0), 0, 0, 0)$.

3.1.3 Reproduction Number

The basic reproduction number, R_0, is defined as the average number of secondary infected cells generated by a single infected cell placed in an uninfected cell population.

There are several methods that are used in the calculation of the basic reproduction number such as using the determinant of the Jacobian matrix, the next generation matrix, the survival function and many others. For our model we use the next generation matrix as presented in [16]. The

system (3.1) can be written as

$$x^{'} = \mathcal{F}(x) - \mathcal{V}(x)$$

where

$$\mathcal{F}(x) = \begin{pmatrix} \gamma T V_1 \\ 0 \end{pmatrix}$$

and

$$\mathcal{V}(x) = \begin{pmatrix} \mu_{T^*} T^* \\ -\omega\pi\mu_{T^*} T^* + \mu_v V_1 \end{pmatrix}$$

The matrices for new infection terms (F) and the transfer terms (V) at the disease free equilibrium (DFE) are as follows;

$$F = \begin{pmatrix} 0 & \frac{\gamma\alpha\lambda}{(\alpha+\mu_Q)(\rho+\mu_T)-\rho\alpha} \\ 0 & 0 \end{pmatrix},$$

and

$$V = \begin{pmatrix} \mu_{T^*} & 0 \\ -\omega\pi\mu_{T^*} & \mu_v \end{pmatrix}.$$

Computing the inverse of V we have

$$V^{-1} = \begin{pmatrix} \frac{1}{\mu_{T^*}} & 0 \\ \frac{\omega\pi}{\mu_v} & \frac{1}{\mu_v} \end{pmatrix}$$

The product FV^{-1} given by

$$FV^{-1} = \begin{pmatrix} \frac{\gamma\alpha\lambda(\omega\pi)}{\mu_v[(\alpha+\mu_Q)(\rho+\mu_T)-\rho\alpha]} & \frac{\gamma\alpha\lambda}{\mu_v[(\alpha+\mu_Q)(\rho+\mu_T)-\rho\alpha]} \\ 0 & 0 \end{pmatrix}$$

is the next generation matrix. The reproduction number (R_0) is defined as the spectral radius of the next generation matrix [16]. The spectral radius of the matrix represents the dominant eigenvalue of that matrix. Hence the dominant eigenvalues of matrix FV^{-1} is R_0 given by

$$R_0 = \frac{\gamma\alpha\lambda(\omega\pi)}{\mu_v[(\alpha+\mu_Q)(\rho+\mu_T)-\rho\alpha]}.$$

When $R_0 < 1$ the infected $CD4$ cells die out and when $R_0 > 1$ the infection will spread hence more $CD4$ cells become infected and the subsequent result is that higher numbers are more likely to cause AIDS. It can immediately be inferred from the expression of R_0 that an effective way

to reduce infected $CD4$ cells is to reduce γ, α, ω. One can also think of reducing λ and π but some strategies are more realistic and feasible than others thus practicability is also one factor to consider when designing control strategies. For example λ is the rate of Q cell production which may not be that easy to alter.

3.1.4 Effective Reproduction Number (R_e)

The effective reproduction number (R_e) is the actual average number of secondary infected cells per primary case observed in the $CD4$ population with an infective $CD4$ cell in the presence of treatment (η). The value of R_e is typically smaller than the value of the basic reproductive number R_0, and it reflects the impact of treatment and depletion of activated $CD4$ cells by the infection. Using the same method used to obtain (R_0) we get

$$R_e = \frac{(1-\eta)\gamma\alpha\lambda(\omega\pi)}{\mu_v[(\alpha+\mu_Q)(\rho+\mu_T)-\rho\alpha]}$$

Note that if $\eta = 1$, $R_e = 0$, showing that if we have a 100% effective treatment there are no new infected secondary cells to be produced. Infact this would imply that the individual is cured. The statistical estimation of R_e is key in monitoring and surveillance of an epidemic in the presence of an intervention. It is an indicator of whether or not the intervention is working or not.

In order to study the qualitative behaviour of the model system 3.1 through computer simulations we need estimates of the model parameters. The problem of parameter estimation for a dynamical system specifically system (3.1) is discussed in Chapter 4.

Chapter 4

Statistical Model

In this chapter we briefly describe a statistical estimation method based on MLE of estimate parameters for a dynamical model for a disease process such as that presented in chapter 3 equation 3.1. This work is based on a relatively complex ODE model for HIV infection and a model of observations including the issue of detection limits [25].

There are two sources of parameter values for this model. We will use the parameter values that were estimated in [25] and some that are found from literature as shown in Table 5.2. The advantage of using these estimated parameter values in [25] is that sound statistical methods were used in their estimation. However it should be noted that strictly speaking the findings are specific to the cohort of individuals in the study and the HIV type which in this case was HIV-1.

The parameters estimated in [25] are shown in Table 5.1. The parameter γ was at the limit of non-identifiability, it was then determined in a plausible range of values by profile likelihood [25]. The natural parameters' vector for subject i was then:

$$\xi^{(i)} = (\lambda^{(i)}, \alpha^{(i)}, \eta^{(i)}, \mu_T^{(i)}, \mu_{T^*}^{(i)}, \pi^{(i)})'.$$

The link functions used in the estimation algorithm were the log transforms for all parameters to ensure that the optimization yields positive estimates of parameters with the exception for $\eta^{(i)}$ for which an inverse logistic function or logit link function was taken (because $0 < \eta^{(i)} < 1$) so that $\tilde{\eta^{(i)}} = log\frac{\eta^{(i)}}{1-\eta^{(i)}}$ where $\tilde{\eta^{(i)}}$ is the associated linear predictor.

4.1 Statistical model for the system of ODE equations

Consider an ordinary differential equations (ODE) model for a population of subjects. For subject i for $i = 1, ..., n$, this can be written as:

$$\left.\begin{array}{l} \frac{d\mathbf{X}^{(i)}(t)}{dt} = f(\mathbf{X}^{(i)}(t), \xi^{(i)}) \\ \mathbf{X}^{(i)}(0) = h(\xi^{(i)}), \end{array}\right\} \tag{4.1}$$

where $\mathbf{X}^{(i)}(t) = (X_1^{(i)}(t), ..., X_K^{(i)}(t))'$ is the vector of the K state variables (or components), where we let $\mathbf{X}(t, \xi^{(i)}) = \mathbf{X}^{(i)}(t)$, to underline the fact that $\xi^{(i)}$ completely determines the trajectories $\mathbf{X}^{(i)}(t)$. We assume that f and h are twice differentiable with respect to $\xi^{(i)}$ where $\xi^{(i)} = (\xi_1^{(i)}, ..., \xi_p^{(i)})'$, ($'$ is transpose) is a vector of p individual parameters that appear naturally in the ODE system and have a biological interpretation generally.

Ideally one can think of a parsimonious model for $\xi^{(i)}$ to allow for interindividual variability . This variability may be explained through explanatory variables, or may represent underlying unobserved effects. If the variability is unexplained, it is accounted for by random effects. Random effects models are now well developed for Gaussian outcomes and to a good extent for non-Gaussian models [15, 42, 58]. Infact generally one can assume that $\xi^{(i)} = \xi_{(0)} + \varepsilon_j$ where $\xi_{(0)}$ is a population average component and ε_j is the individual specific contribution to $\xi^{(i)}$ where $\xi_{(ii)} \sim N(0, \sigma^2)$ say.

In general introduce a link function that relates $\xi^{(i)}$ to a linear model involving explanatory variables and random effects, like in generalized linear mixed models [40]. To simplify the work, one can restrict this component-wise transforms to

$$\left.\begin{array}{ll} \tilde{\xi}_l^{(i)} = \Psi_l(\xi_l^{(i)}), & \\ \tilde{\xi}_l^{(i)} = \phi_l + \mathbf{z^{(i)}}'_l \beta_l + \boldsymbol{\omega}^{(\mathbf{i})'}{}_l \mathbf{b^{(i)}}, & l \leq p, \end{array}\right\} \tag{4.2}$$

where ϕ_l is the intercept, $\mathbf{z_1^{(i)}}$ is a vector of explanatory variables associated with the fixed effects of the lth biological parameter and $\omega_\mathbf{1}^{(\mathbf{i})}$ is a vector of explanatory variables associated with the random effects of the lth biological parameter. β_l's are vectors of regression coefficient associated with the fixed effects. A common assumption used is that $\mathbf{b^{(i)}} \sim N(0, \sum)$, where $\mathbf{b^{(i)}}$ is the individual vector of random effects of dimension q. Let $\mathbf{A} = (a_{l''l'})_{l' \leq l'' \leq q}$ be the lower triangular

matrix with positive diagonal elements such that $\mathbf{AA}' = \sum$ (Cholesky decomposition). Therefore we can write $\mathbf{b^{(i)}} = \mathbf{Au^{(i)}}$ with $u^{(i)} \sim N(0, I_q)$. Substituting this reparameterization of $\mathbf{b^{(i)}}$ into (4.2) simplifies the problem to that of a standard linear model to enable the optimization to converge faster.

4.2 Linking the model for the observations

Usually, not all the components of the system can be observed. Functions $g_m(.)$, $m = 1, ..., M$ of $\mathbb{R}^K$ to $\mathbb{R}$ are introduced to link the potential observations to the original system. These functions are assumed to be twice differentiable. These functions allow observation of only some of the components of the original system, or observation of combinations of several components, for instance, the model may distinguish between non-infected and infected $CD4$, but only the total number of $CD4$ is observed.

Transformations such as the logarithm may also be included in these functions. The $g_m(\cdot)$ are thus assumed to be completely known and are called the observable components. To link the theoretical model to data let Y_{ijm} denote the jth measurements of the mth observable component for subject i at time t_{ijm}; we assume that:

$$\begin{aligned} Y_{ijm} &= g_m(\mathbf{X}(t_{ijm}, \tilde{\xi}^{(i)})) + \epsilon_{ijm} \\ j &= 1, ..., n_{im}, m = 1, ..., M, \end{aligned} \tag{4.3}$$

where ϵ_{ijm} are independent Gaussian with zero mean and variance σ_m^2. Here the ϵ_{ijm}'s are supposed to be independent because they represent measurement errors. We can roughly check this assumption by looking at the correlations among residuals. We assume that the random variables Y_{ijm} all follow a Gaussian distribution thus linear mixed models can readily be used to deal with random effects.

Thus, both the observed dependencies for the within-patient observations of a given biomarker and more generally the correlations among the biomarkers' $\mathbf{Y}$'s are completely determined by the mechanistic relationships among the $\mathbf{X}$'s produced by $f(\cdot)$ in the model 4.1. The model for the observation may be complicated by the problem of detection limits of assays such as undetectable

lower and upper assay limits leading to scenarios similar to left or right censored data . This is the case for HIV RNA concentration here defined as the first observed component $(m = 1)$ of $g_m(.)$ without loss of generality. In this case we either observe Y_{ij1} or the event $\{Y_{ij1} < \zeta\}$, where ζ is the lower detection limit. The model can easily be generalized to upper detection limits or other detection limits depending on time.

4.3 Inference

4.3.1 Log-likelihood

Denoting $\delta_{ij} = I_{Y_{\{ij1}>\zeta\}}$, the latent full individual likelihood $\mathbf{L_i}(\mathbf{u})$ given the random effects $\mathbf{L_u}$ is given by:

$$\begin{aligned}\mathbf{L_i}(\mathbf{u}) = \Pi_{j\leq n_{i1}} &\left\{ \frac{1}{\sigma_1\sqrt{2\pi}} exp\left[-\frac{1}{2}\left(\frac{Y_{ij1} - g_1(\mathbf{X}(t_{ij1}, \tilde{\xi}^{(i)}))}{\sigma_1} \right)^2 \right] \right\}^{\delta_{ij}} \\ &\times \left\{ \Phi\left(\frac{\zeta - g_1(\mathbf{X}(t_{ij1}, \tilde{\xi}^{(i)}))}{\sigma_1} \right) \right\}^{1-\delta_{ij}} \\ &\times \Pi_{m>1, j\leq n_{im}} \left\{ \frac{1}{\sigma_m\sqrt{2\pi}} exp\left[-\frac{1}{2}\left(\frac{Y_{ijm} - g_m(\mathbf{X}(t_{ijm}, \tilde{\xi}^{(i)}))}{\sigma_m} \right)^2 \right] \right\}\end{aligned}$$

where Φ is the cumulative distribution function of the standard univariate normal distribution. The observed individual (marginal) likelihood L_{O_i} is obtained from $\mathbf{L_i}(\mathbf{u})$ as:

$$L_{O_i} = \int_{\mathbb{R}^q} \mathbf{L_i}(\mathbf{u})\phi(\mathbf{u})d\mathbf{u}, \tag{4.4}$$

where ϕ is the multivariate normal density $N(0, I_q)$. Denote $l_i(u) = log\mathbf{L_i}(\mathbf{u})$ and $l_{O_i} = logL_{O_i}$ the full (given random effects) and observed individual log likelihoods, respectively. The global observed log likelihood is $L_O = \Sigma_{i\leq n} L_{O_i}$. The integrand in (4.4) is centered and scaled as suggested for the adaptive Gaussian quadrature in Pinheiro and Bates [48]. The integral is then computed with an efficient algorithm developed by Genz and Keister [22].

4.3.2 Algorithm for Likelihood Maximization

The Newton-Raphson like method that uses only the first derivatives of the log likelihood (the score) was proposed for likelihood inference.

Computation of the score

There are two stages in computation of the score. The first stage is to compute the score of the full likelihood given random effects. The second stage is to compute the score of the observed likelihood by integration using the relationships given by Louis [33] and generalized by Commenges and Rondeau [12]. We assume that there is no censored data, for simplicity.

For subject i at the current point $\theta = ((\phi_l)_{l\leq p}, (\beta_l)_{l=1,p}$, $\mathbf{A} = (a_{ll'})_{l'\leq l\leq q}$, $\sigma = (\sigma_l)_{l\leq M})$, the components of the full score can be written as follows:

$$\begin{aligned}\mathbf{U}^{(\phi_l)}_{\mathbf{F_i}|\mathbf{u^{(i)}}}(\theta) &= \frac{\partial L_i(u)}{\partial \tilde{\xi}_l^{(i)}} \\ &= \sum_{m\leq M, j\leq n_{im}} \frac{1}{\sigma_m^2} \frac{\partial g_m(\mathbf{X}(t_{ijm}, \tilde{\xi}^{(i)}))}{\partial \tilde{\xi}_l^{(i)}} \times \left[Y_{ijm} - g_m(\mathbf{X}(t_{ijm}, \tilde{\xi}^{(i)}))\right],\end{aligned}$$

$$\begin{aligned}\mathbf{U}^{(\beta_\mathbf{l})}_{\mathbf{F_i}|\mathbf{u^{(i)}}}(\theta) &= \frac{\partial L_i(u)}{\partial \beta_\mathbf{l}} \\ &= \mathbf{z_l^{(i)}} \mathbf{U}^{(\phi_l)}_{\mathbf{F_i}|\mathbf{u^{(i)}}}(\theta),\end{aligned}$$

$$\begin{aligned}\mathbf{U}^{(a_{ll'})}_{\mathbf{F_i}|\mathbf{u^{(i)}}}(\theta) &= \frac{\partial L_i(u)}{\partial a_{ll'}} \\ &= \sum_{m\leq M, j\leq n_{im}} \frac{1}{\sigma_m^2}(Y_{ijm} - g_m(\mathbf{X}(t_{ijm}, \tilde{\xi}^{(i)}))) \times \left(u_{l'}^{(i)} \sum_{l''\leq p} w_{l''l}^{(i)} \frac{\partial g_m(\mathbf{X}(t_{ijm}, \tilde{\xi}^{(i)}))}{\partial \tilde{\xi}_{l''}^{(i)}}\right),\end{aligned}$$

$$\begin{aligned}\mathbf{U}^{(\sigma_l)}_{\mathbf{F_i}|\mathbf{u^{(i)}}}(\theta) &= \frac{\partial L_i(u)}{\partial \sigma_l} \\ &= \sum_{j\leq n_{il}} \frac{(Y_{ijl} - g_l(\mathbf{X}(t_{ijl}, \tilde{\xi}^{(i)})))^2}{\sigma_l^3} - \frac{n_{il}}{\sigma_l}.\end{aligned}$$

Using the fact that :

$$\frac{\partial g_m(\mathbf{X}(t, \tilde{\xi}^{(i)}))}{\partial \tilde{\xi}_l^{(i)}} = \sum_{k\leq K} \frac{\partial g_m(\mathbf{X}(t, \tilde{\xi}^{(i)}))}{\partial X^{(k)}} \frac{\partial X^{(k)}(t, \tilde{\xi}^{(i)})}{\partial \tilde{\xi}^{(i)}},$$

the computation of the full score requires to solve numerically the p systems of sensitivity equations $\frac{\partial X^{(k)}(t,\tilde{\xi}^{(i)})}{\partial \tilde{\xi}_l^{(i)}}$. Then the observed scores can be deduced by Louis' formula:

$$\mathbf{U}_{O_i} = \frac{\partial L_{O_i}}{\partial \theta} = (\mathbf{L}_{O_i})^{-1} \int_{\mathbb{R}^q} L_i(u)(\mathbf{u})\mathbf{U}_{\mathbf{F_i}}|\mathbf{u^{(i)}}(\mathbf{u})\phi(\mathbf{u})d\mathbf{u}.$$

The adaptive Gaussian quadrature can be used to compute the integrals, using the same transformation as for the computation of L_{O_i}. Then, the global observed score is $\mathbf{U} = \mathbf{U}_O = \sum_{i \leq n} \mathbf{U}_{O_i}$.

The Maximization Algorithm

The more robust Marquardt algorithm [36], or the Newton-Raphson method, is the most efficient algorithm when the log likelihood is not too far from the quadratic function [25]. This approach requires to compute the Hessian Matrix $\mathbf{H}$ and the score vector $\mathbf{U}$ at each current point θ_k of the maximization procedure. A semi analytical expression for the Hessian could be obtained with the same two-stages approach as for the score, though the computational burden would become unbearable. Therefore an iterative method can be used in which $H(\theta_k)$ is replaced by $G(\theta_k) = \sum_{i \leq n} \mathbf{U}_{O_i}(\theta_k)\mathbf{U}'_{O_i}(\theta_k) + \frac{\nu}{n}\mathbf{U}(\theta_k)\mathbf{U}'(\theta_k)$, where ν is a weighting coefficient. We have that $n^{-1}G(\hat{\theta})$ converges toward $n^{-1}I(\theta^*)$, where $I(\theta^*)$ is the information matrix under the true probability, θ^* being the true parameter value.

Thus $G(\theta_k)$ should be a good approximation of $H(\theta_k)$ near the maximum because $n^{-1}H(\hat{\theta})$ itself converges towards $n^{-1}I(\theta^*)$.

Thus one can use $C(\theta_k) = \mathbf{U}(\theta_k)'G^{-1}(\theta_k)\mathbf{U}(\theta_k)$, for convergence criterion. $C(\theta^*)$ has asymptotically a χ^2_p distribution; which gives an idea of which value should be considered as "small". One may use $G(\hat{\theta})$ as an estimator of $I(\hat{\theta}^*)$ to build confidence intervals and Wald tests, once the convergence is obtained.

We may expect that the variance of $\mathbf{U}$ computed at $\hat{\theta}$ to be a negatively biased estimate of its variance at θ^* (i.e., $I(\theta^*)$), because $\hat{\theta}$ is the value for which $\mathbf{U}(\hat{\theta}) = 0$. In general, this bias is very difficult to estimate. In the linear model with known error variance, it can be shown that $E[\mathbf{U}(\hat{B})\mathbf{U}'(\hat{B})] = \frac{n-dim(\beta)}{n}I(\beta^*)$, where β is the vector of regression coefficients. They proposed to estimate $I(\theta^*)$ by $\frac{n}{n-dim(\theta)}G(\hat{\theta})$ as an unbiased estimator of $I(\theta^*)$.

The whole algorithm (iteration and convergence criterion) has the property that it is invariant under any affine transformation of the parameters .

4.3.3 Expectations and Predictions

The expected trajectory can be obtained by simulating a sample of subjects and averaging, for each time and each marker, over their values.

Individual predicted trajectories can be computed as $\hat{\mathbf{X}}^{(i)}(t) = \mathbf{X}(t, \tilde{\hat{\xi}}^{(i)})$, where $\tilde{\hat{\xi}}_l^{(i)} = \hat{\phi}_l + \mathbf{z_l^{(i)\prime}}\hat{\beta}_\mathbf{l} + \omega_\mathbf{l}^{(i)\prime}\hat{\mathbf{A}}\hat{\mathbf{u}}^{(i)}$ and $\hat{\mathbf{u}}^{(i)}$ is the posterior mode (given the data) of $\mathbf{u}^{(i)}$. From this, the individual predicted trajectories of observed components can be deduced. The fit can then be checked by comparing the predicted values of the components $\hat{Y}_{ijm} = g_m(\hat{\mathbf{X}}^{(i)}(t_{ijm}))$ with the observations Y_{ijm}.

4.3.4 Statistical Analysis of the HIV and AIDS model from the ALBI ANRS 070 Data

The ALBI ANRS 070 trial is a longitudinal study which was carried out in France at Saint Louis hospital. This was an unblinded, randomized controlled trial in which 3 treatment regimens were compared [43].

The first measurement after therapy was performed four weeks later. The vector of natural parameters for subject i was then: $\xi^{(\mathbf{i})} = (\lambda^{(i)}, \alpha^{(i)}, \eta^{(i)}, \mu_T^{(i)}, \mu_{T^*}^{(i)}, \pi^{(i)})'$. The link functions were the log transform for all parameters (because they must be positive), except for $\eta^{(i)}$ for which the inverse logistic function was used (because $0 < \eta^{(i)} < 1$): $\tilde{\eta^{(i)}} = log\frac{\eta^{(i)}}{1-\eta^{(i)}}$.

Observable components g_1 and g_2 were transforms of HIV RNA concentration and total $CD4$ count, respectively, with $g_1 = log_{10}(V_I + V_{NI})$ and $g_2 = (Q + T + T^*)^{0.25}$. To achieve normality and homoscedasticity of measurement error distributions, these transformations of HIV marker's values are commonly used [56].

Chapter 5

Simulations

In this chapter a qualitative and quantitative study of the HIV dynamical model in system 3.1 through simulations is carried out. We do numerical simulations using the ode solver **ode45** that solves initial value problems for ordinary differential equations coded in Matlab. We used the parameters which were estimated in [25] and presented in Table 5.1 while the unestimable parameters were found from literature as shown in Table 5.2, for numerical simulations. The qualitative behaviour of the system was investigated by varying one parameter and holding others fixed at a time.

In [25] the parameters of the statistical model were estimated using repeated measurements of both the viral load and the total $CD4$ count from the ALBI ANRS 070 data on the transformed scales.

Table 5.1: Estimates of the model parameters and their standard deviation. ALBI ANRS 070 clinical trial

Parameters	Estimated Value	Standard deviation
$\tilde{\alpha}$	-3.16	0.15
$\tilde{\lambda}$	2.62	0.12
$\mu_{\tilde{T}^*}$	-0.40	0.11
$\tilde{\pi}$	4.64	0.12
$\mu_{\tilde{T}}$	-2.14	0.087
$\tilde{\eta_0}$	0.96	0.079
β	0.096	0.018
σ_α	0.31	0.025
σ_λ	0.043	0.0059
$\sigma_{\mu_{T^*}}$	0.25	0.028
σ_{CV}	0.42	0.012
σ_{CD4}	0.18	0.0050

Table 5.2: Description of model parameters which were fixed

Parameter	Description	Value	Reference
μ_Q	Death rate of Q cells	0.00014	[38]
μ_v	Clearance of free virions	30.0	[52]
ρ	Rate of reversion to the quiescent state	0.017	[53]
ω	Proportion of infectious virions	0.20	[47]

5.1 Simulations when varying the treatment efficacy η

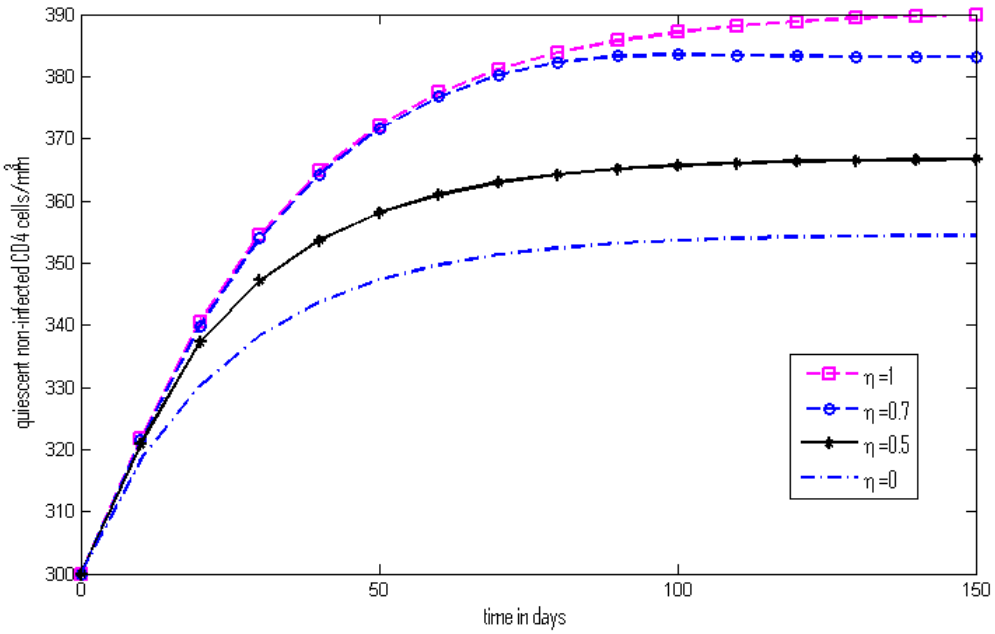

Figure 5.1: Comparison of quiescent non-infected $CD4$ cells at different values of η.

Figure 5.1 shows a plot of quiescent $CD4$ cells with time. We used estimated parameter values and the ones found in literature. For $\eta = 1$, $CD4$ count increases sharply up to around 50 days then it continues increasing at a slower rate. This is due to the fact that when $\eta = 1$, the treatment is 100% or perfectly efficient but in real life it is difficult to find such a treatment. For $\eta = 0.7$, the $CD4$ count increases rapidly in the first 50 days, just the same as for $\eta = 1$, but at around 75 days, we note a difference in the respective $CD4$ count levels. The $CD4$ count for $\eta = 0.7$ levels off at about 380 $CD4cells/mm^3$ and for $\eta = 1$, it levels off at about 390 $CD4cells/mm^3$. This shows that treatment with $\eta = 1$ is more efficient than the one for $\eta = 0.7$. The same thing happens for $\eta = 0.5$ and $\eta = 0$, but the rapid increase in the $CD4$ count is up to 30 days then it increases at a lower rate and it finally levels off. For $\eta = 0.5$ the $CD4$ count levels off at around 360 $CD4cells/mm^3$ and $\eta = 0$ the $CD4$ count levels off at around 350 $CD4cells/mm^3$. This graph makes sense because when the treatment efficacy levels (η levels) decrease, the level at which the $CD4$ cells levels off at lower levels. It should also be noted that for quiescent $CD4$ cells the levels remain high > 200 cells$/mm^3$ irrespective of the treatment efficacy η.

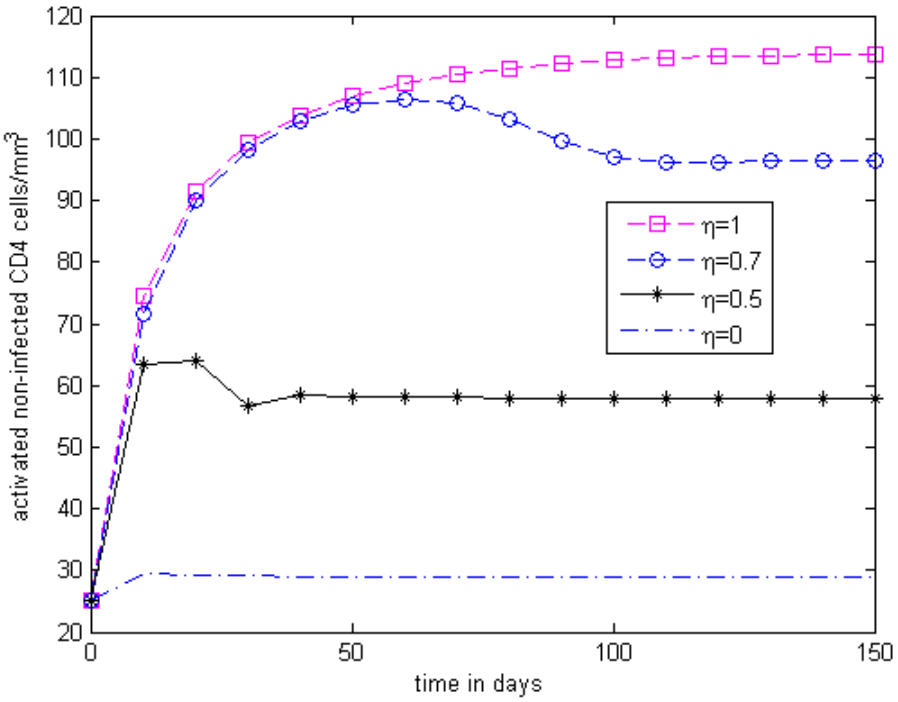

Figure 5.2: Comparison of activated non-infected $CD4$ cells at different values of η.

Figure 5.2 shows the change in activated non-infected $CD4$ cells over time at different levels of η. The simulation results show that non-infected $CD4$ cells increase rapidly in the first 20 days post treatment when $\eta = 1$ and then start increasing at a slow rate thereafter. For $\eta = 0.7$ the activated non-infected $CD4$ cells ultimately settle at a level higher than when $\eta = 0.5$ but lower than that for $\eta = 1$. The set point is lowest when $\eta = 0$ as expected.

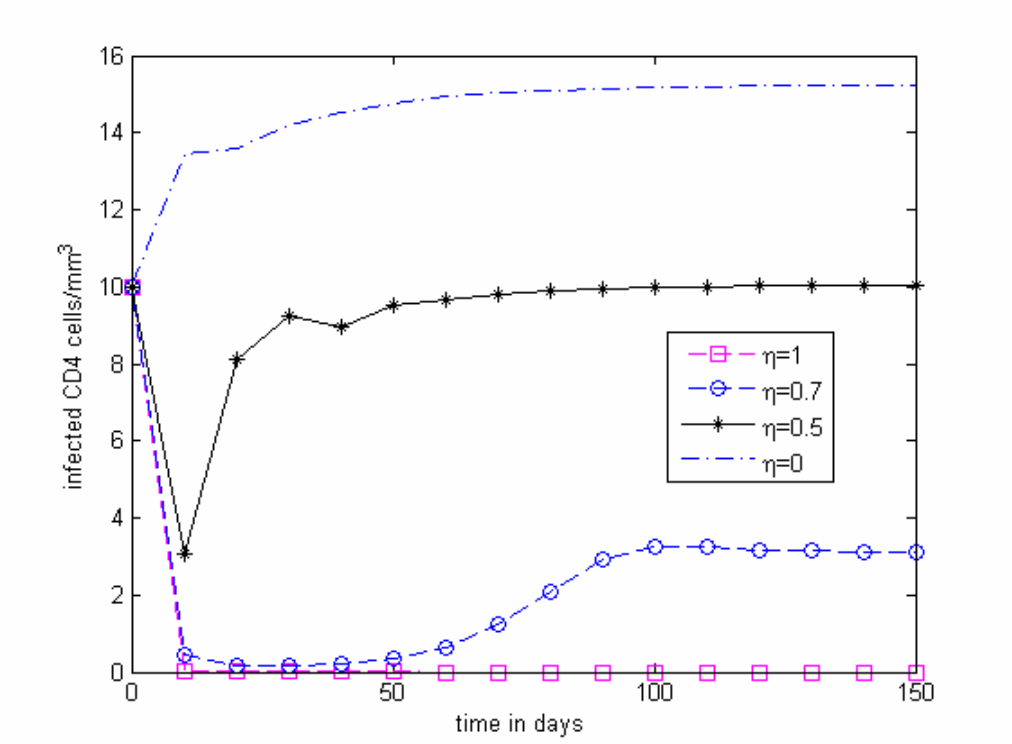

Figure 5.3: Comparison of activated infected $CD4$ cells at different values of η.

Figure 5.3 shows the change in activated infected $CD4$ cells over time at different levels of η. For $\eta = 1$, the treatment is very efficient such that there is a rapid decline of infected $CD4$ cells. The infected $CD4$ cells vanished in 10 days on average. Also for $\eta = 0.7$, there is a rapid decline of infected $CD4$ cells but they do not reach zero. However in this case the infected $CD4$ cells start increasing from 50 days and stay constant from 100 days onwards. This shows that the treatment with $\eta = 0.7$ is not sufficient to suppress the infected $CD4$ cells to zero permanently because after sometime the infected $CD4$ cells increase. As for a treatment, from this observation we may infer that the infected individual needs a treatment boost at this point or prescription to a new effective treatment at day 50. The reason for changing to another type of ARV could be drug resistant related or due to side effects from the first treatment. For $\eta = 0.5$, the infected $CD4$ cells decrease rapidly to $3/mm^3$ for the first 10 days but they increase rapidly up to around 30 days and they stay constant thereafter. This means that infected $CD4$ cells cannot die out if the efficacy of treatment is 0.5. For $\eta = 0$, there is an increase in infected $CD4$ cells in the first 10 days and they stay constant thereafter at a higher level than for any other efficacy levels as

expected (see 5.3).

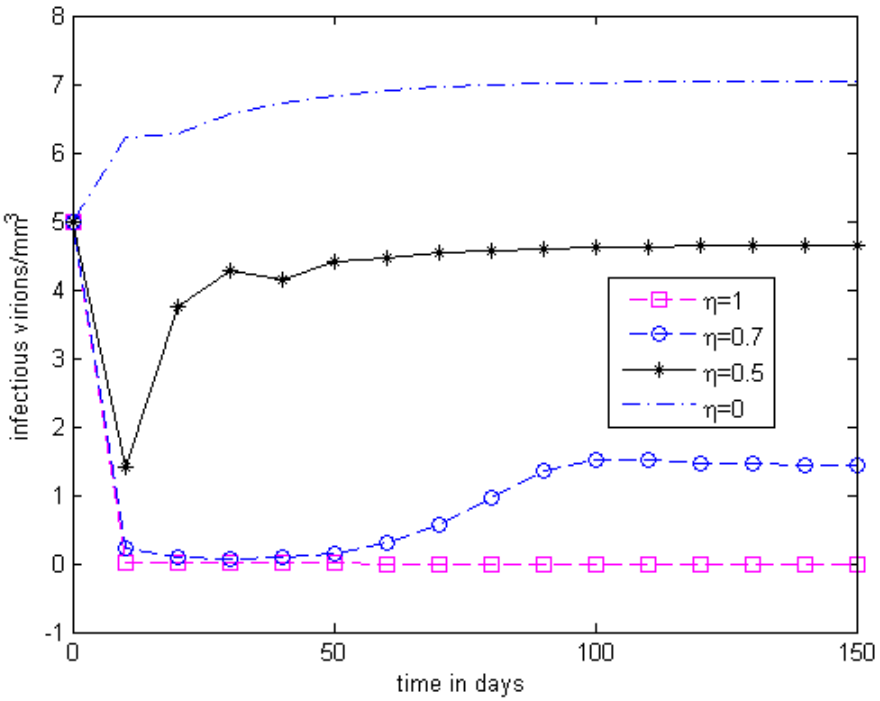

Figure 5.4: Comparison of infectious virions at different values of η.

Figure 5.4 is a plot of total infectious virions (V_I) over time for different levels of η. For $\eta = 1$, the infectious virions decrease rapidly up to zero. This makes sense because if treatment efficacy is one, all infectious virions are going to die off or be eliminated. This is due to the fact that there won't be any infected $CD4$ cells to produce new virions. If $\eta = 0.7$, the virions decrease but they do not reach zero. They start increasing at 50 days. Note that this is the samepoint where infected $CD4$ cells began to increase for $\eta = 0.7$. This may suggest the introduction of a treatment boost or a new ARV to sustain infectious virions suppression. The same happens for $\eta = 0.5$ but the infectious virions start increasing at an earlier time point (i.e 10 days), so another type of ARV or a treatment boost maybe given to the patient at 10 days if possible, so as to ensure the infectious virions remain suppressed.

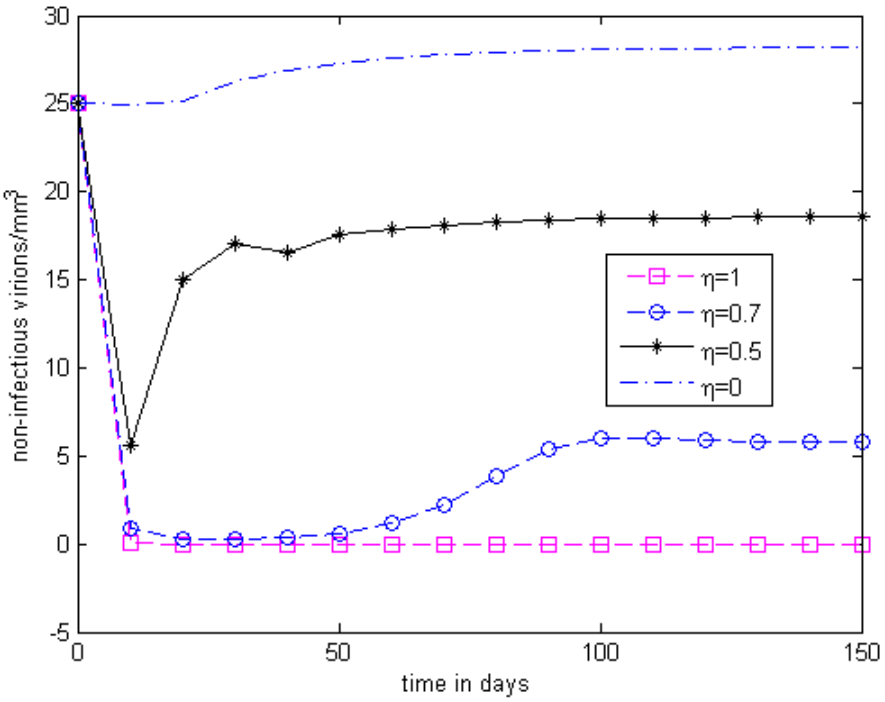

Figure 5.5: Comparison non-infectious virions at different values of η.

Figure 5.5 is a plot of total non-infectious virions (V_{NI}) over time for different levels of η. The same changes experienced by infectious virions are experienced by non-infectious virions. Note that there are a lot more of non-infectious virions than infectious virions because they are predominant in the body system.

5.2 Simulations when varying the $CD4$ cell production rate λ

For η fixed at $\eta = 0.3$ we have the following graph Figure 5.6 for quiescent cells with different values of λ.

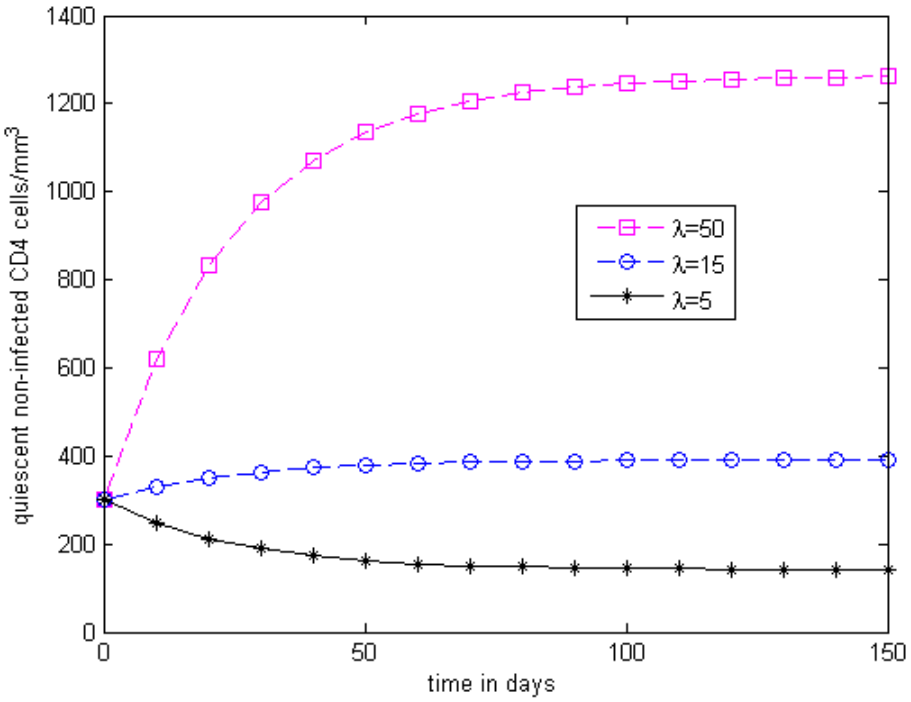

Figure 5.6: Comparison of quiescent non-infected $CD4$ cells at different values of λ.

The higher the rate of quiescent cells production (λ), the higher the quiescent cells (Q) as expected.

Figure 5.7 is a plot of activated non-infected $CD4$ cells, infected $CD4$ cells, infected virions and non-infected virions over time with η and λ fixed at $\eta = 0.3$ and $\lambda = 5$ (both low).

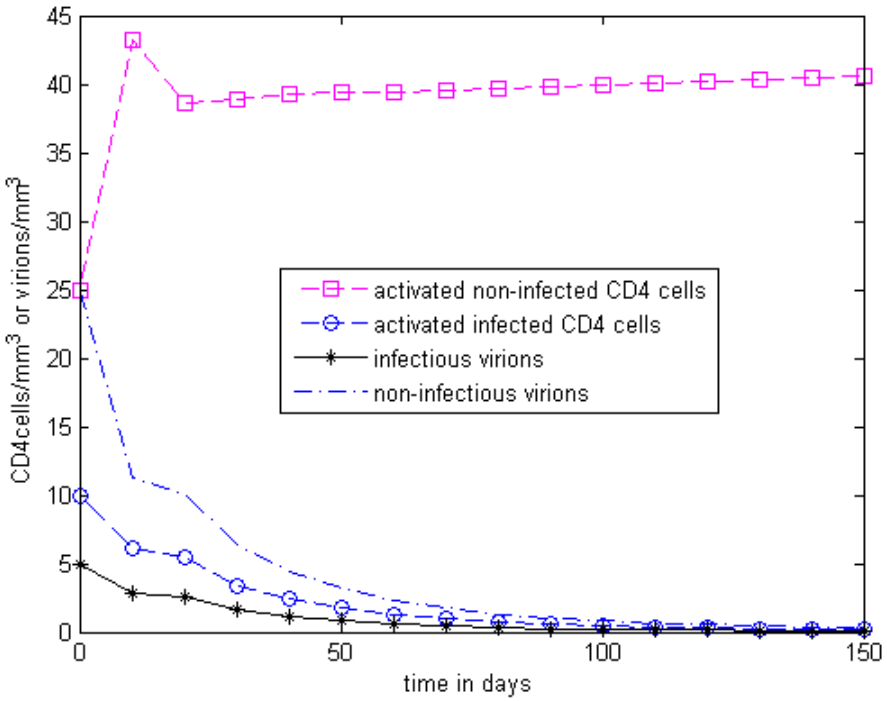

Figure 5.7: The plot showing the change in total virions or $CD4$ cells with time for $\lambda = 5$.

If $\lambda = 5$ and low efficacy of treatment $\eta = 0.3$, the infected $CD4$ cells (T^*) get reduced dramatically. This is because few cells are getting produced. Thus there are few susceptible cells to be infected. Thus infected cells level off at a lower values.

Figure 5.8 is similar to Figure 5.7 but now λ is raised to $\lambda = 15$.

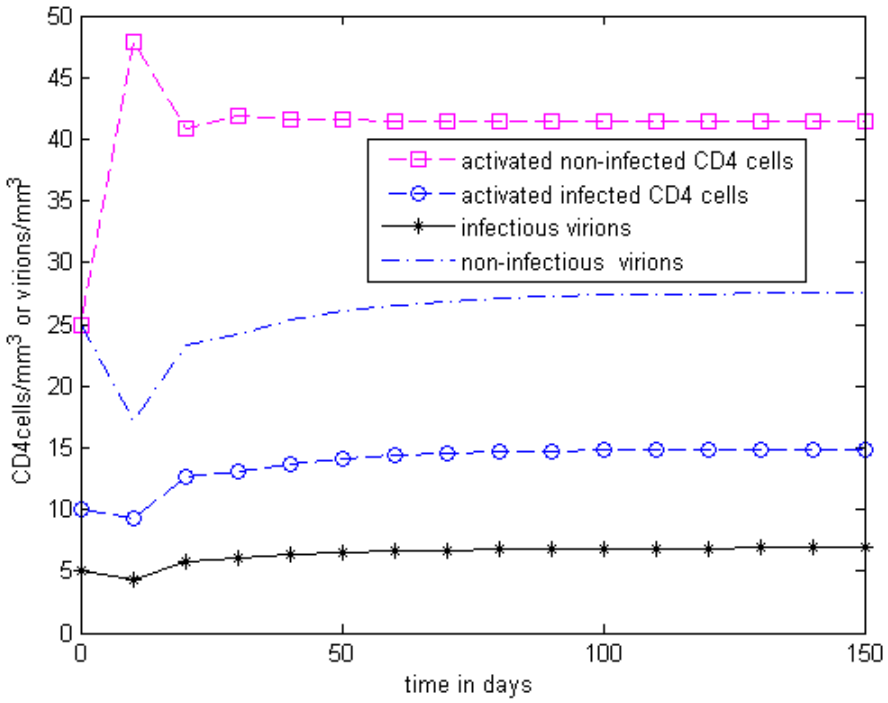

Figure 5.8: The plot showing the change in total virions or $CD4$ cells with time for $\lambda = 15$.

For $\lambda = 15$ and low efficacy of treatment $\eta = 0.3$, a lot of cells get infected. This leads to higher levels of infectious virions and infected $CD4$ cells in the body.

In Figure 5.9, λ is raised to an even higher value $\lambda = 50$ but η is still low at $\eta = 0.3$.

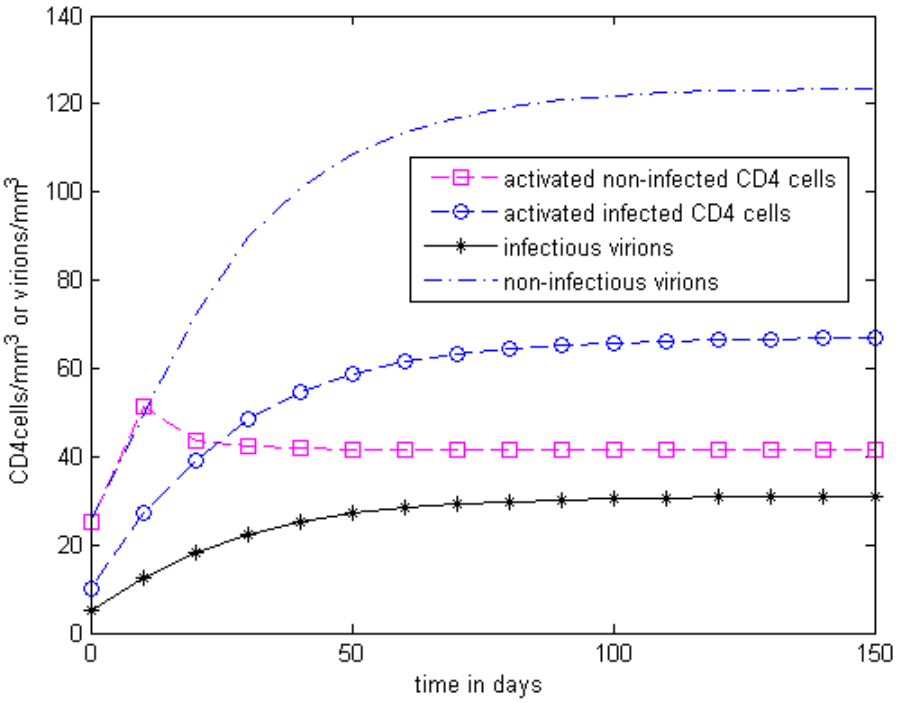

Figure 5.9: The plot showing the change in total virions or $CD4$ cells with time for $\lambda = 50$.

For $\lambda = 50$ and low efficacy of treatment, infected $CD4$ cells increase to a higher level than activated uninfected $CD4$ cells. This will lead to worse disease status than lower value of λ with $\eta = 0.3$ (low treatment efficacy).

In Figure 5.10 a simulation of quiescent $CD4$ cells is performed over time for high treatment efficacy $\eta = 0.8$ and λ varied from $\lambda = 5, 15, 50$.

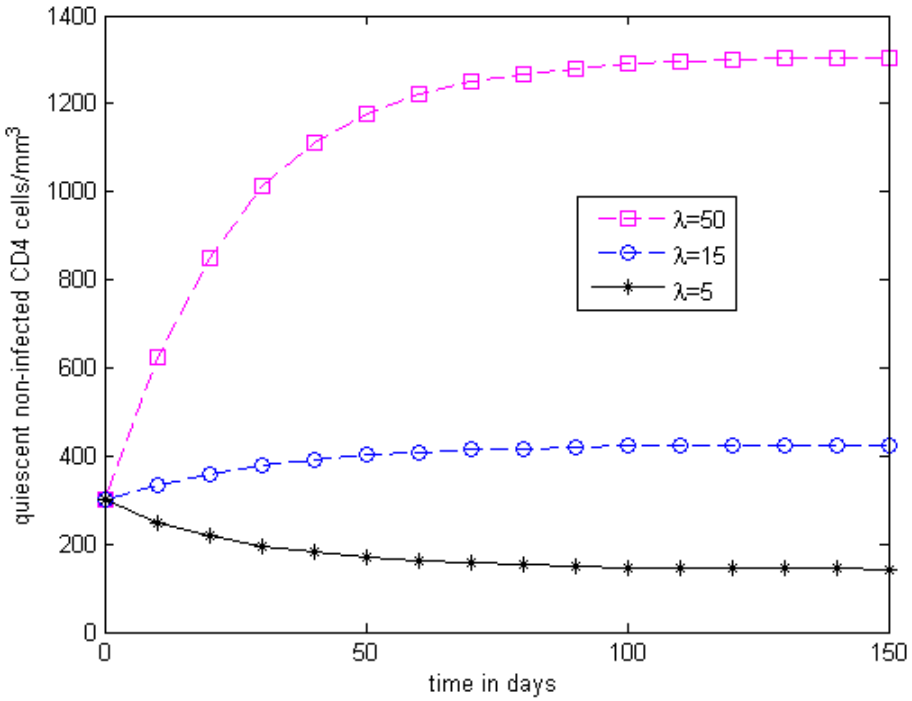

Figure 5.10: Comparison of quiescent non-infected $CD4$ cells at different values of λ.

If we compare the figures 5.6 and 5.10, for the change in quiescent cells with time, employing different values of η (low and high treatment efficacies), the figures are qualitatively the same. This means that η does not affect quiescent cells. This makes sense because quiescent cell production is an individual specific characteristic not affected by treatment. Some individuals may exhibit a low quiescent cell production rate and others a high rate. Thus in the treatment of HIV AND AIDS generalisation of treatment regimes and strategies among individuals may not yield same results on all patients. Thus there is a need to understand individual to individual heterogeneity in designing treatment strategies.

Figure 5.11 shows a plot of different types of population cells over time for a high treatment efficacy ($\eta = 0.8$) and low λ ($\lambda = 5$).

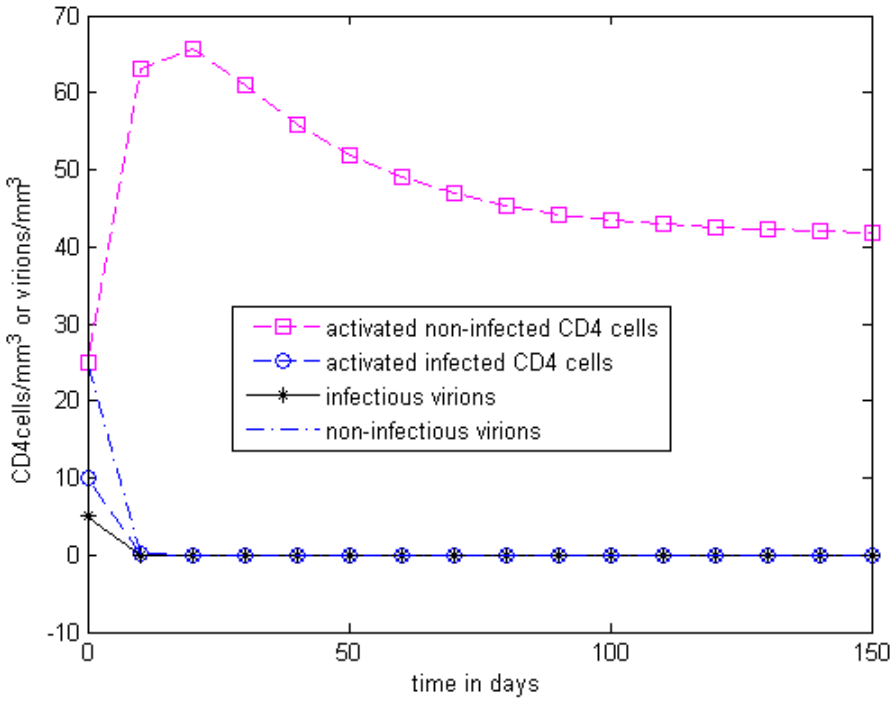

Figure 5.11: The plot showing the change in total virions or $CD4$ cells with time for $\lambda = 5$.

For a low rate of quiescent cells production ($\lambda = 5$) and high treatment efficacy, the infected $CD4$ cells, non-infectious virions and infectious virions are suppressed.

Figure 5.12 is a plot of the different cell population over time and intermediate λ.

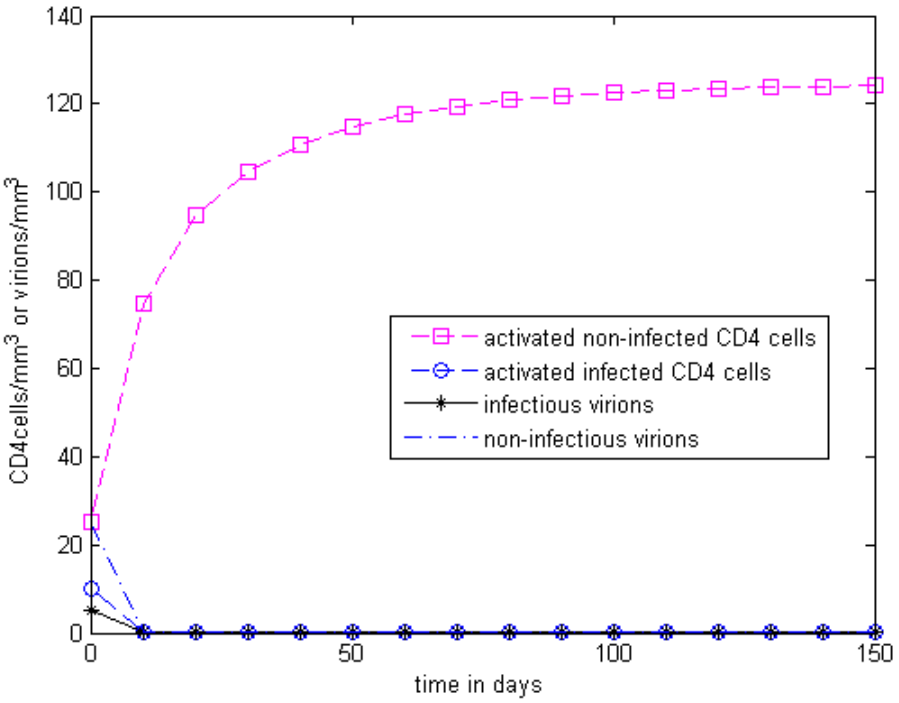

Figure 5.12: The plot showing the change in total virions or $CD4$ cells with time for $\lambda = 15$.

For an average rate of quiescent cells production ($\lambda = 15$), the infected $CD4$, the infected cells, non-infectious virions and infectious virions are still suppressed. The activated non-infected cells increase and level off at a higher level than for the graph for $\lambda = 5$.

Figure 5.13 is a plot of the four cell population over time but now quiescent cell production is higher ($\lambda = 50$).

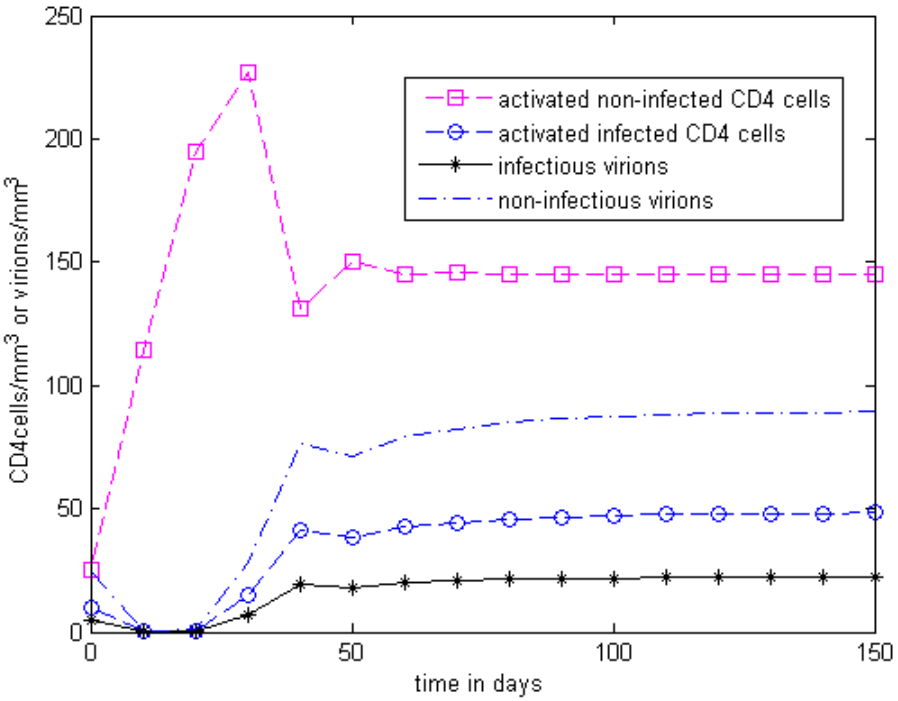

Figure 5.13: The plot showing the change in total virions or $CD4$ cells with time for $\lambda = 50$.

For a high rate of quiescent cell production ($\lambda = 50$) and high treatment efficacy, the infected $CD4$ cells increase and level off at around $50\ CD4cells/mm^3$. The infected $CD4$ cells get suppressed though they do not remain suppressed throughout time as was with the case for $\lambda = 5, 15$. We also note that infectious virions stabilize at higher levels than when $\lambda = 5, 15$ for η fixed at $\eta = 0.8$. Thus one may infer that high treatment efficacy is advantageous for individuals with a controlled production rate λ for quiescent $CD4$ cells production.

5.3 Simulations when varying the infectious virion production rate ω

For η at $\eta = 0.3$ and other parameters fixed as estimated in [25], we have the following graph (Figure 5.14) for quiescent cells with time for different values of ω, where ω is the proportion of infectious virions.

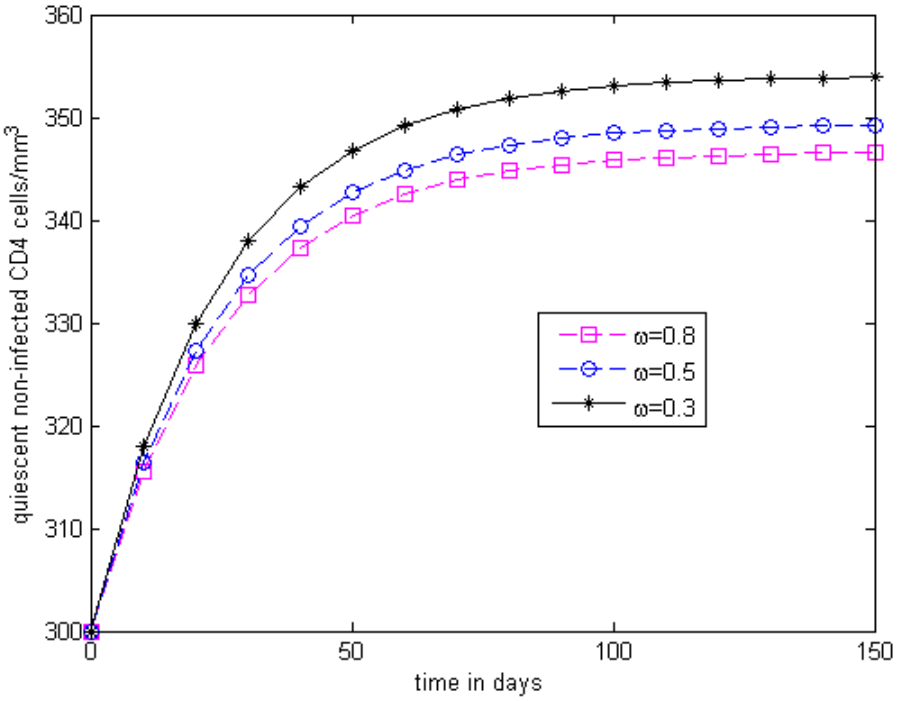

Figure 5.14: Comparison of quiescent non-infected $CD4$ cells at different values of ω.

The higher the proportion of infectious virions the lower the total $CD4$ cells (quiescent cells) in the body system. This is probably due to the fact that a lot of quiescent cells get activated and then infected by these infectious virions. In addition the treatment efficacy is very low, hence allowing the infectious virions to be more effective in infecting $CD4$ cells, i.e allowing them to infect a lot of cells.

Figure 5.15 below shows a plot of the other cell population when η is set at $\eta = 0.3$ and $\omega = 0.3$ (both are low).

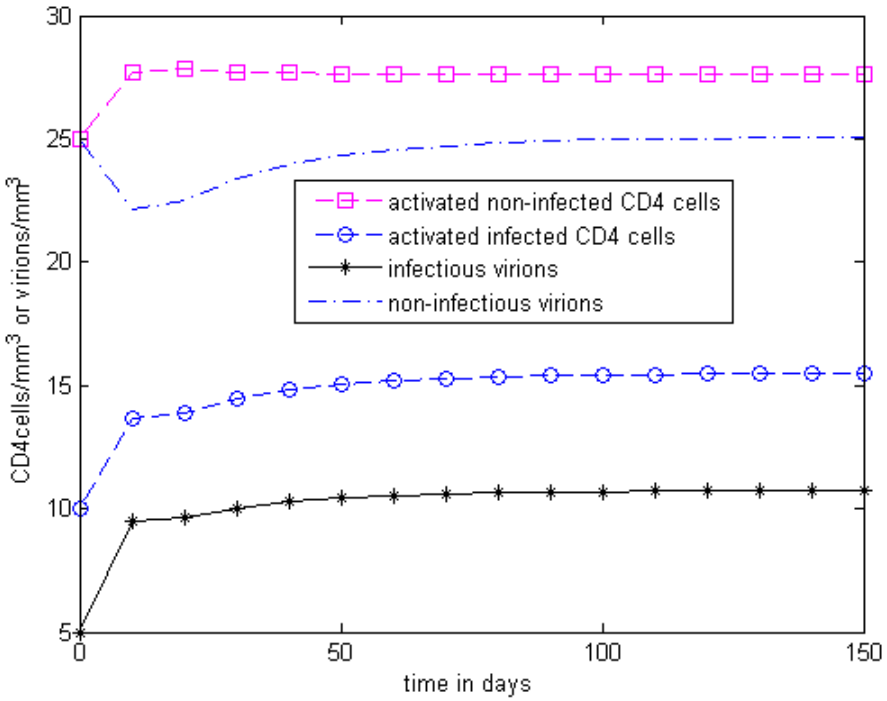

Figure 5.15: The plot showing the change in total virions or $CD4$ cells with time for $\omega = 0.3$.

Even though the treatment efficacy is low, the proportion of infectious virions is very low ($\omega = 0.3$), such that a minimum amount of $CD4$ cells get infected and they level off at a lower value.

In Figure 5.16 the proportion of infectious virions is raised to $\omega = 0.5$.

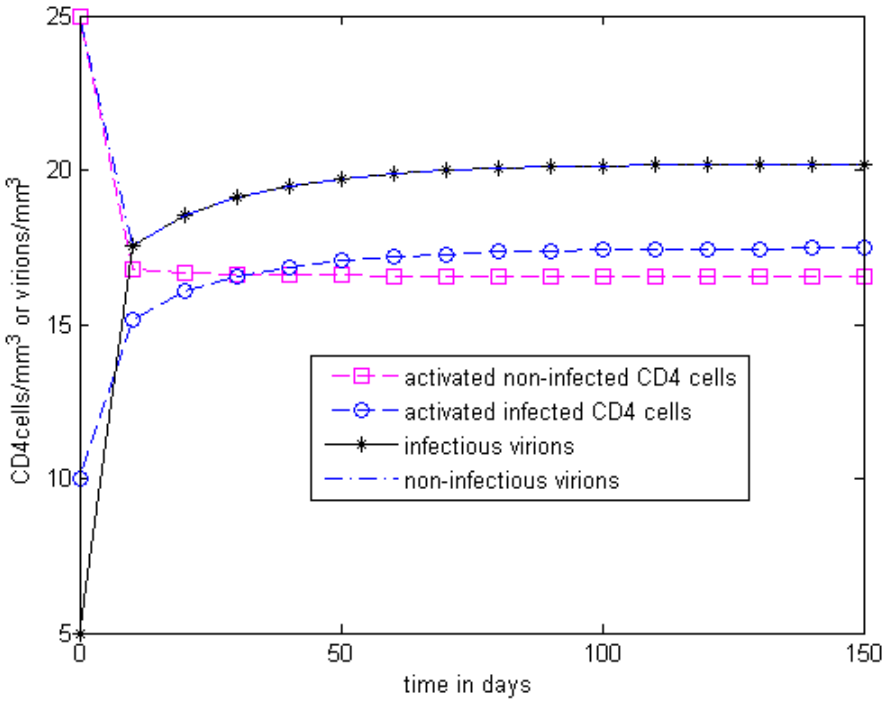

Figure 5.16: The plot showing the change in total virions or $CD4$ cells with time for $\omega = 0.5$.

Now with $\omega = 0.5$, a lot more $CD4$ cells get infected. Activated non-infected $CD4$ cells get reduced to a lower value than infected $CD4$ cells compared to the scenario in Figure 5.15.

In Figure 5.17 ω is raised further to $\omega = 0.8$ but η is still maintained at $\eta = 0.3$.

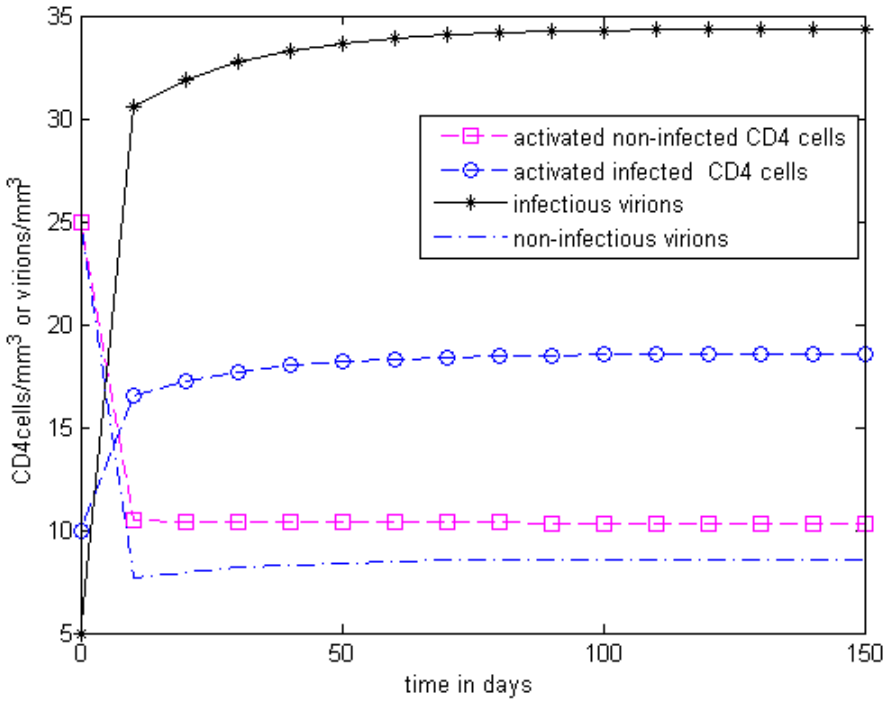

Figure 5.17: The plot showing the change in total virions or $CD4$ cells with time for $\omega = 0.8$.

If $\omega = 0.8$, the proportion of infectious virions is high such that a lot of $CD4$ cells get infected, thereby reducing the activated $CD4$ cells. Also, the treatment efficacy is very low, making the situation even more destructive to the immune system.

Figure 5.18 now plots the change in quiescent $CD4$ cells over time at different levels of ω when η is set at $\eta = 0.8$.

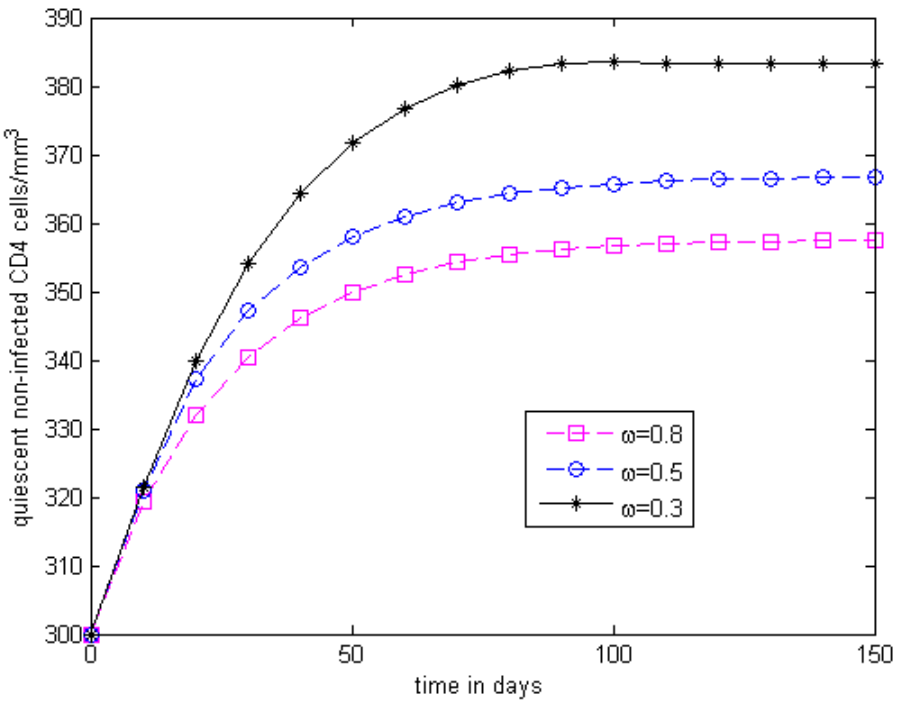

Figure 5.18: Comparison of quiescent non-infected $CD4$ cells at different values of ω.

If $\eta = 0.8$, the treatment is more efficacious such that when it combines with a lower ω, like $\omega = 0.3$, the $CD4$ cells (quiescent cells) stay at a relatively high level. This is because of the fact that if the proportion of infectious virions is low, a small amount of $CD4$ cells get infected. But if the proportion of infectious virions is high, even if η is high, a lot of $CD4$ cells get infected and the quiescent cells are reduced.

Figure 5.19 is a plot of the other four cell production over time when η is set at $\eta = 0.8$ (high) and $\omega = 0.3$ (low).

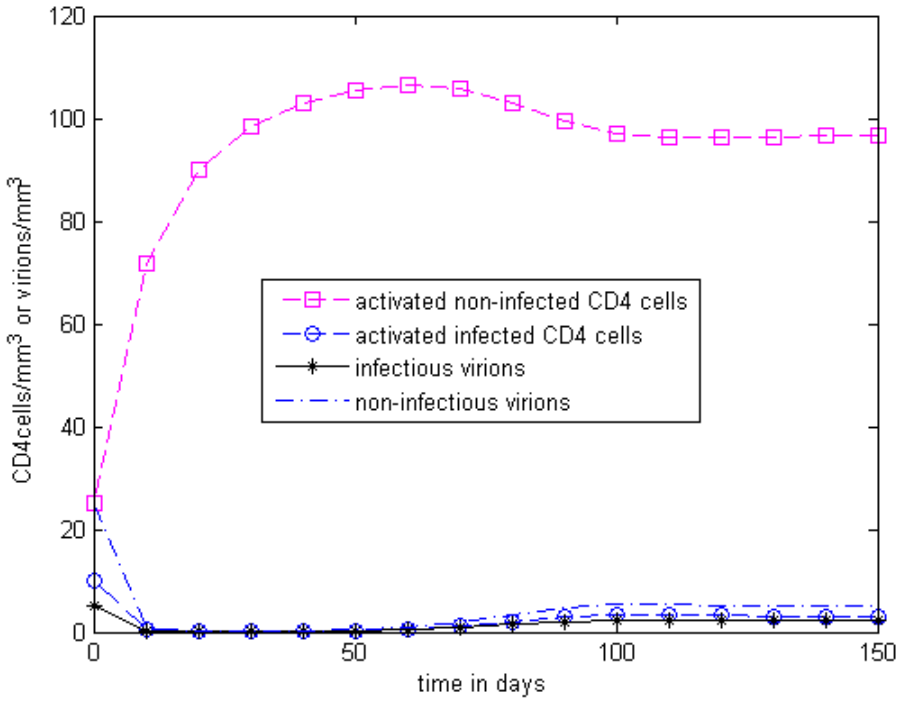

Figure 5.19: The plot showing the change in total virions or $CD4$ cells with time for $\omega = 0.3$.

The proportion of infectious virions is very low ($\omega = 0.3$), and η is high, this promotes the suppression of the infected $CD4$ cells and non-infectious virions, with activated non-infected $CD4$ cells leveling off at a high level.

In Figure 5.20 the level of infectious virions is increased to $\omega = 0.5$ while η is maintained at $\eta = 0.8$. Now the proportion of infectious and non-infectious virions is the same ($\omega = 0.5$).

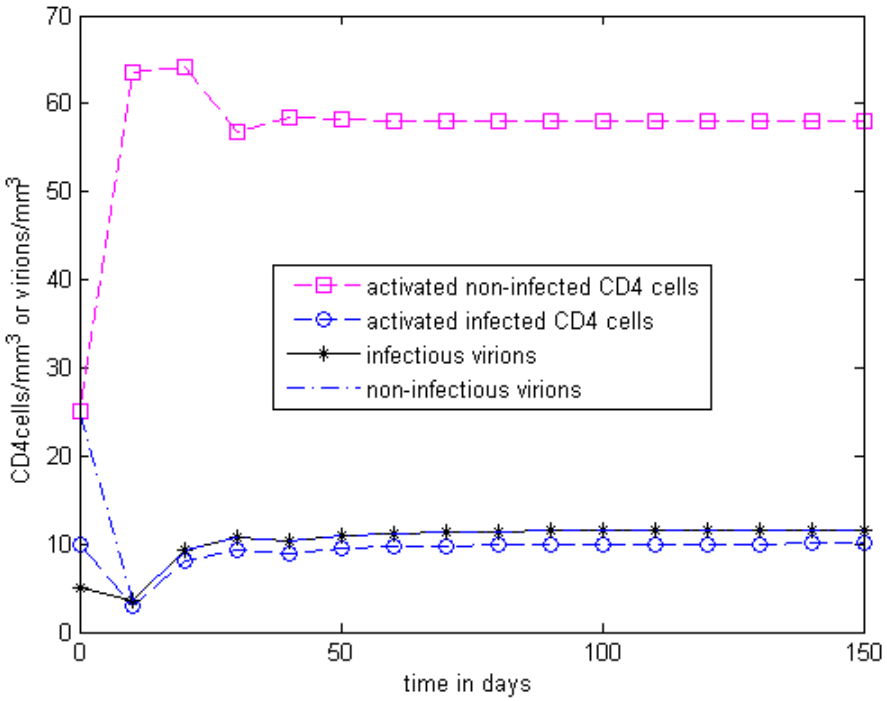

Figure 5.20: The plot showing the change in total virions or $CD4$ cells with time for $\omega = 0.5$.

We see that the activated $CD4$ non-infectious cells get reduced as compared to the case where $\omega = 0.3$ because now there is a higher proportion of infectious virions, implying that more activated $CD4$ cells would get infected. However now η is high therefore controlling the $CD4$ cells from getting infected.

Figure 5.21 is a plot of activated non-infected cells, infected cells, infectious and non-infectious virions. When both η and ω are high $\eta = \omega = 0.8$, but working antagonistically.

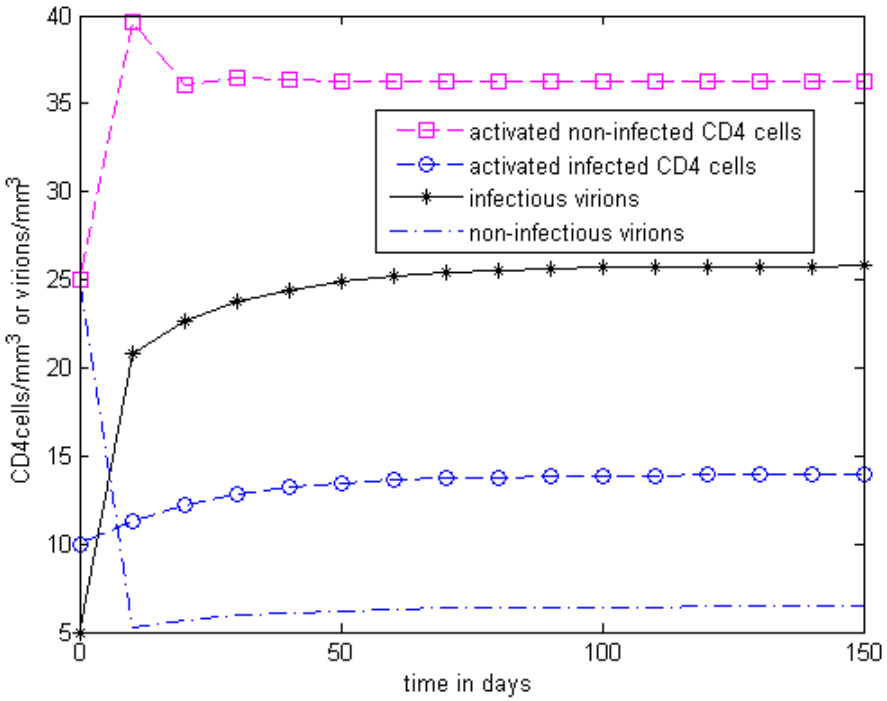

Figure 5.21: The plot showing the change in total virions or $CD4$ cells with time for $\omega = 0.8$.

The treatment efficacy is very high, thereby blocking the production of a lot of infected $CD4$ cells, but at the same time the proportion of infectious virions is very high. Some of these infectious virions get blocked from infecting the $CD4$ cells but the treatment efficacy is not one, therefore this allows some cells to get infected but nonetheless their level remains suppressed.

5.4 Discussion

In this section, we presented some simulations on an existing HIV and AIDS model. The interaction of the immune system and HIV consists of quiescent non-infected $CD4$ cells, activated non-infected $CD4$ cells, infected $CD4$ cells, infectious virions and non-infectious virions. From the simulations we found that the higher the treatment efficacy, the lower the infected cells present in the body. But as long as the treatment is not 100% efficacious, the infected cells remain or persist in the body. The higher the treatment efficacy, the more suppressed the infected cells are.

When we varied ω, the proportion of infectious virions, we discovered that infected cells tend to increase substantially if the treatment efficacy is low (e.g. $\eta = 0.3$) and the proportion of infectious virions is high (e.g. $\omega = 0.8$). For the same proportion of infectious virions and higher treatment efficacy (e.g. $\eta = 0.8$), the infected cells increase but are suppressed to a lower level as compared to the simulations with lower treatment efficacy. This has an implication when considering individuals specific dynamics. Certain individual can have a very high ω (proportion of infectious virions) naturally and another individual can have a very low one. Thus we deduce that due to individual heterogeneity the amount of $CD4$ count in the body can vary from patient to patient despite being given the same form of treatment. It seems more advantageous to have a low ω because it means one will produce few infectious virions to infect healthy cells. We can also deduce that individuals react differently to treatments depending on individual makeup. Even though this is the case, high treatment efficacy is good for any individual.

One of the limitations here, is that it was not possible to estimate all parameters from available data in [25]. Also, our λ (the production rate of quiescent cells) was constant, resulting in graphs not showing the AIDS stage. In future work, it would be better to show a varying λ so as to have graphs which have the AIDS stage. One approach to achieve this is to make $\lambda = \lambda(t)$ as a declining function of time. This better represent reality because now as the disease progresses production of quiescent non-infected cells decline. However one can also think of a scenario where there is a decline and then a restoration to near its original level to signify a high treatment efficacy and quality. The advantage of the model adopted from [25] is that we are able to model and simulate the process where the sub-population of quiescent non-infected $CD4$ cells is included

which basically represents reality.

Chapter 6

Causal Inference with Application to Modelling HIV Related Mortality

Medical practitioners are often faced with the question of whether the $CD4$ count and viral load are good biomarkers/surrogate endpoints for HIV AND AIDS. In the dynamical mathematical model 3.1 in chapter 3, these two biomarkers were used as state variables. But if they are not good surrogate markers, we would not get proper insights about the dynamics of the disease (HIV AND AIDS). Only medical trials and immunological studies backed with sound statistical methods can help identify good surrogate markers. In this chapter we will review methods for surrogate marker validation for a single trial. However the application is constrained by the type of data available. We also review the Generalized estimating equation approach because that is the type of linear model used to assess causality. The data used in the current work CAPRISA CAT study. The data consists of longitudinal measured disease markers for HIV namely viral loads and $CD4$ counts. In the current analysis the disease markers are used as predictors of survival or death, a binary outcome.

6.1 Surrogate Endpoints/Markers

Surrogate endpoints are measures that can be collected in a shorter time period and/or using fewer subjects than those normally considered in a classical clinical trial (e.g survival).The purpose of a surrogate endpoint is to draw conclusions about the effect of intervention on true endpoint without having to observe the true endpoint. Because there is a need to evaluate treatment benefits as fast as possible on easily measurable endpoints, this has always been a preoccupation in clinical research. In general, surrogate endpoint is prespecified as being of primary interest and serves to determine the significance of any observed treatment benefit. Surrogate endpoints are typically proposed based on biological considerations with a certain progression model of disease. They should be most clinically relevant, but consideration of time and cost may force the investigators to use some other surrogate endpoints instead. Examples of surrogates are $CD4$ counts in AIDS. The $CD4$ count in this case can potentially serve as a surrogate endpoint for death. If we consider disease recurrence after surgical removal of early cancers; here disease recurrence can serve as a surrogate endpoint to death. Tumor shrinkage can serve as a surrogate endpoint for survival in advanced cancer studies. Progression to AIDS can serve as a surrogate endpoint for death. Sometimes the surrogate endpoint is merely a biological marker of the disease process leading to the final endpoint, however, in some cases, the surrogate endpoint directly affects the patient's condition and is therefore itself of clinical relevance.

The validity of using one endpoint as a surrogate for another has been raised and studied over the last few years, while the practice of looking at multiple endpoints is not recent in clinical research. The dramatic surge of the AIDS epidemic, the impressive therapeutic results obtained early on with zidovudine drug, and the pressure for an accelerated evaluation of new therapies have all played a major role in focusing attention to the need for a formal definition of surrogate endpoints along with practical methods to validate them [35]. In the research of cardiovascular disease, there was an unsettling discovery that the two major antiarrhythmic drugs encanaide and flecanaide reduced arrhythmia but caused more than threefold increase in overall mortality. This necessitated the need for caution in using non validated surrogate markers in the evaluation of the possible clinical benefits of new drugs [8].

While many would like to avoid surrogate endpoints, sometimes surrogates will be the only reasonable alternative, especially when the true endpoint is rare and/or distant in time. It is then best to use validated surrogates. An application trial is a trial in which the surrogate, but not the true endpoint is observed and used to evaluate the effect of intervention on true endpoint. Before a surrogate endpoint can be confidently used in an application trial, it must be validated using a validation trial in which both the surrogate and true endpoints are observed [4]. In a general sense, a surrogate endpoint is validated if the surrogate endpoint approach and an approach using the true endpoint yield similar conclusions about the effect of intervention on the true endpoint in a validation trial. The problem is that it may be difficult to decide what levels of these measures (e.g proportion of treatment effect or PTE) indicate an appropriately validated surrogate endpoint.

6.1.1 Justification

We need to know the best statistical methods of validating the accuracy of surrogate markers to a disease outcome. This is because there is a need to evaluate treatment benefits as fast as possible on easily measurable endpoints. Also if we can manage to get accurate markers for a disease, this can help us in understanding the pathogenesis of the disease. This can help to better design control and treatment strategies. Accurate markers can also save lives if the true endpoint such as death can be detected in advance before the actual event occurs. Since death is irreversible, once it occurs we can no longer help the patient.

6.2 Methods for Validating a Surrogate Marker

In this chapter the emphasis will be on surrogate marker validation. Methods to evaluate treatment effects are many, so we chose to review methods of surrogate marker validation from statistical point of view in a single trial. Single arm trial is used to assess treatment benefit in the CAPRISA CAT study. We strongly believe that validation or understanding of key surrogate markers for a disease can help understand disease progression and therefore making modelling

better.

6.2.1 Prentice Criterion

Statistical methods for validation of surrogate markers can be dated back to Prentice [49]. The statistical definition of a surrogate marker, by Prentice, requires that the conditional distribution of the clinical outcome given the surrogate marker alone is the same as the conditional distribution of the clinical outcome given the surrogate marker and treatment. The Prentice criterion ensures the rejection of the null hypothesis as of no effect if intervention on the surrogate endpoint implies rejection of the null hypothesis of no effect of intervention on the true endpoint. Let S denote the putative surrogate marker, Z the (randomized) treatment, and Y the outcome of interest. For simplicity, we assume that both S and Y are both measured once, at fixed times, with S measured before Y, and the data on S or Y are never missing (e.g., due to dropout). Prentice [49] proposed these operational criteria for validation of S as a "true surrogate":

1. S must be affected by Z, i.e. treatment has an impact on the surrogate endpoint,
2. Y must be affected by Z, i.e. treatment has an impact on the true endpoint,
3. S must be correlated with the true outcome Y, i.e. the surrogate endpoint has an impact on the true endpoint.
4. The outcome and treatment should be conditionally independent given S (the full effect of treatment upon the true endpoint is captured by the surrogate); i.e,

$$Y \perp Z \mid S. \tag{6.1}$$

We can model Y given Z directly:

$$g\{E(Y \mid Z)\} = \beta_0 + \beta_Z Z; \tag{6.2}$$

where g denotes the link function for a generalized linear model. For example, $g(w) = logit(w)$ for logistic regression and and $g(w) = w$ for linear regression. β_Z represents the overall effect of

treatment on the clinical outcome. We can also model the "effect" of Z conditional on S:

$$g\{E(Y \mid Z, S)\} = \gamma_0 + \gamma_Z Z + \gamma_S S. \tag{6.3}$$

Here the expression γ_Z is viewed as the "effect" of Z controlling for the putative surrogate S. Thus S is considered to be a true surrogate in the Prentice framework only if $\gamma_Z = 0$. Criteria 1 and 2 are Prentice's original criteria. It is clear from (6.2) that generalized linear models and their extension to correlated data play a major role in the assessment of the adequacy of surrogate markers. This is the approach adopted in this thesis.

6.2.2 Proportion of Treatment Effect (PTE)

Freedman et al [19]., proposed focusing attention on the proportion of the treatment effect explained by the surrogate. This method is for one surrogate marker. A good surrogate is one that explains a large proportion of that effect. The PTE explained by a surrogate marker was defined as

$$PTE = 1 - \frac{\gamma_Z}{\beta_Z}. \tag{6.4}$$

The PTE, compares the regression coefficient with and without adjusting for the surrogate marker S. Its advantages are:

1. One might be prepared to use the biomarker as a surrogate endpoint in the next study of a similar type of drug, if the PTE is close to 1.

2. The PTE may be useful to compare two potential surrogate endpoints [62].

Note that a true surrogate according to Prentice [49], implies PTE$= 1$.

Disadvantages of PTE

1. PTE is not bounded,

2. It is based on data from a single trial,

3. The variability in the estimator of γ_Z is large when there is strong collinearity between Z and S (to obtain a 95% confidence interval with a reasonable width may require as many patients as a trial using the clinical endpoint [32]).

4. The PTE becomes ill-defined when, for example, an intervention in HIV reduces HIV-related mortality but is responsible for additional death due to other causes [17].

Due to the fact that the treatment is generally effective in improving the surrogate marker, a strong correlation between treatment and the surrogate marker might be expected. The PTE that is near 1 is insufficient for inference that the surrogate is a valid one, although a value close to 0 indicates an invalid surrogate [17]. PTE combines 3 quantities (adjusted association, relative effect, and the ratio of the variances, conditional on treatment group, of the surrogate and final endpoints) and hence it is difficult to interpret. Similar disadvantages were found even in the case of a model which accounted for multiple surrogate markers and to allow cause-effect relationships between surrogate markers [9, 50]. We need measures for validation of surrogate endpoints which are based on multiple trials involving surrogate and true endpoints because these measures (meta-analytic measures) better capture the uncertainty in relating surrogate and true endpoints than measures derived from a single trial.

6.2.3 Relative Effect and Adjusted Association

The other two alternative measures of surrogacy are relative effect (RE) and the adjusted association (AA)[35]. They base their argument on the premise that surrogates are meaningful quantities to consider if the effect of treatment on the surrogate endpoint can be used to predict treatment effects on the true endpoint. The RE is defined as

$$RE = \frac{\beta_Z}{\gamma_Z} \tag{6.5}$$

where γ_Z and β_Z are estimated from (6.2) and (6.3) respectively. Therefore a new treatment could then be tested through its effect on the surrogate endpoint and declared efficacious if its predicted effect on the true endpoint were sufficiently large to be of clinical interest [18]. The RE value of one corresponds to the surrogate being useful, whereas that of zero corresponds to

the surrogate being useless [23]. The adjusted association (AA) is the correlation between the true and surrogate endpoint, adjusting for the treatment effect. AA measures the correlation between the two endpoints at individual level. We would expect a strong association to reflect some biological pathway from the surrogate endpoint to the clinical endpoint.

Disadvantages of RE

1. Its confidence limits may be too wide to permit clinically useful predictions.
2. Its value may depend on γ_Z , (in other words, since RE is the slope of a regression line between γ_Z and β_Z, the linearity of this regression may be questioned).

RE is best estimated from meta-analysis of several trials. Single trial assessments of RE can be misleading, but from meta-analysis we may check the underlying assumptions that treatment effect magnitudes on the clinical endpoint are proportional to the effects on the surrogate [62]. We can then plot $\gamma_Z$and β_Z, for each trial and the estimated slope is the RE.

In meta-analysis, the AA is calculated by fitting the same models as in the single trial case but stratifying by trial. Thus the applicability and use of the above measures is also strongly dependent on the data available, the more general regression type models are worth considering.

Comparison of PTE and RE, AA

PTE can be easily defined regardless of the nature of the endpoints considered [32], but the interpretations of RE and AA are model dependent.

This difficulty, however, reveals the complexity of the problem, since the notion of a perfect surrogate is hard to define when the surrogate and the true endpoints are not of the same nature [35]. If, for example, the true endpoint was continuos and the surrogate endpoint was binary, the surrogate could never be a perfect surrogate for the true endpoint, except in degenerate cases, because not all the variability in the true endpoint would be accounted for. The opposite case may be more informative (i.e. true endpoint is binary and the surrogate is continuous). The PTE is equivalent to AA/RE, in case of normally distributed errors. The PTE is therefore a

composite measure. Further information on surrogate validity may be gleaned by considering the AA and RE separately [62].

6.2.4 Coefficient of Determination (R^2)

Consider a trial with multiple observations per subject or patient. R^2_{trial} is the coefficient of determination, it measures the strength of association. If R^2_{trial} is close to one the surrogate is trial-level valid [7].

R^2_{indiv} is the coefficient of determination, it is the subject level association between endpoints. If R^2_{indiv} is sufficiently close to one, then a surrogate is individual-level valid.

A surrogate must be trial- and individual-level valid, for overall validity. This approach also yields a prediction of the effect of treatment on the true endpoint. This method is therefore more informative than a simple classification of a surrogate as valid or invalid.

6.2.5 Likelihood Reduction Factor (LRF)

Alonso et al. [1] proposed using the LRF to quantify the treatment effect via surrogate markers. They noted the lack of a unified approach to applying R^2_{trial} and R^2_{indiv} when neither the surrogate nor the clinical endpoint is normally distributed [62]. LRF is related to the generalized correlation between two variables proposed by Kent [29] and the likelihood ratio test (LRT) statistic. LRF, measures the general correlation between two variables [29], such as to quantify the association between surrogate markers and clinical endpoint. It is a measure of individual-level association which may be applied under any generalized linear model to a single trial or meta-analysis. The LRF equals R^2_{indiv} in the special case of normally distributed surrogate and clinical endpoint [62]. Alonso used models 6.2 and 6.3 to define the LRF as

$$LRF(Z, S : Z) = 1 - exp\{-LRT(Z, S : Z)/n\}, \tag{6.6}$$

where $LRF(Z, S : Z)$ is the LRT statistic based on 6.2 and 6.3, and n is the total number of observations [51].

The advantage of this approach is that $LRF(Z, S : Z)$ is bounded by [0,1]. Alonso argued that as the correlation between S and the clinical outcome Y becomes stronger, the estimate of LRF will get larger. But, the quantity LRF(Z,S:Z) reflects the correlation between S and Y after adjusting for the treatment Z, but not the unconditioned correlation between S and Y. In cases where there is a strong correlation between treatment Z and S, the correlation between S and Y may not be strong after adjusting for Z. To observe this, we assume the true models are

$$g(E[Y \mid Z, S]) = g(E[Y \mid S]) = \beta_0 + \beta_S S, \tag{6.7}$$

and

$$S = a_0 + a_Z Z + \varepsilon \tag{6.8}$$

where $\varepsilon \sim NI(0, \sigma^2)$. Models defined by 6.7 and 6.8 indicate that the effect of treatment acts solely through the surrogate marker S. Prentice definition for a valid surrogate marker is satisfied by the surrogate marker S in models 6.7 and 6.8. As σ^2 goes to 0, LRF(Z,S:Z) based on models 6.2 and 6.3 will approach 0 because the collinearity between Z and S is getting stronger, but the relationship between S and Y is unchanged. This shows that the LRF statistic does not reflect the correlation between the surrogate marker and the clinical outcome.

For some generalized linear models, the upper bound of LRF is less than 1 [29]. The adjusted LRF defined by Alonso et al [1] is

$$LRF_a(S, Z : Z) = \frac{LRF(S, Z : Z)}{LRF_{max}}, \tag{6.9}$$

where LRF_{max} is the LRF value for the best-possible fitted model. $LRF_{max} = 1$, for linear regression. For logistic regression, the quantity LRF_{max} can be estimated by the LRF value based on the full model and the simplest model that only includes the intercept as the independent variable. If we assume model 6.3 is the full model, the quantity LRF_{max} can be estimated by

$$\widehat{LRF_{max}} = 1 - exp(-LRT(S, Z : 1)/n), \tag{6.10}$$

where $LRT(S, Z : 1)$ is the LRT statistic based on model 6.3 and the regression model only including the intercept as an independent variable [51]. $LRF_a(S, Z : Z)$ is bounded by $[0, 1]$ with the possibility of reaching 0 and 1.

6.2.6 Adjusted LRF and Proportion of the Information Gain (PIG)

According to Prentice's definition, a key condition for S to be a perfect surrogate marker is $f(Y \mid S, Z) = f(Y \mid S)$. We may compare models 6.3 and 6.7 by a LRT, to check this condition. A generalized correlation between Y and Z after adjusting for S, according to Kent, can be defined as

$$LRF(S, Z : S) = 1 - exp(-LRT(S, Z : S)/n). \tag{6.11}$$

The correlation in 6.11 is related to $LRT(S, Z : S)$, which indicates how much information is gained from model 6.7 by adding an additional variable Z. If $LRT(S, Z : S) = 0$, there is no additional information gained after adding Z, therefore $f(Y \mid S, Z) = f(Y \mid S)$ and S is a perfect surrogate marker [51]. An adjusted LRF is defined by

$$LRF_a(S, Z : Z) = \frac{LRF(S, Z : Z)}{LRF_{max}}, \tag{6.12}$$

which is similar to 6.9 and is bounded by [0,1]. LRF_{max} is estimated by 6.10.

Kullback-Leibler information gain is another measure to quantify the relationship between the surrogate marker and the clinical outcome. It is obtained by comparing models6.3 and 6.7.

The estimated Kullback-Leibler information gain when comparing model 6.7 and the model including only the intercept is $LRT(S : 1)/(2n)$, and the estimated estimated Kullback-Leibler information gain comparing model 6.3 and the model including only the intercept is $LRT(S, Z : 1)/(2n)$. An estimator for the proportion of the information gain is

$$PIG = \frac{LRT(S : 1)}{LRT(S, Z : 1)}. \tag{6.13}$$

LRF and PIG are closely related. Applying the approximation of $exp(t) \approx 1 + t$ when t is small, it is easy to show that

$$1 - LRF_a(S, Z : S) \approx PIG. \tag{6.14}$$

LRF and PIG can be defined similarly for multiple surrogate markers $S_1, S_2, ..., S_k$. For example, $LRF_a(Z, S_1, S_2, ..., S_k : S_1, S_2, ..., S_k)$ can be used to quantify the goodness of k surrogate markers simultaneously.

The advantage of PIG is that it is robust to collinearity between treatment and surrogate markers while PE is not.

Limitations of PIG

Some of the limitations of other quantities used to evaluate surrogate markers (Bakers,[3]) are also applicable to PIG.

1. PIG only applies to individual studies and does not recognize variation in the relationship between the surrogate marker and the clinical outcome across studies,
2. PIG has no established cut-off point, above which the surrogate marker is said to be valid.

It is very unfortunate that there may not be a single answer to this question as each candidate surrogate marker has a unique biological properties that will affect its relationship to the clinical outcome of interest and consequently affect the magnitude of any estimator.

Therefore we need more research in statistical science and a better understanding of the mechanisms by which any given surrogate marker is affecting a clinical outcome so as to overcome these two drawbacks.

6.3 Generalized Estimating Equations

In a longitudinal study design, data are characterized by repeated observations over time on the same unit(s). Observations from the same unit are more likely to exhibit more correlation than those from two different units. If we analyse such data and ignore such within unit correlation, we would get invalid results. The analysis of correlated data arising from repeated measurements when measurements are assumed to be multivariate normal has been studied extensively. But, the normality assumption is not always reasonable, for example, a different methodology must be used in the data analysis when the responses are discrete and correlated. The usual approach used to analyse correlated outcome data is the Generalized Estimating Equations (GEE's) and it provides a practical method to analyze such data with reasonable statistical efficiency. The GEE approach was introduced by Zeger and Liang (1986) [31, 65].

The GEE approach is a method which extends generalized linear models to handle longitudinal or clustered correlated data. The GEE's fall under marginal models, sometimes referred to as

population-averaged models. The target of inference is the population, as opposed to cluster or subject specific models. For example to answer the question whether a treatment is effective or not in the general population, a marginal model is the most appropriate. The term marginal , indicates that the model for the mean response does not depend on any random effects or previous responses but depends only on covariates of interest [21]. The GEE's have frequently been applied in biomedical and health sciences since their invention [31, 15]. No distributional assumptions are required for GEE models, only a regression model for the mean response and a working correlation structure are required [15, 27, 31, 65]. Here, the "working correlation" structure, accounts for the correlation within each unit and it is to be specified for each analysis.

The GEE approach is generally applicable to both continuous and discrete responses. It provides a non-likelihood based or a quasi-likelihood approach for modelling responses which are correlated. The models are an extension of generalized linear models (GLMs)[40].

It is important to consider the type of outcome or response variable (i.e continuous or discrete) in modeling correlated data statistically.

The advantage of the GEE method is that it can be used with both discrete and continuous explanatory variables, a large number of categorical variables, missing response values, and/or time-dependent covariates, but the weighted least squares does not apply. In this context it is more advanced than the general linear model.

6.3.1 Binary Longitudinal Data Model with GEE

A binary response is one that has two possible outcomes e.g. 1 or 0, depending on whether an event of interest occurred or not. A GEE model for a binary response observed longitudinally is an extension of the standard logistic regression model from the generalized linear model approach for independent observations [39].

Let Y_{ij} be the response from participant i at time t_{ij}, for $i = 1, 2, ..., N$ and $j = 1, ..., n_i$. Also, define $Pr(Y_{ij} = 1) = \mu_{ij}$ as the marginal mean of Y_{ij}. This is the probability of observing the event of interest $(Y_{ij} = 1)$ for participants i at time t_{ij}.

There are three steps to follow when using a GEE approach. The first step is to set up a regression model (i.e a model to relate the dependent or response variable to a linear combination of explanatory variables or independent variables). A regression model type of analysis is used to check if any of the explanatory variables are associated with the response variable.

Since the mean of a binary outcome is a probability (i.e must be between 0 and 1), a transformation on the mean through a link function $g(\mu_{ij})$ is needed to ensure all possible values of the linear predictor model map onto the real line [27].

The logit (the log of an odds) scale is commonly used link function for binomial data. In this case the model is given by

$$log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_0 + \beta_1 X_{ij1}, ..., +\beta_k X_{ijk};$$

where $X_{ij1}, ..., X_{ijk}$ are the explanatory variables for participant i at time t_{ij}, and $\beta_0, \beta_1, ..., \beta_k$ are the intercept and the regression coefficient parameters of the explanatory variables that should be estimated, respectively. Note that in the above model specification, covariates appear as time dependent but in reality some covariates are only measured at baseline. It should be noted that the coefficient β_p, $p = 1, 2, ..., k$ are interpreted on logit scale on the response.

The second step, is to define the variance of Y_{ij} in the GEE approach. The variance is completely determined by the mean of Y_{ij} as

$$Var(Y_{ij}) = \mu_{ij}(1-\mu_{ij}),$$

for a binary response variable. In the final step, you have to choose the working correlation structure between observations on the same participant. The correlation structure $\gamma_{jp} = Corr(Y_{ij}, Y_{ip})$, $j \neq p$ may depend on a vector of unknown parameters α, which is assumed to be equal for all participants [27]. However in practice, the true underlying correlation structure of repeated measurements is usually unknown, so it is difficult to decide which constrained structure is to be used in advance [26]. This may often lead to the correlation matrix structure being misspecified and thereby distorting the results [14]. To solve this problem, the GEE method can be used for valid parameter estimation.

Even when the working correlation is misspecified, the GEE approach generally produces consistent estimators of the true variance of the estimated parameters [27]. This is called the robust or

empirical covariance estimator. For this reason the GEE method can still be used for valid parameter estimation.

Working Correlation Matrix

A model that relates a marginal mean to the linear predictor $x^{'}\beta$ through a link function need to be chosen. The generalized estimating equations for estimating β, as an extension of the generalized linear model (GLM) equation, is given by

$$\sum_{i=1}^{n} \frac{\partial \mu^{'}}{\partial \beta} \mathbf{V}_i^{-1}(\mathbf{Y}_i - \mu_i(\beta)) = 0$$

where μ_i=$(\mu_{n_i}), ..., \mu_{it_i})^{'}$, $\mathbf{Y}_i = (y_{i1}, y_{i2}, y_{i3}, ..., y_{n_i})$ and $\mathbf{V}_i$ is the assumed or model based covariance matrix of $\mathbf{Y}_i$. GLM estimating equations are similar to these equations except that, here with GEEs we have multiple outcomes. They include a vector of means instead of a single mean and a covariance matrix instead of a scalar variance [37]. The covariance matrix of $\mathbf{Y}_i$ is modelled as follows:

$$\mathbf{V}_i = \phi \mathbf{A}_i^{\frac{1}{2}} \mathbf{R}_i(\alpha) \mathbf{A}_i^{\frac{1}{2}}$$

where $\mathbf{A}_i$ is a $n_i \times n_i$ diagonal matrix with $v(\mu_{ij})$ as the jth diagonal element. The quantity ϕ is a dispersion parameter and $\mathbf{R}_i$ is the working correlation matrix. The $(j, j^{'})$ element of $\mathbf{R}_i(\alpha)$ is the assumed or proposed working correlation between y_{ij} and $y_{ij'}$ which depends on α.

The widely used or assumed working correlation structures are exchangeable, auto-regressive(AR1), independent and unstructured. A few of these are briefly discussed below to further add clarity.

- Exchangeable structure:- This structure allows for the assumption that the correlations between any pair of observation from the same individual be the same, irrespective of the length of the time interval. Thus for four repeated measurements $Y_{ij}, j = 1, ..., n_i$ we have:

$$Corr(Y_{ij}, Y_{i,j}) = \begin{cases} 1 & j = j \\ \rho & j \neq j \end{cases}$$

$$R = \begin{pmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{pmatrix}$$

This structure requires only one estimated parameter. Although the specification of constant correlation between any two repeated measurements may not be justified in a longitudinal study (closeby observations are more correlated than far apart observations), it is often reasonable in situations in which the repeated measures are not obtained over time [15, 31, 65]. It is probably reasonable when there are a few repeated measurements not taken over a wide time space [37]. This is a commonly used structure and it is relatively easy to explain to investigators.

- Autoregressive structure: -The correlation between two observations or measurements which are m time units apart is ρ^m, where $0 < \rho < 1$. The greater the power m, the smaller the correlation. Therefore the correlation diminishes for further apart observations. With four repeated measurements equally spaced we have:

$$Corr(Y_{ij}, Y_{i,j+s}) = \rho^s$$

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

Similarly this structure requires one estimated parameter.

- Independence structure:- It adopts the working assumption that repeated observations for a subject are independent. In this case solving the GEE is the same as fitting the usual regression models for independent data and the resulting parameter estimates are the same, but the standard errors are different [15]. The GEE method still accounts for correlation by operating at the cluster level.

$$R = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

- Unstructured structure:- It assumes that all correlations are different. It is the least restrictive correlation structure but it requires $n_i(n_i - 1)/2$ parameters to be estimated where n_i is the number of observations for subject i. Here

$$Corr(Y_{ij}, Y_{i,k}) = \begin{cases} 1 & j = j \\ \rho_{jk} & j \neq j \end{cases}$$

$$R = \begin{pmatrix} 1 & \rho_{21} & \rho_{31} & \rho_{41} \\ \rho_{21} & 1 & \rho_{32} & \rho_{42} \\ \rho_{31} & \rho_{32} & 1 & \rho_{43} \\ \rho_{41} & \rho_{42} & \rho_{43} & 1 \end{pmatrix}$$

It is therefore very important to have a good idea of the most appropriate correlation structure for an analysis. An analysis that uses an unstructured correlation matrix will be less efficient than an analysis that uses the proper structure. The problem, though, is knowing which structure to use, but minimizing the degree of misspecification helps to model the problem much easier.

6.3.2 Application to the CAPRISA AIDS Cohort Study

Between 28 June 2004 and 22 January 2009, CAPRISA (Centre for the AIDS Programme of Research in South Africa) enrolled 1011 HIV positive patients from Vulindlela area in Pietermaritzburg. The data was collected as part of CAPRISA's AIDS treatment project. For inclusion HIV positive men and women ≥ 14 years with CD4+ count < 200 cells/μL or who were at WHO stage IV disease were enrolled. There were 351 men and 660 women involved in the study, thus females were overpresented. Weight, height, viral load, $CD4$ count, age and gender were collected at baseline and the patients were put on ARV treatment. Thereafter six monthly visits

were scheduled for all participants, where their $CD4$ count and viral load were measured. The data used here is for baseline, 6 months, 12 months and 18 months. For analysis viral loads and $CD4$ counts were treated first as continuous variables. The age variable (age at time of entry into the study) is a continuous variable. Sex was by default a binary predictor.

Explanatory variables in the model are gender (x_{ij2}), age (x_{ij3}), $CD4$ count (x_{ij4}), viral load (x_{ij5}), height (x_{ij6}) and weight (x_{ij7}). The model also included the intercept (x_{ij1}). In a further analysis $CD4$ and viral load count variables were redefined as categorical variables as shown below to provide an alternative analysis compared to when the two biomarkers the are treated as continuous:

$$X_{ij4} = \begin{cases} 1 & CD4count < 200 \\ 2 & 200 < CD4count < 400 \\ 3 & CD4count \geq 400 \end{cases}$$

and

$$X_{ij5} = \begin{cases} 1 & viralload < 400 \\ 2 & viralload \geq 400 \end{cases}$$

The outcome of interest used in the current analysis is the patient's survival status at visit j. Suppose y_{ij} represents the survival status (response variable) of patient i at the jth visit, for $i = 1, ..., 1011$ and $j = 0 \quad (baseline), 6, 12, 18$ months. The response is recorded as 0 for the patient who is alive at each visiting time and 1 if the person died. Now, $\mu_{ij} = E(Y_{ij})$ represents the mean of the survival status, which is the probability that the patient died between the $j-1$ and j^{th} visits. Thus the exact time of death is not known therefore a time to event model is not possible. The response variable is binary, therefore we use the logit link function for the binomial distribution $g(\mu_{ij})$=$log\{\mu_{ij}/(1-\mu_{ij})\}$. The general mean model is

$$g(\mu_{ij}) = x_{ij}^{'}\beta$$

where β is a vector of regression parameters to be estimated.

In total 109 patients died, representing $10,8\%$ mortality overall. The study is still ongoing since it was designed as an open observational study to administer patient care. There was missing data due to known and unknown various reasons. Some patients asked to be transferred to the nearest

clinics so that they could collect their ARV's early, thereby causing missing for those patients in the original study. Others died before they could go for the next appointment and some patients decided to stop treatment on their own, an action that could be due to possible side effects or some other non-adherence reasons. Other patients would miss intermittently for example get treatment after baseline and not return for medical examination after 6 months, but come back after 12 months and this would happen for different patients. Thus the analysis is faced with complex missing data problems requiring methods that are capable of using the available data correctly. The core problem in the current chapter is to study the relationship between survival and key disease biomarkers namely $CD4$ count and viral loads.

The GENMOD procedure in SAS was used to analyse the data, using the GEE method. The assurance that the GEE method can handle missing data under the missing completely at random (MCAR) assumption is an advantage. We first used the exchangeable working correlation structure in the analysis. Measures to assess the model goodness of fit similar to maximum likelihood estimation (MLE) are not available in the current GENMOD procedure because the methodology is based on quasi-likelihood estimation. The MCAR assumption in missing data methodology means the probability of a missing observation depends on observed covariates and not on the observed or unobserved outcomes [55].

The GENMOD procedure in SAS was used to analyse the data, using the GEE method. We first used the exchangeable working correlation structure in the analysis. Measures to assess the model goodness of fit similar to maximum likelihood estimation (MLE) are not available in the current GENMOD procedure because the methodology is based on quasi-likelihood estimation. Missing data is a common problem in the analysis of longitudinal data. Data can be said to be missing at random (MAR) if the mechanism causing the missing data depends only on the observed outcomes. Data is said to be missing not at random (MNAR) if the mechanism causing the missing data depends on both observed and unobserved outcomes. Finally data is said to be missing completely at random (MCAR) if the mechanism causing the missing data does not depend on either observed or unobserved outcomes [58, 31]. Inorder to explore the missingness assumption in the current analysis, first a dropout indicator R_{ij} was created which took the value 1 if the outcome Y_{ij} was missing and 0 otherwise. A model relating the response and the dropout

indicator including measured and observed covariates was fitted and the dependence between the two was found not to be significant. Thus in the current analysis the MCAR assumption was used. This was an assurance because the GEE method works under the MCAR assumption but its strength lies in its capability to account for correlation in the data. In addition the fact that it allows for empirical based standard errors is an added advantage.

Descriptive Statistics

Table 6.1 below gives descriptive statistics for the CAPRISA CAT data.

Table 6.1: Descriptive Statistics for CAPRISA data used in the GEE analysis

Variable	N	Mean	Std Dev	Min	Max
AGE	3977	33.836	8.817	14.00	75.00
WEIGHT	3779	60.093	12.462	24.00	125.00
HEIGHT	3714	159.272	11.206	117.00	197.00
CD4count	3882	229.470	153.254	6.00	1820.00
Viral Load	3582	119627.920	504403.590	40.00	9348302.00

Table 6.2 shows that the mean $CD4$ count for males $(203, 68)$ was less than the mean for females $(239, 76)$. The opposite is true for viral loads, the viral load mean for males $(141143, 06)$ s greater than that for females $(111033, 30)$. Also the mean height for males $(166, 54)$ is greater than that for females $(156, 46)$. The means for weight and age are almost equal for both sexes.

Table 6.2: Descriptive Statistics for the Repeated measurements in CAPRISA data categorized by Gender

Gender	Total Observations	Variable	N	Mean	Std Dev
Male	1131	$CD4$count	1100	203.68	140.71
		Viral load	1017	141143.06	601119.45
		Height	1036	166.54	11.28
		Weight	1066	59.64	9.64
		Age	1126	35.36	8.54
Female	2853	$CD4$count	2774	239.76	156.77
		Viral load	2559	111033.30	460668.43
		Height	2678	156.46	9.83
		Weight	2708	60.22	13.33
		Age	2846	33.29	8.75

6.3.3 GEE analysis with viral load and $CD4$ count as continuous variables

The GENMOD procedure was used first treating all of the measurements as independent and fits a generalized linear model, so as to generate a starting solution. These parameter estimates are then used as starting values for the GEE solution [37], the exchangeable correlation structure was used.

Score Statistics

Table 6.3 contains the Type 3 results for the model effects. The results show that height, weight and gender are not significant factors in relation to the probability of death. But, age has a nearly significant association ($p = 0.0618$). We note that $CD4$ count ($p = 0.0004$) and viral load ($p = 0.0001$) have got a highly significant association with survival status.

Table 6.3: Score Statistics for Type 3 GEE Analysis

Source	DF	Chi-Square	Pr>ChiSq
Gender	1	0.15	0.6983
AGE	1	3.49	0.0618
WEIGHT	1	2.45	0.1173
HEIGHT	1	0.03	0.8732
CD4count	1	12.67	0.0004
Viral Load	1	14.53	0.0001

GEE Parameter Estimates

Table 6.4: Analysis of GEE Parameter Estimates: Empirical Standard Error Estimates (with continuous $CD4$ count and viral load)

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr>| Z |
Intercept	-2.9182	1.8928	-6.6280	0.7916	-1.54	0.1231
Gender Male	0.0907	0.2295	-0.3592	0.5406	0.40	0.6928
Age	0.0226	0.0110	0.0011	0.0441	2.06	0.0395
CD4 Count	-0.7963	0.2476	-1.2816	-0.3111	-3.22	0.0013
Viral Load	0.8715	0.2259	0.4287	1.3144	3.86	0.0001
Height	-0.0018	0.0107	-0.0228	0.0192	-0.16	0.8695
Weight	-0.0171	0.0119	-0.0403	0.0061	-1.44	0.1492

Since the effects reported in the Type 3 analysis are single degree of freedom effects, the score statistics in that table are assessing the same hypotheses as the Z statistic in GEE parameter estimates table. In Table 6.4, the p-value for Z for age is 0.0395, compared to the value 0.0618 using the score statistic in "Type 3" statistic. Also the p-value for Z for $CD4$ count is 0.0013, compared to the value 0.0004 from the score statistic in "Type 3" statistic. We would assess the null hypothesis with score statistic, in a strict testing situation [37]. The Z and Wald statistic

generally produce more liberal p-values than score statistics.

The parameter estimate for $CD4$ count is -0.7963. The negative sign indicates that if $CD4$ count increases, the odds and hence probability of death decreases, holding other covariates fixed. This makes sense because with an increase in $CD4$ count, the body is able to fight the diseases more efficiently, thereby reducing the chance of death due to disease. The increase in $CD4$ count can be attributed to the beneficial effect of the ARVs which helps the body immune system to reconstitute itself.

The parameter estimate for viral load is 0.8715. This shows that viral load is positively associated with the probability of death therefore an increase in viral load increases the chance or probability of death. This makes sense because when the viral load increases, the immune system is compromised such that the body can no longer fight diseases, thereby increasing the chance of death of an individual. The interaction between $CD4$ count and viral load was found not to be a significant factor ($p = 0.2399$) for the probability of death.

To interpret the age effect at baseline, let x^* denote age at baseline and let P_{x^*} denote the probability of death holding other covariates fixed. Then

$$logit(P_{x^*}) = \beta_0 + constant + \beta_* x^*$$

older HIV patients at baseline are at high risk of death than younger ones. The odds of death for a patient one year older is $e^{0.0226} = 1.023$ times the odds of death for a patient aged x^*. The same interpretation holds for any other continuous variable measured only at baseline. Thus older patients tend to have a higher probability of death than younger ones in this cohort of HIV infected individuals.

Covariance Matrix

Tables 6.5 and 6.6 are the model-based and empirical covariance matrices of the model parameter estimates respectively. Between the two we give importance to the empirical estimate of the covariance matrix determined empirically because it is data driven. At the same time, we note that even if the correlation between observation structure is misspecified, both the parameter

Table 6.5: Covariance matrix (Model-Based)

	Constant	Gender	Age	CD4 count	Viral load	Height	Weight
Constant	2.73981	0.11438	−0.004269	−0.10589	−0.12036	−0.01411	−0.000846
Gender	0.11438	0.05442	−0.000205	0.004246	0.002422	−0.000904	0.0001715
Age	−0.004269	−0.000205	0.0001143	0.0000908	0.0000968	$3.8379E-6$	−0.000010
CD4 count	−0.10589	0.004246	0.0000908	0.04783	0.02081	0.0001036	−0.000148
Viral load	−0.12036	0.002422	0.0000968	0.02081	0.04889	0.0001022	−0.000059
Height	−0.01411	−0.000904	$3.8379E-6$	0.0001036	0.0001022	0.0000956	−0.000021
Weight	−0.000846	0.0001715	−0.000010	−0.000148	−0.000059	−0.000021	0.0000834

Table 6.6: Covariance matrix (Empirical)

	Constant	Gender	Age	CD4 count	Viral load	Height	Weight
Constant	3.58271	0.11700	−0.005099	−0.17496	−0.14575	−0.01804	−0.001959
Gender	0.11700	0.05269	$7.8808E-6$	0.002857	0.003184	−0.000878	−0.000058
Age	−0.005099	$7.8808E-6$	0.0001202	0.0002385	0.0001448	$1.1811E-6$	$1.8555E-6$
CD4 count	−0.17496	0.002857	0.0002385	0.06129	0.02617	0.0005298	−0.000654
Viral load	−0.14575	0.003184	0.0001448	0.02617	0.05105	0.0005058	−0.000945
Height	−0.01804	−0.000878	$1.1811E-6$	0.0005298	0.0005058	0.0001147	−0.000024
Weight	−0.001959	−0.000058	$1.8555E-6$	−0.000654	−0.000945	−0.000024	0.0001406

Algorithm converged

estimates and their empirical standard errors are consistent, provided that the specification of the mean model is the correct model.

The working correlation structure was estimated as shown below. Thus from Table 6.7, we note

Table 6.7: Working Correlation Matrix

	Col1	Col2	Col3	Col4
Row1	1	0.0026	0.0026	0.0026
Row2	0.0026	1	0.0026	0.0026
Row3	0.0026	0.0026	1	0.0026
Row4	0.0026	0.0026	0.0026	1

that the correlation between any two response Y_{ij}, $Y_{i\mu}$ was estimated as is 0.0026. However as earlier discussed this is not necessarily an estimate of the true underlying correlation but the assumed correlation.

6.3.4 GEE analysis with viral load and $CD4$ count as continuous variables fitted independently

When the viral load variable is removed from the model and we fit $CD4$ count and all the covariates, we found out that $CD4$ count $(p < 0.0001)$ is a highly significant factor in relation to the probability of death, and weight $(p = 0.0055)$ is now significantly associated with the probability of death. Gender, height and age are not significant factors in relation to the probability of death. Also, when $CD4$ count variable is removed from the model and we fit viral load alone and all other covariates, we found out that viral load $(p < 0.0001)$ is a highly significant factor for the probability of death, while weight $(p = 0.0548)$ is a less significant factor. Gender $(p = 0.0227)$ is also found to be a significant factor for the probability of death. Height and age were found to be not significant factors for the probability of death.

6.3.5 GEE analysis with viral load and $CD4$ count as categorical variables

Score Statistics

Table 6.8 contains the Type 3 results for the model effects. The results show that height, weight

Table 6.8: Score Statistics for Type 3 GEE Analysis

Source	DF	Chi-Square	Pr>ChiSq
Gender	1	0.13	0.7174
AGE	1	3.60	0.0579
WEIGHT	1	2.37	0.1233
HEIGHT	1	0.03	0.8668
CD4count	2	17.74	0.0001
Viral Load	1	12.72	0.0004

and gender are not significant factors in relation to the probability of death. But, age has a nearly significant association ($p = 0.0579$), $CD4$count ($p = 0.0001$) and viral load ($p = 0.0004$) have got a highly significant association with survival status.

GEE Parameter Estimates

The logit link model parameters for categorical covariates are based on reference level interpretation. We can exponentiate the parameter estimates to obtain estimates of odds ratios for a category with respect to the reference level. The parameter estimate for $CD4$ count in the category $CD4 < 200$ is 0.9375, this means that the odds of death for those in the category $CD4 < 200$ of $CD4$ count are $e^{0.9375} = 2.554$, times the odds of death for those patients with $CD4$ count≥ 400. Those in category 1 are 2.554 times more likely to die than those in category 3. The parameter estimate for viral load is -0.8139, the odds of death for those in a lower category of viral load is $e^{-0.8139} = 0.6420$, times the odds of death for those patients with higher viral

Table 6.9: Analysis of GEE Parameter Estimates: Empirical Standard Error Estimates with categorical $CD4$ count and viral load)

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr>| Z |
Intercept	-2.9138	1.7088	-6.2630	0.4354	-1.71	0.0882
Gender Male	0.0842	0.2285	-0.3637	0.5320	0.37	0.7127
Age	0.0230	0.0110	0.0015	0.0445	2.10	0.0360
CD4 Count 1	0.9375	0.4504	0.0546	1.8203	2.08	0.0374
CD4 Count 2	-0.2563	0.4809	-1.1989	0.6863	-0.53	0.5941
Viral Load 1	-0.8139	0.2263	-1.2574	-0.3704	-3.60	0.0003
Height	-0.0018	0.0105	-0.0224	0.0187	-0.17	0.8627
Weight	-0.0167	0.0118	-0.0398	0.0064	-1.42	0.1560

load. This means that an increase in viral load increases the chance or probability of death and an increase in $CD4$ count reduces the probability of death. It should be noted that an analysis with categorical $CD4$ count and viral load is more informative than its counterpart where the variables are treated as continuous each contributing 1 degree of freedom. The model better relates the two biomarkers as surrogates for survival. The interaction effects of $CD4$ count and viral load was assessed but found to be insignificant. However both the continuous and categorical analysis show a strong additive independent effects due to the two markers on survival.

6.3.6 GEE analysis with viral load and $CD4$ count as categorical variables fitted independently

When the viral load variable is removed from the model and we fit $CD4$ count and all the covariates, we found out that $CD4$ count $(p < 0.0001)$ is a highly significant factor in relation to the probability of death, and weight $(p = 0.0011)$ is also significant factor. Gender and age are not significant factors in relation to the probability of death. Also, when $CD4$ count variable is removed from the model and we fit viral load and all the covariates, we found out that viral load

$(p < 0.0001)$ is a highly significant factor for the probability of death, while weight $(p = 0.0542)$ is a less significant factor. Height and gender were found to be not significant factors for the probability of death.

6.3.7 GEE parameter estimates using various working correlation structures

For comparison and completeness we fitted the model 6.15 using several working correlations: exchangeable, AR1, unstructured and independent correlation structures. $CD4$ count and viral load are treated as continuous variables. The common mean model fitted is

$$logit(\mu_{ij}) = \beta_0 + \beta_1 Gender + \beta_2 Age + \beta_3 CD4count + \beta_5 Viralload + \beta_6 Height + \beta_7 Weight \tag{6.15}$$

Table 6.10 shows the results of the analysis for model (6.15). We obtained consistent parameter estimates from using different correlation structures. These results are consistent with [27, 31, 37], which stated that GEE methods are robust to an assigned correlation structure; you can misspecify that correlation structure and still obtain consistent parameter estimates and standard errors. This is evident in 6.10 where parameter estimates and standard errors (empirical) for all predictors are quite similar irrespective of the correlation structure. The p-values are also fairly consistent which is an assurance for consistent interpretation.

6.3.8 Conclusion and Discussion

We used the GEE method because it is a standard method for analyzing non-normal longitudinal data and also [27] recommended using the GEE method only if the number of participants is at least 30, and if 3 to 5 data points per participant are assessed. This was the case with CAPRISA data. From the GEE analysis, with both $CD4$ count and viral load treated as continuous and categorical, we discovered that $CD4$ count and viral load can potentially serve as a surrogate endpoint for death. Weight was also found to be a significant factor for the probability of death, when viral load and $CD4$ count were fitted independently. This means that weight at baseline

Table 6.10: Analysis of GEE parameter estimates based on empirical standard error estimates, using various working correlation structures, with death status as outcome variable, and gender, age, $CD4$ count, viral load, weight and height as explanatory variables .

Correlation	Parameter	Estimate	Standard Error	p-value
Exchangeable	Intercept	-2.9182	1.8928	0.1231
	Gender Male	0.0907	0.2295	0.6928
	Age	0.0226	0.0110	0.0395
	$CD4$ count	-0.7963	0.2476	0.0013
	Viral load	0.8715	0.2259	0.0001
	Height	-0.0018	0.0107	0.8695
	Weight	-0.0171	0.0119	0.1492
AR1	Intercept	-2.9248	1.8946	0.1226
	Gender Male	0.0901	0.2296	0.6946
	Age	0.0226	0.0110	0.0393
	$CD4$ count	-0.7983	0.2484	0.0013
	Viral load	0.8734	0.2264	0.0001
	Height	-0.0017	0.0107	0.8704
	Weight	-0.0171	0.0118	0.1494
Unstructured	Intercept	-2.9013	1.8900	0.1248
	Gender Male	0.0927	0.2295	0.6863
	Age	0.0228	0.0110	0.0380
	$CD4$ count	-0.8006	0.2479	0.0012
	Viral load	0.8708	0.2251	0.0001
	Height	-0.0020	0.0107	0.8549
	Weight	-0.0169	0.0118	0.1538
Independent	Intercept	-2.9289	1.8961	0.1224
	Gender Male	0.0900	0.2296	0.6950
	Age	0.0226	0.0110	0.0394
	$CD4$ count	-0.7986	0.2488	0.0013
	Viral load	0.8762	0.2267	0.0001
	Height	-0.0018	0.0107	0.8691
	Weight	-0.0170	0.0118	0.1500

alone cannot be used as a surrogate for survival. However given the apparent association between $CD4$ count and viral load, weight at baseline may predict poor or better performance on ARVs for HIV infected patients.

The exact time of death is not known, this is due to the fact that the data collectors could get the information that a certain person is dead if he does not turn up for the next visit and

relatives are contacted and they confirm that the person is dead. For this reason a time to event analysis is not feasible. Also it is hard to know whether death was caused by an HIV related illness or not because most patients die at home and the data collectors are not able to determine whether death was due to HIV illness or not. In the analysis the assumption is that patients died due to HIV illnesses and if so possibly associated with HIV biomarkers. This was found to be the case in the current analysis but still data quality can be an issue, for example, the issue of incomplete information, here in this study time of death is not known and other causes of death are not known. Though age and weight are time-dependent explanatory variables, they were only measured at baseline in the CAPRISA data. There were more females than males enrolled in the study, this could be due to the fact that many HIV positive women are at child-bearing age. When females get pregnant they are encouraged to get tested for HIV and thereby getting them to know their status. This compels them to start seeking treatment when they are tested positive. As for men, it is up to them to go and get tested and most of them do not prefer to get tested due to fear of stigma. Thus tools such as education campaign or voluntary counselling and testing (VCTs) should continue to be applied to improve this situation.

The assay used to detect viral load could only go as low as $40/mm^3$ meaning that if the viral load dipped below $40/mm^3$ it could not be detected. This leads to the problem of lower detection limit in determining viral loads. A better modelling approach to deal with this uncertainty is a possible future problem.

In the analysis weight was available only at baseline. We believe the association of weight and survival (alive or dead) could have been understood better if the covariate weight was available repeatedly over time (time dependent covariate).

In this work we first analyzed a dynamical HIV model including treatment. We investigated three cases i.e varying the rate of quiescent cells production (λ), efficacy of treatment (η), and proportion of infectious virions ω, on an existing HIV model. We determined the basic reproduction number and the effective reproduction number. From the simulations, we discovered that the higher the treatment efficacy, the lower the number of infected cells are left in the body. It would be a great discovery, if treatment with treatment efficacy of one would be found, because that would mean that all infected $CD4$ cells would be killed in the body, and that HIV would be

curable. At the moment, this "magic bullet" is still a dream.

This study gives a basic insight to understand the impact of HIV treatments on the infected $CD4$ cells. The model had a constant λ (the production rate of quiescent cells), resulting in graphs not showing the AIDS stage. Further work based on this dissertation would have to extend this basic model to incorporate a non-constant λ. Nonetheless this dissertation is a good example to show that mathematical analysis will continue to play a huge and major role in solving problems in biology and epidemiology.

Further statistical analysis on the CAPRISA data was carried out using the GEE modelling method. From the analysis we discovered that $CD4$ count< 200 and high viral load can potentially serve as surrogate endpoints for survival in HIV studies. Categorizing variables seemed to be more efficient than just using them as continuous variables because categorizing gives us cut off points to signify where a variable starts being a significant factor.

Finally the important area of causal inference and surrogate marker validation for studies and data generated in Africa still remains under-developed and efforts to enhance capacity in the area are an important undertaking. We hope the current study has initiated this process.

Bibliography

[1] Alonso, A., Molenberghs, G., Burzykowski, T., Renard, D., Geys, H., Shkedy, Z., Tibaldi, F., Abrahantes, J. C. and Buyse, M. (2004). Prentice's approach and the meta-analytic paradigm: A reflection on the role of statistics in evaluation of surrogate endpoints. *Biometrics* **60**, 724-728.

[2] Anderson, R. W., Ascher, M. S. and Sheppard, H. W. (1998). Direct HIV cytopathicity cannot account for CD4 decline in AIDS in the presence of homeostasis: A worst-case dynamic analysis. *Journal of acquired immune deficiency syndromes and human retrovirology* **17**, 245-252.

[3] Baker, S. G. (2006b). Surrogate endpoints: Wishful thinking or reality. *Journal of the National Cancer Institute* **98**, 502-503.

[4] Baker, S. G. (2005). A simple meta-analytic approach for using a binary surrogate endpoint to predict the effect of intervention on true endpoint. *Biostatistics* **7,1**, 58-70.

[5] Blattner, W. A. (1991). HIV epidemiology: past, present, future. *FASEB Journal* **5**, 2340-2348.

[6] Burton, G. F., Keele, B. F., Estes, J. D., Thacker, T. C. and Gartner, S. (2002). Follicular dendritic cell contributions to HIV pathogenesis. *Seminars in Immunology.*, **14 (4)**, 275284, doi:10.1016/S1044-5323(02)00060-X, PMID 12163303.

[7] Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D. and Geys, H. (2000). Statistical validation of surrogate endpoints: problems and proposals. *Drug Information Journal* **34**, 447-454.

[8] Cardiac Arrhythmia Suppression Trial (CAST) Investigators. (1989). Preliminary report: Effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardinal infraction. *New England Journal of Medicine* **321**, 406-412.

[9] Chen, C., Wang, H. and Snapinn, S. M. (2003). Proportion of treatment effect (PTE) explained by a surrogate marker. *Statistics in Medicine* **22**, 3349-3359.

[10] Chopra, M., Lawn, J. E., Sanders, D., Barron, P., Karim, A. S. S., Bradshaw, D., Jewkes, R., Karim, A. Q., Flisher, A. L., Mayosi, M. B., Tollman, S. M., Churchyard, G. J. and Coovadia, H. (2009). Achieving the health Millenium Development Goals for South Africa:challenges and priorities. *The Lancet* **374**, 1023-1031.

[11] Chun, T. W., Carruth, L. and Finzi, D. (1997). Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. *Nature* **387**, 183-188.

[12] Commenges, D. and Rondeau, V. (2006). Relationship between derivatived of the observed and full loglikelihoods and application to Newton-Raphson algorithm. *International Journal of Biostatistics* **2**, http://www.bepress.com/ijb/vol2/iss1/4/.

[13] Daar, E. S., Little, S., Pitt, J., Eric S., Santangelo, J., Pauline, Ho., Harawa, N., Peter, K., Janis, V. G., Jiexin, M. D; Paula, G., Douglas, D. R., Susan, M. and Stephen, N. (2001). Diagnosis of primary HIV-1 infection, Los Angeles County Primary HIV Infection Recruitment Network. *Annals of Internal Medicine* **134 (1)**, 925. PMID 11187417.

[14] Diggle, P. J.(1998). An approach to the analysis of repeated measurements. *Biometrics* **44**, 959-971.

[15] Diggle, P. J., Heagarty, P., Liang, K-Y., and Zeger, S. (2002). Analysis of longitudinal data. *The Journal of Applied Statistics in the Pharmaceutical industry*. **3,2**, 147-148.

[16] Driessche, P. V. d. and Watmough, J. (2002). Reproduction numbers and sub-threshhold endemic equilibria for compartmental models of disease transmission. *Mathematical Biosciences* **180**, 29-48.

[17] DeGruttola, V., Fleming, TR., Lin, DY. and Coombs, R. (1997). Perspective: validating surrogate markers-are we being naive? *The Journal of Infectious Diseases* **175**, 237-246.

[18] Ellenberg, S. S. (1991). Surrogate endpoints in clinical trials: Getting closer to identifying markers for survival in AIDS. *British Medical Journal* **302**, 63-64.

[19] Freedman, L. S., Graubard, B. I. and Schatzkin, A. (1992). Surrogate endpoints in clinical trials: Definitions and operational criteria. *Statistics in Medicine* **11**, 167-178.

[20] Gallo, R. and Montaigner, L. (2002). Prospects for the future. *Science* **298**, 1730-1731.

[21] Garrett, M. F., Nan, M. L. and James, H. W. (2004). *Applied longitudinal analysis analysis*. Wiley Series in Probability and Statistics **QA278.F575.**, 391-321.

[22] Genz, A. and Keister, B. D. (1996). Fully symmetric interpolatory rules for multiple integrals over infinite regions with gaussian weight. *Journal of Computational and Applied Mathematics* **71**, 299-311.

[23] Gosh, Debashis. (2009). On assessing surrogacy in a single trial setting using a semicompeting risks paradigm. *Biometrics* **65**, 521-529.

[24] Grossman, Z., Meier-Schellersheim, M., Sousa, A. E., Victorino, R. M. M. and Paul, W. E. (2002). CD4+ T-cell depletion in HIV infection: Are we closer to understanding the cause? *Nature Medicine* **8**, 319-323.

[25] Guedj, J., Thiebaut, R. and Commenges, D. (2007). Maximum likelihood estimation in dynamical models of HIV. *Biometrics* **63**, 1198-1206.

[26] Heungsun, Hwang. and Yoshio Takane. (2005). Estimation of growth curve models with structured error covariances by generalized estimating equations. *Behaviormetrika* **32,2**, 155-163.

[27] Ji-Hyun, Lee., Thaddeus, A. Herzog., Cathy, D. Meade., Monica, S. Webb. and Thomas, H. Brandon. (2007). The use of GEE for analyzing longitudinal binomial data: A primer using data from a tobacco intervention. *Science Direct Addictive Behaviors* **32**, 187-193.

[28] Jos, T. (2003). *Applied longitudinal data analysis for epidemiology system: A practical guide*. Cambridge university press, 133-144.

[29] Kent, J. (1983). Information gain and a general measure of correlation. *Biometrika* **70**, 163-173.

[30] Kuznetsov, Y. G., Victoria, J. G. and Robinson, W. E. (2003). Atomic force microscopy investigation of human immunodeficiency virus (HIV) and HIV-infected lymphocytes. *Journal Virology* **77**, 11896-11909.

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

[32] Lin, DY., Fleming, TR. and DeGruttola, V. (1999). Estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in Medicine* **18**, 107-15.

[33] Louis T. (1982). Finding the observed Information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B* **44**, 226-233.

[34] Lurie, M., Williams, B., Zuma, K., Mkaya-Mwaburi, D., Garnett, G.P., Sturm, A.W., Sweat, M.D., Gittelsohn, J., and Abdool Karim, S. (2003). The impact of migration on HIV1 transmission in South Africa: A study of migrant and nonmigrant men and their partners. *Sexually Transmitted Diseases*, **30(2)**, 149-156.

[35] Marc, B and Geert, M. (1998). Criteria for validation of surrogate endpoints in randomized experiments. *Biometrics* **54**, 1014-1029.

[36] Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics* **11**, 431-441.

[37] Maura, E. S., Charles, S. D. and Gary, G. K. (2003). Categorical data analysis using SAS system: 2nd edition. *SAS Publishing*, 471-550.

[38] Mclean, A. R. and Michie, C. A. (1995). In vivo estimates of division and death rates of human t lymphocytes. *Proceedings of the National Academy of Sciences of the USA* **92**, 3707-3711.

[39] Mcullagh, P. (1983). Quasi-likelihood functions. *The Annals of Statistics* **11**, 59-67.

[40] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Second Edition, London: Chapman and Hal.

[41] Mohri, H., Bonhoeffer, S., Monard, S., Perelson, A. S. and Ho, D. D., (1998). Rapid turnover of T lymphocytes in SIV-infected rhesus macaques. *Science* **279**, 1223-1227.

[42] Molenberghs, G. and Verbeke, G. (2005). *Models for discrete longitudinal data*. Springer Series in Statistics, Springer-Verlag, New-York.

[43] Molina, J. M., Journot, V., Ferchal, F., Pellegrin, I., Rancinan, C., Sombardier, M. N. and Decazes, J. M. (1998). *ALBI (ANRS 070)*. International Conference on AIDS, 12.

[44] Mwambi, H., Ramroop, S., White, L. J., Okiro, E. A., Nokes, D. J., Shkedy, Z. and Molenberghs, G. (2010). A frequentist approach to estimating the force of infection for a respiratory disease using repated measurements data from a birth cohort. *Statistical Methods in Medical Research* (In press).

[45] Pantaleo, G., Demarest, J. F., Schacker, T., Vaccarezza, M., Cohen, O. J., Daucher, M., Graziosi, C., Schnittman, S. S., Quinn, T. C., Shaw, G. M., Perrin, L., Tambussi, G., Lazzarin, A., Sekaly, R. P., Soudeyns, H., Corey, L. and Fauci, A. S. (1997). The qualitative nature of the primary immune response to HIV infection is a prognosticator of disease progression independent of the initial level of plasma viremia. *Proceedings of the National Academy of Sciences of the United States of America* **94(1)**, 254258, doi:10.1073/pnas.94.1.254. PMID 8990195.

[46] Perelson, A. S. and Nelson, W. P. (1999). Mathematical analysis of HIV-1 dynamics in vivo. *SIAM REVIEW* **41, 1**, 3-44.

[47] Piatak, M., Jr, Saag, M. S., Yang, L. C., Clark, S. J., Kappes, J. C., Luk, K. C., Hahn, B. H., Shaw, G. M. and Lifson, J.D. (1993). High levels of HIV-1 in plasma during all stages of infection determined by competitive PCR *Science* **259 (5102)**, 17491754, doi:10.1126/science.8096089. PMID 8096089.

[48] Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-effects models in S and* $S-PLUS$. London:Springer.

[49] Prentice, R. L. (1989). Surrogate endpoints in clinical trials: Definitions and operational criteria. *Statistics in Medicine* **8**, 415-425.

[50] Qu, Y. and Case, M. (2006). Quantifying the indirect treatment effects via surrogate markers. *Statistics in Medicine* **25**, 223-231.

[51] Qu, Y. and Case, M. (2007). Quantifying the effect of the surrogate marker by information gain. *Biometrics* **63**, 958-963.

[52] Ramratnam, B., Bonhoeffer, S., Binley, J., Hurley, A., Zhang, L., Mittler, J. E., Markowitz, M., Moore, J. P., Perelson, A. S. and Ho D. D. (1999). Rapid production and clearance of HIV-1 and hapatitis C virus assessed by large volume plasma apheresis. *The Lancet* **354**, 1782-1786.

[53] Riberio, R. M., Mohri, H., Ho, D. D. and Perelson, A. S. (2002). In vivo dynamics of T cell activation, proliferation, and death in HIV-1 infection: Why are CD4 but not CD8 T cells depleted? *Proceedings of the National Academy of Sciences* **24,15**, 1572-15, 577.

[54] The Ripple effect. (2009). It is time for a "No nonsense" approach to HIV. *http://www.rippleweb.co.za.*

[55] Rosenbaum, P. R. and Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* **70**, 41-55.

[56] Thiebaut, R., Jacqmin-Gadda, H., Leport, C., Katlama, C., Costagliola, D., Le Moing, V., Morlat, P., Chene, G. and the APROCO Study Group. (2003). Bivariate longitudinal model for the analysis of evolution of HIV RNA and $CD4$ cell count in HIV infection taking into account left censoring of HIV RNA measures. *Journal of Biopharmaceutical Statistics* **13**, 271-282.

[57] UNAIDS. (2008). Report on the Global AIDS Epidemic. Geneva.

[58] Verbeke, G. and Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. Springer Series in Statistics Springer-Verlag, New-York.

[59] Volkow, P., Mohar, A. and Terrazas, J. J. (2002). Changing risk factors for HIV infection. *Arch Med Res* **33**, 61-66.

[60] Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss-Newton Method. *Biometrika* **61**, 439-447.

[61] World Health Organisation. Interim proposal for a WHO staging system for HIV infection and disease, *Weekly Epidemiological Record* **65**, 221-224.

[62] Weir, C. J. and Walley, R. J. (2006). Statistical evaluation of biomarkers as surrogate endpoints: a literature review. *Statistics in Medicine* **25**, 183-203.

[63] Werner, L. (2009). *Modelling acute HIV infection using longitudinally measured biomarker data including informative drop-out*. MSc thesis.

[64] Wikipedia. HIV. *http://en.wikipedia.org/wiki/HIV*.

[65] Zeger, S. L., and Liang, K. Y. (1986). The analysis of discrete and continuous longitudinal data. *Biometrics* **42**, 121-130.

APPENDIX

SAS code used to fit the GEE models as described in the text. Code1 is for finding the data means categorized by gender. Code 2 is for fitting the GEE model for death status with $CD4$ count and viral load as continuous variables. Code 3 is fitting the GEE model and the contrast estimate results. Code 4 is fitting the GEE model for death status with $CD4$ count and viral load as categorical variables.

```
/* Code 1*/
proc means data=oln N mean std maxdec=2;
Title "Output from proc means";
class Gender;
var CD4count Viralload Height Weight Age;
run;


/* Code 2*/
proc genmod data=death descending;
class Pid Gender Visit;
model DeathStatus=Gender Age CD4count Viralload Height Weight / link=logit dist=bin type3;
repeated subject=Pid/ corr=exch covb corrw;
run;


/*Code 3*/
ods select Estimates;
proc genmod data=death descending;
class Pid Gender Visit;
model DeathStatus=Gender Age CD4count Viralload Height Weight / link=logit dist=bin type3;
repeated subject=Pid/ corr=exch covb corrw;
estimate 'CD4 count' CD4count 1/exp;
estimate 'viral load' Viralload 1/exp;
```

```
run;


/* Code 4*/
proc genmod data=death descending;
class Pid Gender Visit CD4count Viralload;
model DeathStatus=Gender Age CD4count Viralload Height Weight / link=logit dist=bin type3;
repeated subject=Pid/ corr=exch covb corrw;
run;
```