UNIVERSITY OF CAPE TOWN

Modelling relationships between clinical markers of the Human Immunodeficiency Virus disease in a South African population

Dept of Statistical Science

Faculty of Business Science

Freedom N. Gumedze

TABLE OF CONTENTS

DECLARATION

I declare that the work on which this thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other University.

I empower the University to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

.....

Freedom N. Gumedze

ACKNOWLEDGEMENTS

My many thanks to Assoc. Professor June Juritz for her expert guidance, encouragement and support in supervising this work. While leaving me to my own creative processes, she provided direction and shared invaluable insights during the preparation of this thesis.

I wish to also extend my appreciation to Dr. Robin Wood for providing me with the data for this thesis and for his assistance with the biological interpretations of the results.

Thanks to staff and post-graduate students of the Department of Statistical Sciences, for they have enriched my experience more than they know, especially Franseca Little for the computing support. I owe a debt of gratitude to Assoc. Professor Tim Dunne for providing me with opportunities to advance my academic career.

My friends and family have given me all the encouragement and assistance that they possibly could while I was working on this thesis, but none more so than Mduduzi and Lungile, for whose support I am most grateful.

I am indebted to the Central Statistical Office of the Swaziland Government for providing the funding for this work.

Thank you Lulu for your patience, love and encouragement in seeing this project through with me.

ABSTRACT

This study investigated relationships between the CD4 count and other clinical markers of the HIV disease, total lymphocyte count and viral load, in a South African population.

The CD4 count has been an important clinical marker of disease progression in HIV infected individuals and has been the focus of many studies in developed countries. Most of the studies reported in the literature have been done using data from well-defined cohorts of HIV patients. Similar studies in Africa do not appear to have been done.

This study used clinical records of HIV infected individuals attending the Somerset Hospital HIV Clinic, over the period 1984-97, to study the relationship between the CD4 count and total lymphocyte count. From a practical perspective this relationship is important in South Africa for two reasons. Firstly, a majority of the HIV infected population is poor and can not afford the higher costs associated with the measurement of the CD4 count instead of the total lymphocyte count. Secondly, in many small clinics or hospitals in South Africa the equipment for measuring the CD4 count is generally not available but the equipment for measuring the total lymphocyte count is widely available.

Random effects models were used to examine the relationship between the CD4 count and the total lymphocyte count. The analysis revealed that there was a weak relationship between the CD4 count and the TLC. The relationship was not strong enough to allow for satisfactory prediction of the CD4 count from the TLC.

In clinical practice different thresholds of the CD4 count are usually used to determine whether initiating treatment is desirable. A CD4 count < 200 is generally used as a criterion for commencing treatment against toxoplasmosis and pneumocystis carinii pneumonia. The results from the analysis were assessed for use in clinical practice by calculating sensitivity, specificity and predictive value measures for the total lymphocyte count for detecting specified thresholds of CD4 counts.

The study also investigated the relationship the CD4 count and the viral load using data from HIV patients enrolled in a clinical trial at the Somerset Hospital HIV Clinic. The results from this analysis were in agreement with those from other similar studies conducted in developed countries. This implies that the pathogenesis of HIV is the same in both African patients and those in developed countries.

Other approaches to modelling the relationships between the CD4 count and the total lymphocyte count, and the viral load were briefly investigated in the final chapter of the study. These approaches included hierarchical generalized linear models and Bayesian models for the CD4 counts.

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Background

The Human Immunodeficiency Virus (HIV) spreads from one person to another through sexual intercourse, direct exposure to contaminated blood, or transmission from a mother to her baby. In the body the virus invades certain cells of the immune system, replicates inside them and spreads to other cells. The body tries to compensate for the loss by making new cells, but the immune system remains under constant siege and eventually fails to keep up.

HIV infection is normally monitored by laboratory and clinical markers of disease progression. In general a marker of disease progression indicates the state of disease advancement in an individual. Markers are useful in disease staging and assessing future prognosis. When an individual is infected with HIV, the state of the immune system reflects the state of the disease in the host. CD4 lymphocyte count is the most quoted marker of HIV disease progression; other markers include total lymphocyte count (TLC) and viral load. In the context of monitoring HIV patients, markers are used for initiating and monitoring antiretroviral therapy, and for assessing therapeutic effects in clinical trials.

Since 1985 a plethora of statistical methods has been used to study markers of HIV disease progression, especially CD4 counts (DeGruttola *et al.*, 1991; Lange *et al.*, 1989; Moss *et al.*, 1989; Munoz *et al.*, 1988; Segal *et al.*, 1994; Taylor *et al.*, 1989 and 1994; Vittinghof *et al.*, 1994). However, even after several years of statistical endeavour and innovation there are no 'standard operating procedures' for the analysis of markers (e.g. CD4 cell count) in relation to HIV progression.

An approach to the analysis of markers of HIV disease progression is to use regression methods to describe the histories of markers. In this approach, periodic observations of a variable thought to reflect underlying progression of disease are available on each of a number of patients. The focus of the modelling effort in this approach is characterization of the underlying stochastic process generating the observations i.e. the model describes changes in the marker of disease progression over time. Repeated measurements of markers have also been used to model relationship between the marker markers and time to an event such as AIDS and or death, adjusting for fixed and times dependent covariates.

In this study, we investigate models that describe the dependence of one marker of disease progression (CD4 cell count) on either TLC count or viral load.

1.2 Objectives of the study

The primary objectives of the research were to:

i) investigate the relationship between the CD4 count and TLC in HIV infected individuals,

ii) examine the relationship between the CD4 count and viral load.

1.3 Plan of Thesis

Chapter 2 presents a model for the relationship between CD4 cell counts and TLC counts, in HIV infected individuals. The aim of this Chapter is to investigate if CD4 could be predicted from TLC, given the HIV stage. In Chapter 3 the relationship between CD4 counts and viral load is investigated using data from two clinical studies. Initial investigation of the relationship between the CD4 count and viral load is conducted using simple linear regression on single measurements of CD4 and viral load for a number of HIV patients. Repeated measures of the markers on a number of subjects are then used to examine the relationship between the two markers. Chapter 4 presents some other approaches to modelling the relationship between the CD4 count and TLC and the association between the CD4 count and viral load. These approaches make particular distributional assumptions about the CD4 counts and employ different statistical techniques for the estimation of the unknown regression parameters. A number of possible models are proposed and their results are compared. Chapter 5 draws conclusions from the analyses and also puts forward suggestions for further research in modelling the CD4 counts.

1.4 Data Sources

1.4.1 Medical records of patients attending the Somerset Hospital HIV Clinic, 1984-97

Somerset Hospital HIV Clinic has been a referral centre for HIV seropositive patients from the Western Cape since 1984, following the identification of the first AIDS cases in Cape Town. Retrospective clinical chart reviews were performed on patient records prior to January 1992. Demographic details, HIV risk behaviours, laboratory results and clinical data of patients were extracted and entered into a computer database using the Epi Info software package. From January 1992, the data were collected prospectively.

The patients were clinically staged according to the World Health Organization (WHO) HIV staging system, in which stage 4 is equivalent to the 1987 Center for Disease Control definition of AIDS. CD4 counts were measured approximately every six months, by flow cytometry and total lymphocyte counts were determined by automated blood cell counter. For purposes of this study the file has been closed on July 1997, in which 1686 HIV-infected individuals had presented to Somerset Hospital HIV Clinic i.e. the data set consists of all patients presenting to the clinic over the 13-year period from 1984 to 1997.

This data set was used to examine the relationship between the CD4 count and TLC.

1.4.1.1 Key features of the data

The data possess special features that distinguish them from data that have been previously used to study HIV disease progression. The study was essentially an observational study and not a designed study. The following is an outline of the features of the data:

(a) The study subjects are neither a cohort nor a sample from a larger cohort of HIV infectives.
(b) The duration of HIV infection (the length of time a patient has been infected with HIV) is not known.
(c) Both patients with AIDS and HIV-infected individuals, who are at various stages of the disease, are included in the study.
(d) Time intervals for clinical and laboratory review vary by patient and clinical and laboratory information may be missing/incomplete for some patients.

The features of the data impose certain restrictions on the statistical analyses that may be performed. Some of these restrictions may be overcome; on the other hand there may not be practical remedies for some of the constraints imposed by the nature of the data. For instance, missing values for the markers for a particular patient lead to an unbalanced data structure. Any model of CD4 counts needs to take into account the unbalance structure of the data in the estimation of unknown regression parameters.

1.4.2 Clinical studies

The data came from two clinical studies, involving HIV patients, conducted at the Somerset Hospital HIV Clinic. The first study was a cross-sectional one, involving 123 HIV patients, and the data with consisted of single measurements of CD4 counts and viral load. The second study was a clinical trial, involving 116 HIV patients, and the data consisted of repeated measurements of CD4 counts and viral load. In both studies selected patients were neither on antiretroviral therapy nor did they suffer any intercurrent infections prior to study entry.

The data from the cross-sectional study was used for preliminary investigation of the relationship between the CD4 count and viral load. While the data from the clinical trial was used to investigate aspects of variability in the CD4 count and to examine the association between repeatedly measured CD4 counts and viral load.

CHAPTER 2

The relationship between the CD4 count and Total lymphocyte count

2.1 Introduction

The CD4 count is considered to be an important prognostic indicator for disease progression in HIV infected individuals; however it (CD4 count) is costly to measure especially in resource poor countries. It is postulated that total lymphocyte count can also provide equally useful information about the immune status of HIV infected individuals. The total lymphocyte count, TLC, has been found to be a good predictor of CD4 count (Moss *et al.*, 1988; Post *et al.*, 1996; Sloan *et al.*, 1991). Several authors have also found the TLC to be a useful marker for staging the HIV disease (Montaner *et al.*, 1992; World Health Organisation global programme on AIDS, 1993). Post *et al.*, (1996), compared TLC with CD4 counts as predictors of AIDS onset and death, in HIV-positive patients and they concluded that CD4 and TLC were equally good predictors of HIV disease progression.

A common approach to analysing the CD4 count or TLC is to use survival analysis. This is done by estimating Kaplan Meier survival functions or using the proportional hazards model (Bachetti *et al.*, 1992; Coates *et al.*, 1992; Graham *et al.*, 1993; Post *et al.*, 1996) to establish a relationship between the marker some event of interest during the HIV infection period such as AIDS or death. Raboud *et al.*, (1993) extended this approach by using CD4 count as a time-dependent covariate in a proportional hazards model. The time-dependent covariate model was found to be inadequate because marker values were highly variable which led to biased estimates of the regression parameters. Other authors have simultaneously modelled the CD4 values as a function of time and the risk of CD4 cell count on disease progression (DeGruttola *et al.*, 1991; Faucet and Thomas, 1996; Tsiatis *et al.*, 1992). Here, the focus of the modelling effort was to describe the evolution of the CD4 count over time and its relationship to either onset of AIDS or death. The strength of these studies was that the variation in the CD4 count was explicitly taken into account. Several authors have also considered the variability of markers of HIV disease progression, especially CD4 counts (DeGruttola and Tu, 1992; Hughes *et al.*, 1994; Malone *et al.*, 1990; Self and Pawitan, 1992).

Although other markers have been evaluated as predictors of CD4 count, such as CD8 count, heamoglobin, platelet counts, cytomegalovirus (CMV) and hepatitis B surface antigen. (Munoz, *et al.*, 1988); the TLC has not been assessed extensively as a potential clinical marker for HIV disease progression.

Moreover, most of the studies have been done using data from developed countries, where the cost of clinical examinations are affordable and the data are collected from well defined cohorts of HIV patients. Similar studies in Africa are hindered by the unavailability of reliable data.

In Africa, a majority of the HIV infected population is poor and can not afford the costs of necessary clinical examinations such as the measuring of CD4 counts (Post *et al.*, 1996; Maartens *et al.*, 1997). For instance, in South Africa the costs of measuring CD4 count and TLC are R110 and R40, respectively. Furthermore, in many clinics or hospitals in Africa the equipment for measuring clinical markers, especially the CD4 count, is usually not available. Since it is cheaper to measure TLC than CD4 count, in a clinical setting there is a need to estimate CD4 count given TLC and some other clinical information on the host. This requires the use of appropriate statistical methods that allow for such estimation of CD4 count from TLC, in a systematic manner.

In this study, CD4 count in a given patient was expressed as a function of TLC, HIV stage, and age at first clinic visit, presence of extrapulmonary tubercolosis and time since first clinic visit. It was also of interest to distinguish components of variation in the CD4 count that may be due to random error or changes in laboratory technique from those components of change that are due to biological factors in the host.

The CD4 cell counts were modelled using a natural logarithmic transformation. This scale improves the satisfaction of the assumption of normality over that with analysis of untransformed CD4 cell counts. A number of different transformations of CD4 counts have been considered in the literature. Taylor *et al.*, (1994) transformed the CD4 counts by a fourth-root power to achieve homogeneity of within-subject variance. Hughes *et al.*, (1994) used a natural logarithmic transformation to study CD4 cell variation in HIV infection. Coates *et al.*, (1992) have argued that untransformed CD4 cell counts are adequate. Sabin, (1995) asserted that logarithmic or square-root transformations of CD4 count provided more biologically plausible results. In this study, the logarithmic transformation of the CD4 count was used.

2.2 Objectives

The goals in this analysis were to:

i) investigate the relationship between CD4 count and TLC, i.e. to provide an insight into the question of whether TLC can be used as a surrogate for CD4 count in HIV patients;

ii) to investigate whether the relationship between CD4 count and TLC is the same at all stages of the disease, or whether the relationship becomes stronger or weaker with disease progression and whether the relationship depends on patient's age, the presence of extrapulmonary tubercolosis and time since first visit;

iii) estimate variation in the CD4 counts that is due to the patient (biological variation) and that, which is due to random error.

2.3 Methodology

2.3.1 Description of the data

The data used for this study was from the clinical records of 1686 patients presenting with HIV infection at the Somerset Hospital HIV Clinic over the period 1984 to 1997. Table 2.1 gives the distribution of patients for selected clinical and demographic characteristics. Not all patients had data available in all variables. In Cape Town, the HIV epidemic affected mostly whites and the homosexual population, before 1990 (Wood *et al.*, 1996). Around the beginning of 1990, HIV disease developed into a disease for heterosexuals and began to infect women on a much larger scale. During this period the disease also started to invade the African population, vigorously. Wood *et al.*, (1996) presented a lucid description of the profile of patients presenting to the Somerset Hospital HIV Clinic during the period 1984 to 1995. They showed that the change in HIV transmission pattern from homosexual to heterosexual, over this period, resulted in changes in both the demographic profile of HIV patients and their HIV presentation. They reported that the frequency of opportunistic infections differed from both Central Africa and developed countries but was dominated by a high prevalence of tuberculosis in the population studied.

Although clinical examinations were supposed to be performed every six months, patients often missed clinic visits and as a result CD4 count, TLC and HIV stage were not recorded. The reasons for the missing of clinic visits by patients were mainly of a social nature, so that missing CD4 counts and TLC could be considered to be missing at random. In addition the CD4 and TLC were not ascertained at all visits the patient made to the clinic. Before 1993, it was hospital policy not to measure CD4 count on all patients at each visit made to the hospital. This decision was made in order to reduce the costs of clinical examinations that are necessary for patient monitoring.

In this analysis, only records with both CD4 and TLC measurements were considered. There were 1333 patients with both CD4 and TLC measurements. A random sample of 376 patients was drawn from the 1333 patients. The data set from the sampling contained 937 observations. Table 2.2 shows the distribution of the 1333 patients and those from the random sample by selected characteristics. Not all 1333 patients had data available in all variables. Patients with complete cases were not different from those sampled in a way that might affect the results. The sampled data was then used to describe the relationship and to investigate how well CD4 counts could be estimated from TLC.

The reasons for taking simple random sample of the data were:

i) The sample would be representative of the type of data on HIV infected individuals obtained in clinical practice in South Africa.

ii) The use of the random sample would allow us to check the model using the rest of the data.

iii) If we had employed a more complex sampling scheme such as stratified or cluster sampling, we would have had to allow for this in the regression methods we have used to analyze the data.

Tables 2.3 and 2.4 present the summary statistics of the CD4 and the log-transformed CD4 counts by HIV stage, respectively, from the sampled data. While Tables 2.5 and 2.6 show summary statistics of the TLC and log-transformed TLC by HIV stage.

CD4 counts ranged from 5 to 1384 counts, with a mean and standard deviation of 306.2 and 225.7, respectively. There was a large variation in the TLC measurements ranging from 120 to 7850. The mean and standard deviation of the TLC measurements were found to be 1741 and 917.4, respectively. Variation in the both log-transformed CD4 and TLC was found to increase with stage of HIV disease, after stage II (Table 2.4 and 2.6).

Table 2.1 Frequency distribution of HIV patients attending the Somerset Hospital HIV Clinic during the period 1984-1997 by selected characteristics

Characteristic	Period of first clinic visit							
	1984-89		1990-93		1994-97		Total	
	No. of patients	Percent	No. of patients	Percent	No. of patients	Percent	No. of patients	Percent
Gender								
Female	3	2.6	178	30.3	501	51.0	682	59.5
Male	112	97.4	410	69.7	481	49.0	1003	40.5
Sexual orientation								
Homosexual	69	60.0	157	26.9	84	9.6	310	19.7
Heterosexual	11	9.6	338	57.9	742	84.7	1091	69.3
Bisexual	21	18.3	43	7.4	12	1.4	76	4.8
Other	14	12.2	46	7.9	38	4.3	98	6.2
Population group								
White	85	73.9	186	31.9	115	12.5	386	23.9
Coloured	24	20.9	167	28.6	179	19.5	370	22.9
African	3	2.61	227	38.9	615	67.1	845	52.3
Other	3	2.6	3	0.5	8	0.9	14	0.9
HIV Stage								
I	42	36.5	241	41.7	320	36.5	603	38.8
II	23	20.0	66	11.4	155	17.7	244	15.7
III	25	21.7	168	29.1	261	29.8	454	29.2
IV	25	21.7	97	16.8	133	15.2	255	16.4
Age								
10-25	19	16.8	132	22.7	231	23.9	382	23.0
25-35	45	39.8	268	46.1	415	43.0	728	43.9
35-45	36	31.9	124	21.3	237	24.5	397	23.9
45+	13	11.5	57	9.8	83	8.6	153	9.2
TB status								
With TB	7	6.1	67	11.3	89	9.1	163	9.7
Without TB	108	93.9	521	88.7	894	90.9	1523	90.3

Table 2.2 Frequency distribution of patients, with both CD4 and TLC measurements for all the observations and the random sample, by selected characteristics

Characteristic	All the data		Sampled data	
	No. of Patients	Percentage	No. of Patients	Percentage
HIV Stage				
I	368	27.6	168	44.6
II	274	20.6	65	17.3
III	443	33.2	80	21.3
IV	248	18.6	63	16.8
Total	1333	100.0	376	100.0
Age				
10-25	302	22.7	82	21.8
25-35	586	44.0	172	45.7
35-45	307	23.1	92	24.5
45+	136	10.2	30	8.0
Total	1331	100.0	376	100.0
TB status				
With TB	150	11.3	41	10.9
Without TB	1183	88.7	335	89.1
Total	1333	100.0	376	100.0
Number of visits per patient				
1	694	52.1	194	51.6
2	267	20.0	81	21.5
3	141	10.6	28	7.4
4	79	5.9	16	4.3
> 5	152	11.4	57	15.2
Total	1333	100.0	376	100.0

Table 2.3 Summary Statistics of CD4 counts by HIV stage

Stage	No. of obs.	Mean	Min.	Max.	Std. Dev.	Med.	Coeff. of var.
I	294	409.9	30	1216	215.95	381	0.53
II	188	408.2	52	1384	241.30	374	0.59
III	288	223.3	10	968	117.05	190	0.52
IV	167	253.5	5	755	150.02	106	0.99
Total	937	306.21	5	1384	225.68	262	0.74

Table 2.4 Summary Statistics of Log CD4 counts by HIV stage

Stage	No. of obs.	Mean	Min.	Max.	Std. Dev.	Med.	Coeff. Of var.
I	294	5.9	3.4	7.1	0.61	5.9	0.10
II	188	5.8	4.0	7.2	0.68	5.9	0.12
III	288	5.1	2.3	6.9	0.86	5.2	0.17
IV	167	4.4	1.6	6.6	1.20	4.7	0.27
Total	937	5.4	1.6	7.2	1.01	5.6	0.19

Table 2.5 Summary Statistics of TLC by HIV stage

Stage	No. of obs.	Mean	Min.	Max.	Std. Dev.	Med.	Coeff. of var.
I	294	1966	350	4750	783.09	1850	0.40
II	188	1946	310	3900	689.91	1835	0.35
III	288	1570	230	7270	907.54	1390	0.58
IV	167	1411	120	7850	1185.25	1130	0.84
Total	937	1741	120	7850	917.36	1620	0.53

Table 2.6 Summary Statistics of Log TLC by HIV stage

Stage	No. of obs.	Mean	Min.	Max.	Std. Dev.	Med.	Coeff. of var.
I	294	7.5	5.9	8.5	0.41	7.5	0.06
II	188	7.5	5.7	8.3	0.40	7.5	0.05
III	288	7.2	5.4	8.9	0.55	7.2	0.08
IV	167	7.0	4.9	9.0	0.78	7.0	0.11
Total	937	7.3	4.9	9.0	0.57	7.4	0.08

Figure 2.1 and Figure 2.2 show distributions of the log CD4 count and log TLC by HIV stage, respectively. Figure 2.3 and Figure 2.4 show the Box –Whisker plots of CD4 counts and TLC, respectively, while Figure 2.5 and 2.6 are Box-Whisker plots of the log CD4 count and log TLC, by HIV stage, respectively. The plots show that both CD4 and TLC start to decline steadily after stage II. The rate of decline for both CD4 and TLC then accelerates between stage III and IV.

A scatter plot of CD4 count against TLC is shown in Figure 2.7. A plot of the log-CD4 counts against log TLC suggested a linear relationship between the two markers (Figure 2.8). The scatterplot of the log-transformed variables by HIV stage also shows that observations of patients 68, 122 and 335 were extreme. These observations were suspected to be outliers. Further examination of the data revealed that these observations were low CD4 counts with high corresponding TLC values and they were taken on patients in advanced stages of the HIV disease (stage III and IV), (Figure 2.9).

Figure 2.1. Histogram of Log CD4 count by HIV Stage

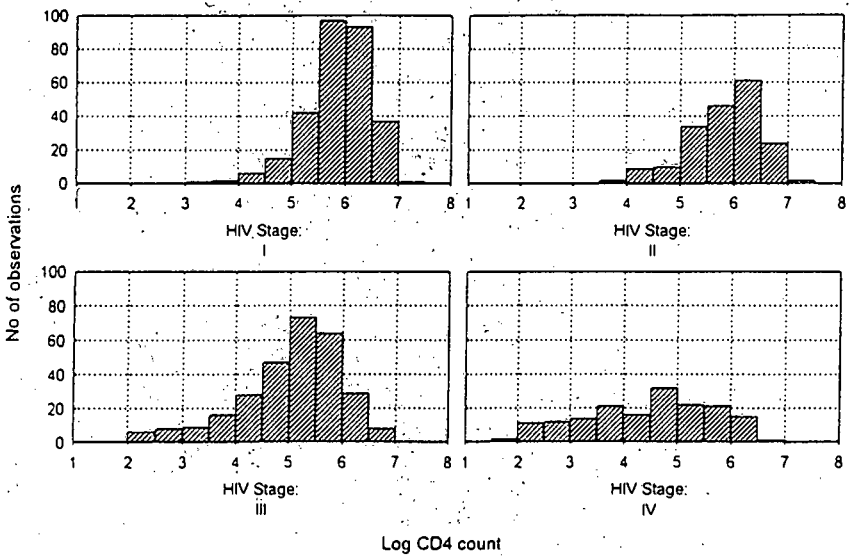


Figure 2.2. Histogram of Log TLC by HIV Stage

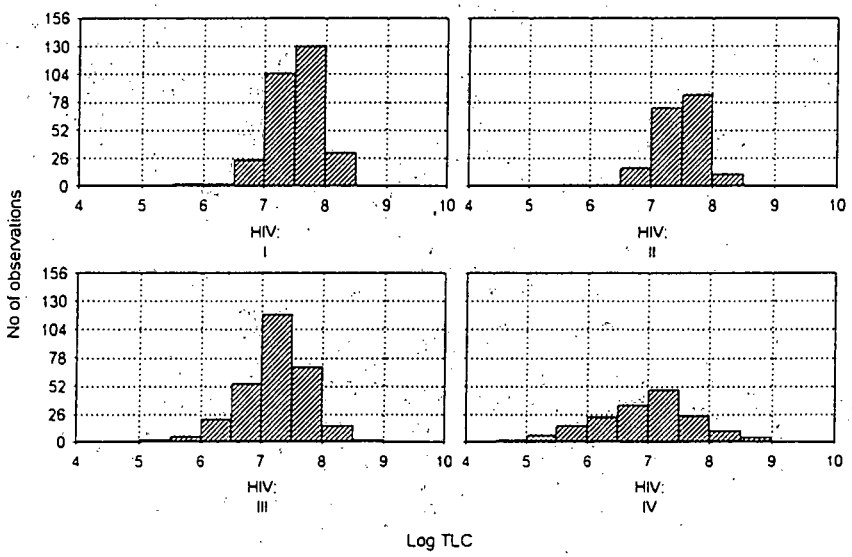


Figure 2.3. Box & Whisker Plot of CD4 count by HIV Stage

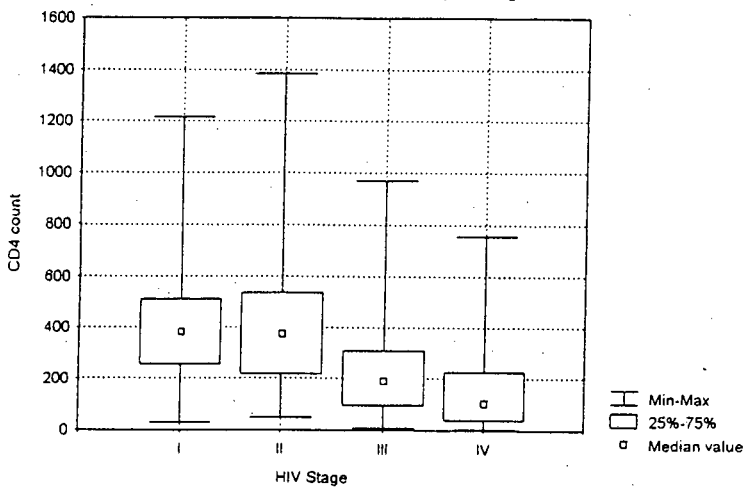


Figure 2.4. Box & Whisker Plot of TLC by HIV Stage

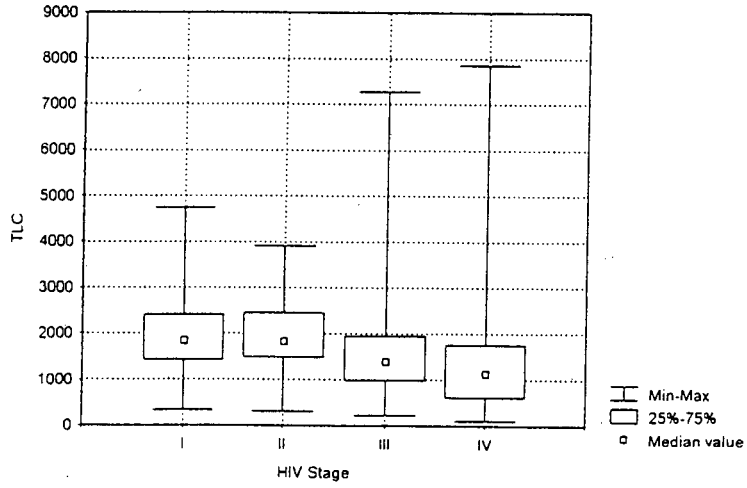


Figure 2.5. Box & Whisker Plot of Log CD4 count by HIV Stage

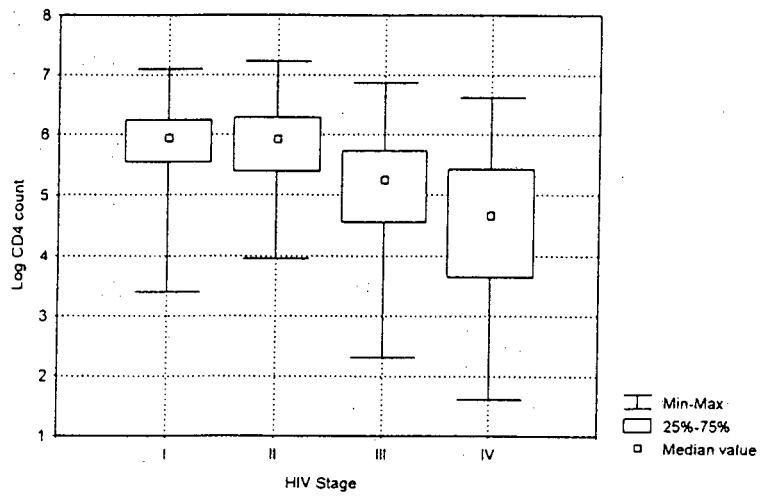


Figure 2.6. Box & Whisker Plot og LOG TLC by HIV Stage

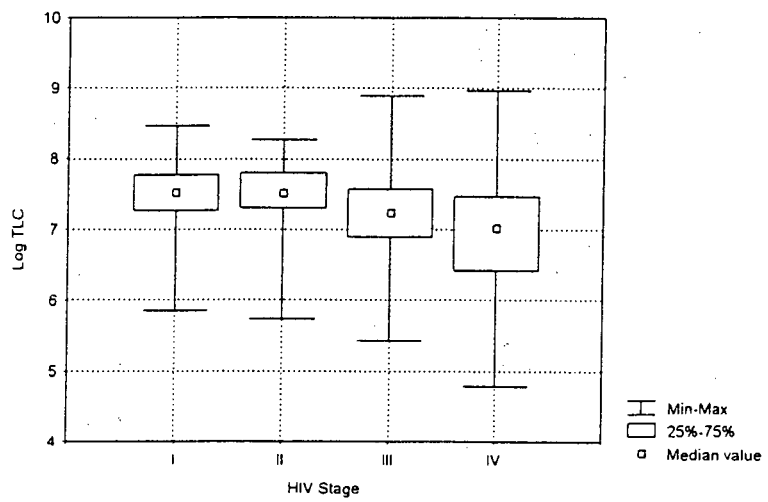


Figure 2.7. Scatterplot of CD4 count against TLC

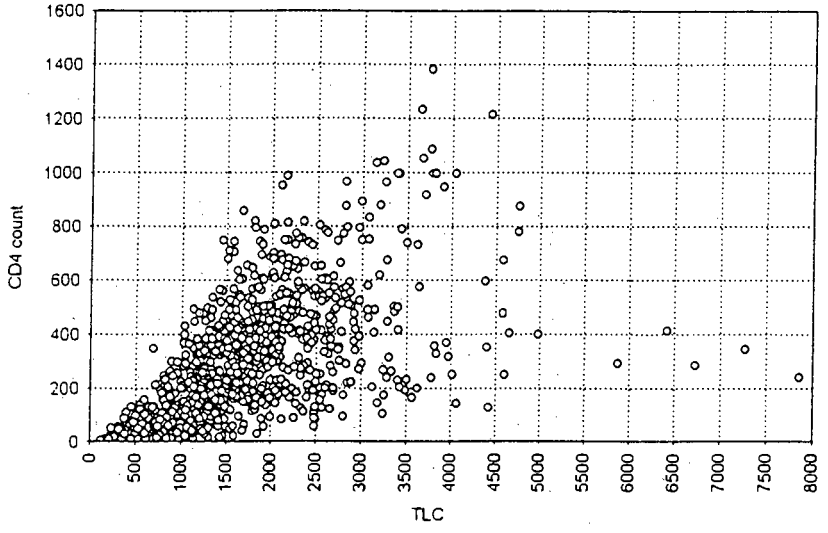


Figure 2.8. Scatterplot of Log CD4 count against Log TLC

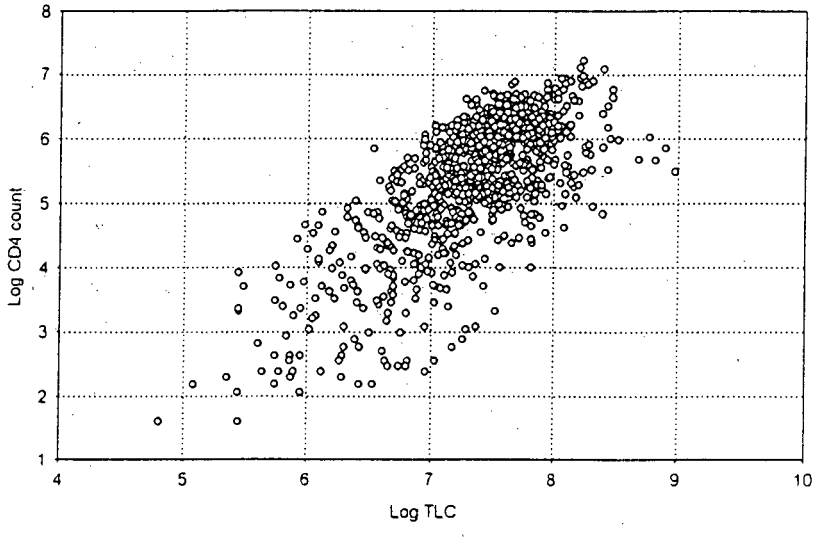
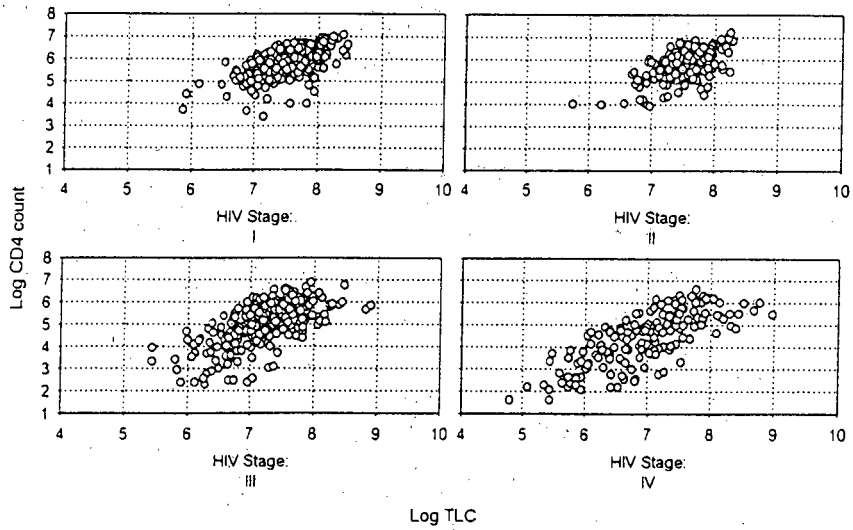


Figure 2.9. Scatterplot of Log CD4 count against Log TLC
by HIV Stage



2.3.2 Statistical Methods

One of the primary objectives of this study was to examine the relationship between two markers (CD4 count and TLC) and also identify covariates that might affect the relationship between the two markers. We also wanted to account for the correlation among observations on each patient.

The timing and total number of measurements recorded varied from patient to patient so that the data was unbalanced and unequally spaced.

We considered the linear mixed model (random effects) to be appropriate for this data. The following is a brief description of the general form of a linear mixed model we employ in the analysis.

The linear mixed model has the form (Laird and Ware, 1982; Lindstrom and Bates, 1988)

$$y = X\beta + Zu + e$$

where:

y is a vector of n observations. β and u are vectors of p and q unknown fixed and random effects parameters, respectively. X and Z are known design matrices for the vector of fixed and random effects, respectively. e is an unknown random error vector.

Both random effects and random error are assumed to be Normal distributed such that $E(u) = 0$, $E(e) = 0$ and

$$\mathrm{Var}\begin{pmatrix} u \\ e \end{pmatrix} = \begin{pmatrix} G & 0 \\ 0 & R \end{pmatrix}\sigma^2$$

where G and R are known positive definite matrices and σ^2 is a positive constant. The matrices G and R can be allowed to have a more general structures to model correlated errors. u is independent of e.

If the matrix R is taken to be an identity matrix, then estimates of β and u are solutions to the mixed model equations (Henderson, 1950):

$$\begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z + G^{-1} \end{pmatrix}\begin{pmatrix} \beta \\ u \end{pmatrix} = \begin{pmatrix} X'Y \\ Z'Y \end{pmatrix}$$

The fixed effects are then estimated by generalized least squares:

$$\hat{\beta} = \left(X'V^{-1}X\right)^{-1}X'\hat{V}^{-1}Y$$

with $\mathrm{Var}\left(\hat{\beta}\right) = \left(X'\hat{V}^{-1}X\right)^{-1}$

where $V = (ZGZ' + I)\sigma^2$, is the variance of the data y.

Estimates of the random effects are given by:

$$\hat{u} = (Z'Z + G^{-1})^{-1} Z'(Y - X\hat{\beta})$$

with $\text{Var}(\hat{u}) = \left(Z'Z + G^{-1}\right)^{-1}\sigma^2$

The estimates of the fixed effects are BLUE (best linear unbiased estimators) while estimates of the random effects are BLUP (best linear unbiased predictors).

The variance covariance matrix of the fixed and random effects is given by (Henderson, 1975):

$$\text{Var}\begin{pmatrix}\beta \\ u\end{pmatrix} = \sigma^2 \begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z + G^{-1}\end{pmatrix}^{-1}$$

The most common statistical technique for estimating model parameters is Maximum likelihood (ML). A known property of ML estimation is that in estimating the variance components it takes no account of the degrees of freedom involved in estimating the fixed effects, so that the variance components are biased. For example when data $x_1, \ldots, x_n$ are n independent observations in a simple random sample from a Normal distribution with mean μ and variance σ^2, the usual unbiased estimate of σ^2 is $\frac{\sum_i (x_i - \bar{x})^2}{n-1}$ whereas the MLE, $\hat{\sigma}^2$, is $\frac{\sum_i (x_i - \bar{x})^2}{n}$.

Residual (or restricted) maximum likelihood (REML) overcomes the problem with ML estimation of not taking into account the degrees of freedom used in estimating the fixed effects when estimating the variance components. The method was introduced in random effects models by Patterson and Thompson (1971 and 1974) and has received extensive attention in the literature (Gilmour *et al.*, 1995; Harville, 1974 and 1977; Khuri *et al.*, 1985; Robinson, 1987; Searle, Casella and McCulloch, 1992; Thompson, 1979). The basic idea behind REML is that the likelihood used to estimate the variance components is based on the residuals calculated after fitting only the fixed effects part of the model. REML includes no procedure for the estimation of fixed effects. However, it seems reasonable to use the REML estimates of the variance components in estimating, V, the variance of the data y (Searle, Casella and McCulloch, 1992). Thus the parameter estimates, $\hat{\beta}$, are unbiased estimates of the fixed effects, β, and their estimated standard errors are also unbiased because variability in the estimates of the variance components has been taking into account. Another way of looking at REML is that the method maximizes that part of the likelihood which is invariant to the fixed effects, β (Casella and Berger, 1990).

REML has wide applicability in scientific research. It is used to analyse data that arise in agricultural, biological, educational, industrial and medical research. The method

provides estimates of the sizes of variation and partitions the variability to different sources. It also allows for the estimation of random effects, BLUPS (Henderson, 1975; Robinson, 1991). The BLUPs are interpreted as predictions of the random effects given the data. REML is usually the preferred method of estimating unknown model parameters when the data is unbalanced or when there is more than one source of variation in the data.

The GENSTAT package was used to perform the analysis in this study. Estimation of variance components and fixed effects was conducted using the "REML" directive in GENSTAT (GENSTAT 5 Committee, 1993). Hypothesis tests of fixed effects were done by using the Wald test. Residual plots were used to check model adequacy.

2.4 Results

2.4.1 Model Specification

A random effects model for describing the relationship between CD4 count and TLC could be formulated as:

Let y_{ijkl} be the CD4 cell count at visit j for patient i in HIV stage k with TB status l

x_{ijkl} be the total lymphocyte count at visit j for patient i in HIV stage k with TB status l

The linear model for the i^{th} patient is given by:

$$\log y_{ijkl} = \mu + \beta_1 \log x_{ijkl} + \beta_2 age_i + \beta_3 t_{ij} + \alpha_k^{stage} + \delta_l^{TB} + u_i^{patient} + b_i x_{ij} + e_{ijkl} ,$$

for $j = 1,\ldots,n_i$ and $i = 1,\ldots,376$

where the **fixed** terms in the model are:

μ is the constant term,
β_1 is the effect of log TLC,
β_2 is the effect of age (years),
β_3 is the effect of time since first visit,
α_k is the effect of stage k, for $k = 1,\ldots,4$ with $\alpha_1 = 0$,
δ_l is the effect of extrapulmonary tuberculosis, for with $\delta_1 = 0$, if extrapulmonary tubercolosis is absent,

the **random** terms are:

u_i is the random effect for patient i, $i = 1,\ldots,376$
b_i is the effect of effect of log TLC on the i^{th} patient

and e_{ij} is the random error.

The model assumes that u_i is normally distributed with mean zero and variance σ_u^2, and e_{ijkl} is also normally distributed with mean zero and variance σ_e^2.

The variance of the CD4 count measured on patient i at visit j is given by

$$\text{var}(y_{ij}) = \sigma_e^2 + \sigma_u^2,$$

Since CD4 counts from the same patient are correlated; we have

$$\text{cov}(y_{ij}, y_{ij'}) = \sigma_u^2 \quad \text{for } j \neq j'$$

But observations from different patients are independent, i.e. the observations satisfy the condition

$$\text{cov}(y_{ij}, y_{i'j}) = 0 \quad \text{for } i \neq i'$$

The correlation between of two CD4 counts from the same patient is calculated as

$$\rho = \frac{\sigma_u^2}{\sigma_e^2 + \sigma_u^2}$$

This quantity is referred as the within patient correlation or variance component ratio (Longford, 1993, page 27). The variance component ratio measures the fraction of the total residual variation, which is due to between-patient variation.

The ratio of the variance components or variance ratio is defined as

$$\omega = \frac{\sigma_u^2}{\sigma_e^2}$$

This variance ratio measures the contribution of patient variation to variation relative to that of random error.

This model could also be considered as a random coefficient regression model (Longford, 1993; Swamy, 1971). This model would have allowed each patient to have random an intercept and slope, u_i and b_i, respectively. We fitted this model to the CD4 count data but it did not converge and so it was not considered further. The analysis was then done using the random effects model.

2.4.2 Model Selection

The relationship between CD4 count and TLC was assumed to depend on a patient's HIV stage, age, TB status and time on study. The purpose of the inclusion of the other variables (age, TB status and time since first visit) in the model was to assess whether these variables affected the relationship between the CD4 count and TLC. Tuberculosis has been found to interact with the HIV disease (De Cock *et al.*, 1992; Harries, 1990, Wood, 1999, *personal communication*). Post *et al.*, (1995) investigated the relationship between pulmonary tuberculosis in HIV infected patients and the CD4 count. They found that specific patterns of tuberculosis had positive predictive values of up to 100 percent for identifying patients with CD4 counts < 200. In this study we postulate that tuberculosis affects the relationship between the CD4 count and TLC.

The model selection process entailed fitting random effects models sequentially (Table 2.7). We examined the significance of all the main effects by fitting a random effects model to the data with patients as random effects and denoted this model as Model A. The Wald statistics for the fixed effects from this model are shown in Table 2.7. Only log TLC and HIV stage were found to be statistically significant and they were retained in the model. The other variables were then dropped from the model. In model B, we assessed the significance of an interaction between log TLC and HIV in a model. The Wald test indicated that this interaction term was not statistically significant and we therefore excluded it from the model. This meant that the relationship between CD4 count and TLC did not depend on the HIV stage of the patient i.e. CD4 count decline occurred with similar slopes in patients with different values of TLC, but intercepts were different for patients at different HIV stages. Figure 2.10 clearly depicts this finding.

Model C was the final fitted model. The final model included only log TLC and HIV stage as fixed effects and patients as random effects. We based all inferences on this model.

Table 2.8 gives the parameter obtained from fitting Model C (Table 2.7). The positive coefficient for the fixed effect of log TLC indicated that there was a positive relationship between the logarithm of CD4 and the logarithm of TLC i.e. a decrease in the log-transformed CD4 measurement was associated with a decrease in the log-transformed TLC.

We found that the log CD4 counts of patients in HIV stage II were not statistically different from those of patients in HIV stage I. (Table 2.8). However log CD4 counts of patients in HIV stage III and IV were statistically different from those of patients in HIV stage I. Furthermore, log CD4 counts of patients in HIV stage III were not statistically different from those of patients in HIV stage IV.

The within patient correlation in the CD4 counts, was found to be 0.576. This finding was not surprising given the wide intervals between CD4 count measurements. The variance ratio, the variation due to patient as a fraction of random error, was estimated to be 1.358; i.e. variation due to patient was found to be 1.358 times the variation due to random error. A plot of the fitted and observed log CD4 against log TLC is given in Figure 2.10. Table 2.9 presents correlation coefficients from the final fitted model by

HIV stage. The relationship seemed to improve with stage of HIV and it was strongest in stage IV, with correlation coefficient of 0.781; this implies that TLC could explain 62 percent of the variation in CD4 counts in patients who were in stage IV of the HIV disease.

In this analysis we have not explicitly taken into account the duration of HIV infection (the length of time a patient has been infected with HIV) because the dates of infection with HIV, for the study subjects, were not known nor could they be estimated. The bias that it could represent was afforded by the inclusion of the variable, time since first visit in the model. However we found no association between time since first visit and the CD4 count.

Table 2.7 Regression results of models fitted to the CD4 count data during the model selection process

Model A			Model B			Model C*		
Fixed term	**Wald statistic**	**d.f.**	**Fixed term**	**Wald statistic**	**d.f.**	**Fixed term**	**Wald statistic**	**d.f.**
Log TLC	662.1	1	Log TLC	1088.2	1	Log TLC	1086.7	1
$(\text{Log TLC})^2$	3.3	1	Stage	144.4	3	Stage	144.1	3
Stage	127.0	3	Log TLC × Stage	4.7	3			
Age	1.0	1						
TB	0.1	1						
Time	2.3	1						
Random term	**Variance Component**	**Std error**	**Random term**	**Variance Component**	**Std error**	**Random term**	**Variance Component**	**Std error**
Patient	0.235	0.025	Patient	0.239	0.026	Patient	0.239	0.026
Random error	0.173	0.010	Random error	0.176	0.010	Random error	0.176	0.010

* *Final model fitted to the CD4 count data*

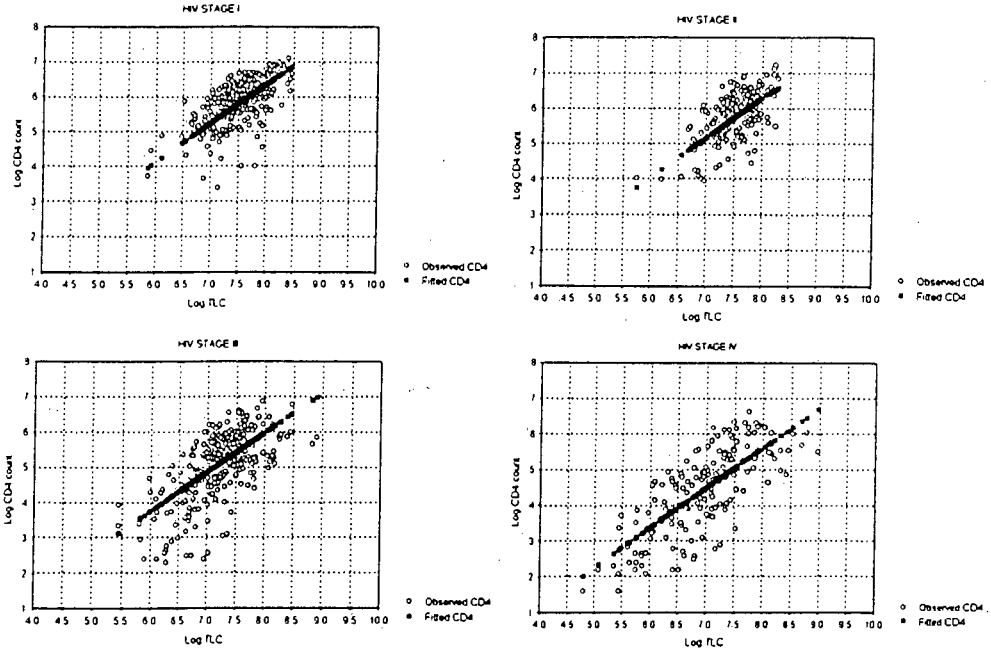
Table 2.8 Parameter estimates from the final model fitted to the CD4 count data

Fixed term	**Parameter estimate**	**Standard error**	**Z-value**
Constant	-2.583	0.310	-8.332
Log TLC	1.117	0.041	27.244
Stage I	0.000		
Stage II	-0.074	0.055	-1.345
Stage III	-0.369	0.051	-7.235
Stage IV	-0.756	0.068	-11.118
Random term	**Variance Component**	**Standard error**	
Patient	0.239	0.026	
Random error	0.176	0.010	
Variance ratio	1.358		
Within patient correlation	0.576		
Stratum variance	**Variance**	**Degrees of freedom**	
Patient	0.605	361.92	
Random error	0.176	570.08	

Table 2.9 Correlation coefficients from final fitted model by HIV Stage

HIV Stage	Correlation coefficient	Coefficient of determination (%)
I	0.576	33.1
II	0.618	38.2
III	0.652	42.5
IV	0.750	56.3
All Stages	0.781	61.0

Figure 2.10. Scatterplot of Log CD4 count and Fitted values against Log TLC by HIV Stage



2.4.3 Model Checking

A plot of residuals against fitted values from the model is shown in Figure 2.11.The plot shows no systematic pattern in the log CD4 count residuals. Figure 2.12 shows a plot of the residuals against HIV stage. The plots do not appear to indicate an increase in variance by HIV stage. However, observations of patients 210 in stage II and of patient 18 and 135, both in stage III, appeared to be outliers.

Although the residuals were quite small in magnitude, the normal probability plots indicated substantial departure from normality, for HIV stage I, II and III, log CD4 counts (Figure 2.13). However, the normal plots of the residuals do not appear to contradict the assumption of normality for CD4 counts taken on stage IV patients. We found it informative to inspect the tails of the normal probability plot since it indicated outlying observations. It was also observed that observations corresponding to the extremes of the normal probability plots belonged to different patients, that is, outlying observations did not always come from the same patient(s).

Figure 2.14 shows an index plot of modified Cook's statistics by HIV stage. There was no evidence of influential observations for CD4 counts belonging to patients in HIV stage II. In stage III, observations of patients 18 and 135 were found to the influential. Although the model appeared to be adequate for stage IV CD4 counts, according to the residual analysis, there were few influential observations to the parameters in the model. These observations belonged to patients, 140, 253 and 322. Exclusion of these influential observations had a negligible effect on the regression results (Table 2.10).

Table 2.10 Parameter estimates of the final model fitted for the CD4 count data with influential observations excluded

Fixed term	**Parameter estimate**	**Standard error**	**Z-value**
Constant	-2.411	0.293	-8.229
Log TLC	1.098	0.039	28.154
Stage I	0.000		
Stage II	-0.087	0.051	-1.706
Stage III	-0.375	0.048	-7.813
Stage IV	-0.747	0.064	-11.672
Random term	**Variance Component**	**Standard error**	
Patient	0.212	0.023	
Random error	0.157	0.009	
Variance ratio	1.350		
Within patient correlation	0.575		
Stratum variance	**Variance**	**Degrees of freedom**	
Patient	0.537	357.02	
Random error	0.157	565.98	

Figure 2.11. Scatterplot of Log CD4 residuals against fitted values by HIV Stage

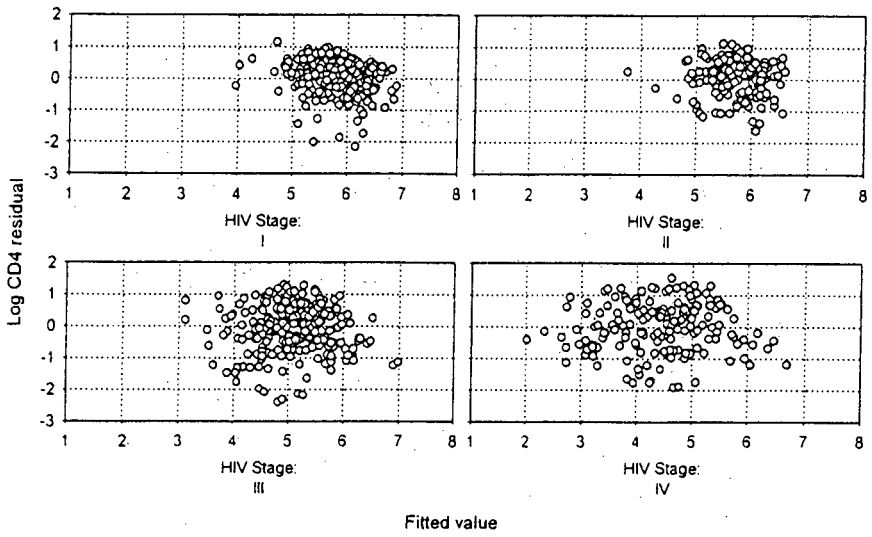


Figure 2.12. Plot of Log CD4 residuals against HIV Stage

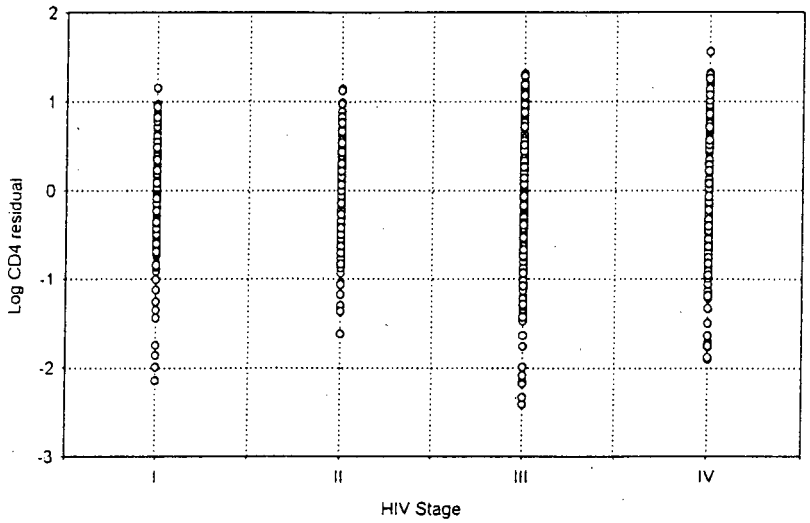


Figure 2.13. Normal probability plots of Log CD4 residuals by HIV Stage

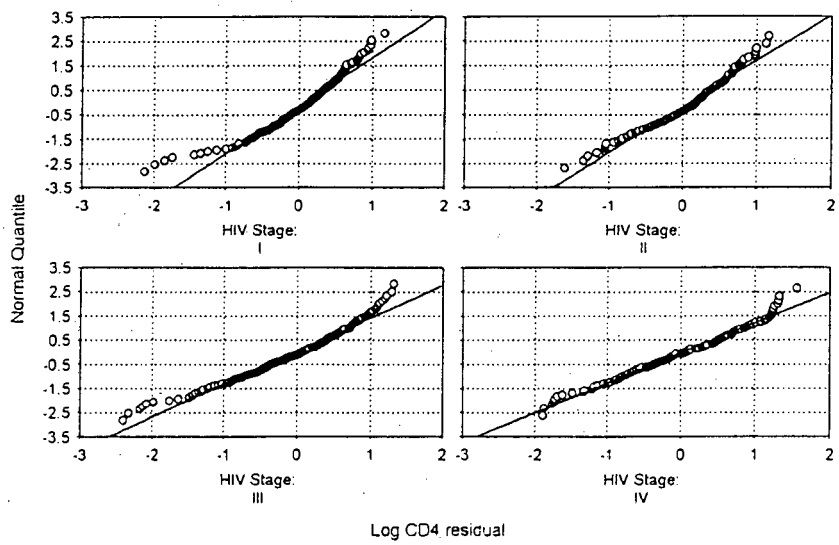
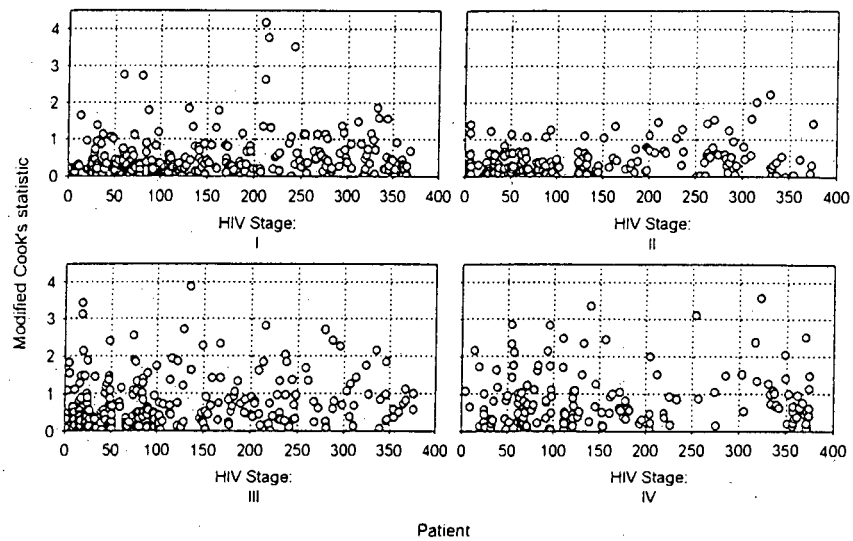


Figure 2.14. Index plot of Modified Cook's statistics by HIV Stage



2.4.4 Model Application

2.4.4.1 Estimating the CD4 count on 'new' data using the model

One of the objectives of the research was to investigate how well TLC estimated CD4 count in a regression model. Given a patient's total lymphocyte count and HIV stage, from the fitted model the estimation equations were given by:

Let y_{ij} be the CD4 cell count at visit j for patient i

x_{ij} be the total lymphocyte count at visit j for patient i

Stage I: $\log y_{ij} = -2.583 + 1.117 \times \log x_{ij}$

Stage II: $\log y_{ij} = -2.658 + 1.117 \times \log x_{ij}$

Stage III: $\log y_{ij} = -2.953 + 1.117 \times \log x_{ij}$

Stage IV: $\log y_{ij} = -3.339 + 1.117 \times \log x_{ij}$

Since the fitted CD4 values were on the log-scale, to make them comparable to the observed CD4 counts they were back-transformed by taking exponents resulting in the following set of equations:

Stage I: $y_{ij} = exp(-2.583 + 1.117 \times \log x_{ij})$

Stage II: $y_{ij} = exp(-2.658 + 1.117 \times \log x_{ij})$

Stage III: $y_{ij} = exp(-2.953 + 1.117 \times \log x_{ij})$

Stage IV: $y_{ij} = exp(-3.339 + 1.117 \times \log x_{ij})$

To assess the practical validity of the fitted model, the rest of the data (records not used in fitting the model), was used to estimate CD4 counts using the above equations. Table 2.11. gives a distribution of the relative distances of the estimated values from the observed CD4 counts. The model was more satisfactory in estimating CD4 counts taken on patients in earlier stages of the HIV disease than those in advanced stages of the disease. For instance, 80 percent of the predicted CD4 counts were within 10 percent of the observed value of CD4 count in stage I compared to only 46 percent in stage IV. The reason for this finding might be that the CD4 counts were more variable at advanced stages of the HIV disease than at earlier stages. Thus, the new data set might present with extreme observations, which the model could not fit, yet they were genuine observations.

Table 2.11 Distribution of relative distances of estimated CD4 counts to observed CD4 counts

Within k % of observed CD4	Stage I		Stage II		Stage III		Stage IV		All Stages	
	Freq.	Percent	Freq.	Percent	Freq.	Percent	Freq.	Percent	Freq.	Percent
< 10	438	79.9	332	76.1	397	56.2	145	40.1	1312	63.9
10-20	88	16.1	73	16.7	192	27.2	100	27.6	453	22.1
20-30	10	1.8	12	2.8	44	6.2	47	13.0	113	5.5
> 30	12	2.2	19	4.4	74	10.5	70	19.3	175	8.5
Total	548	100.0	436	100.0	707	100.0	362	100.0	2053	100.0

2.4.4.2 Evaluating the model for use in clinical practice

In monitoring HIV patients, thresholds of the CD4 count are often used to indicate the onset of particular diseases, such as AIDS, or to commence treatment of particular infections. A CD4 count < 200 is generally used as a criterion for commencing treatment against toxoplasmosis and pneumocystis carinii pneumonia. HIV patients with CD4 counts < 50 are prone to more serious diseases or infections associated with the HIV disease.

Post *et al.*, (1996) suggest that a TLC < 1250 could be used, instead of CD4 < 200, as a criterion for commencing cotrimoxazole prophylaxis (treatment against toxoplasmosis and pneumocystis carinii pneumonia). We found a TLC < 1250 to be 64 percent sensitive and 88 specific for a CD4 count < 200, using the raw data and ignoring HIV stage information. These estimates were consistent with those obtained by Post *et al.*, (1996).

In this study, we determined different TLC cut-offs for the different HIV stages, using the model. The TLC cut-offs were calculated using the model equations for each stage. We employed inverse regression methodology (Draper and Smith, 1981, page 47) to calculate TLC values given a CD4 count of 200 in each stage. The equations for estimating the TLC cut-offs for a CD4 count of 200 were given by (on the log-scale)

Stage I: $$\log \text{TLC} = \frac{\log 200 + 2.583}{1.117} = 7.056$$

Stage II: $$\log \text{TLC} = \frac{\log 200 + 2.658}{1.117} = 7.123$$

Stage III: $$\log \text{TLC} = \frac{\log 200 + 2.953}{1.117} = 7.387$$

Stage IV: $$\log \text{TLC} = \frac{\log 200 + 3.339}{1.117} = 7.733$$

or equivalently (on the original scale of the data)

Stage I: $$TLC = exp\left[\frac{\log 200 + 2.583}{1.117}\right] = 1159.797$$

Stage II: $$TLC = exp\left[\frac{\log 200 + 2.658}{1.117}\right] = 1240.165$$

Stage III: $$TLC = exp\left[\frac{\log 200 + 2.953}{1.117}\right] = 1614.854$$

Stage IV: $$TLC = exp\left[\frac{\log 200 + 3.339}{1.117}\right] = 2282.439$$

The estimates of TLC values given a CD4 count of 200 pointed toward different TLC cut-offs for each stage (Table 2.12). According to the model, this implied that different values of TLC for each stage of the HIV disease should used as thresholds for commencing treatment. This finding was probably due the presence of higher TLC values for some patients, at advanced stages of the HIV disease (stage III and IV), (Figure 2.9) . It was noted that the TLC cut-off of 1250 corresponded to an estimated TLC cut-off for HIV stage II of 1240.

We then investigated how good the TLC cut-offs were at detecting a CD4 count of < 200 or > 200. This was done by classifying the CD4 counts from the rest of the data, using the TLC cut-offs. The rest of the data consisted of 2053 observations from 957 patients (data not used in fitting the model). Proportions of CD4 counts correctly classified and those incorrectly classified by the TLC cut-offs were calculated. The proportions were defined as follows (Altman, 1994, page 410):

Sensitivity was the proportion of CD4 counts < 200 correctly identified by the TLC cut-off.

Specificity was the proportion of CD4 counts > 200 correctly identified by the TLC cut-off.

Positive predictive value (PPV) was the proportion of TLC values less than the TLC cut-off and had CD4 counts < 200.

Negative predictive value (NPV) was the proportion of TLC values greater than the TLC cut-off and had CD4 counts > 200

False negative rate was to the proportion of CD4 counts < 200 and had TLC values > TLC cut-off.

False positive rate was the proportion of CD4 counts > 200 and had TLC values < TLC cut-off.

The results of the calculations of the above proportions are given in Table 2.12. In HIV stage IV, a TLC < 2281 was found to be 96 percent sensitive and 35 percent specific for an observed CD4 count < 200 (Table 2.12), with a positive predictive value of 85 percent. This implies that according to the model, from which the TLC cut-offs were calculated, we would expect 96 percent of the observed CD4 counts < 200 to have TLC values < 2281, while 35 percent of those with observed CD4 counts > 200 would have TLC values > 2281.

The false negative rate was estimated to be 4.18 percent in HIV stage IV. In this situation, this proportion represents the proportion of patients who would not be given treatment, if a TLC cut-off of 2281 were used as a criterion for commencing prophylasis treatment on patients in HIV stage IV. On the other hand the false negative rate represents the proportion of patients who would be given the treatment, yet they probably do not need it since their 'true' CD4 counts would be above 200. This proportion gives an indication of the costs (not necessarily financial) associated with unnecessary treatment of patients. If a TLC cut-off of 2281, in HIV stage IV, were used to decide on which patients to treat; we would be treating about 96 percent of the patients who need treatment but we could also be giving it to 65 percent of the patients who do not need by virtue of their CD4 counts being above 200.

For comparative purposes, we also assessed the ability of a fixed value of TLC < 1250 to detect a CD4 < 200, as suggested by Post *et al.*, (1996). In HIV stage IV, a TLC < 1250 was found to be 72 percent sensitive and 83 percent specific for an observed CD4 count < 200 (Table 2.13), with a positive predictive value of 94 percent. If TLC cut-off of 1250 were used as a criteria for commencing treatment, 72 percent of the patients needing the treatment would receive it and only 17 percent of patients with CD4 counts > 200 would receive it.

We also determined TLC cut-offs for a CD4 count of 50 to detect HIV patients with acute infections related to the HIV disease. The model again suggested different cut-offs for the TLC for each stage (Table 2.14).

The tables showing the relations between the different TLC values (cut-offs) and the CD4 count < 200 or > 200 are given in Appendix A.

Table 2.12 Sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) of estimated TLC values for detecting CD4 counts < 200 or >200 in each HIV stage*

Stage	TLC value	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	False negative rate (%)	False positive rate (%)
I	1159	42.03 *30.24,54.52*	92.07 *89.65,94.49*	43.28 *31.22,55.96*	91.68 *89.21,94.15*	57.97 *45.48,69.76*	7.93 *5.51,10.35*
II	1240	48.23 *37.26,59.34*	88.89 *85.60,92.18*	50.62 *39.27,61.92*	87.61 *84.18,91.04*	51.76 *40.66,62.74*	11.40 *8.08,14.72*
III	1615	78.75 *74.48,83.02*	68.36 *63.52,73.20*	71.28 *66.79,75.77*	76.34 *71.66,81.01*	21.25 *16.98,25.52*	31.64 *26.80,36.48*
IV	2281	95.82 *93.50,98.14*	34.67 *24.04,46.54*	84.88 *80.98,88.78*	68.42 *53.64,83.20*	4.18 *1.86,6.50*	65.33 *54.56,76.10*

* 95 % confidence intervals are shown in *italics*.

Table 2.13 Sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) of a fixed TLC value of 1250 for detecting CD4 counts < 200 or >200 in each HIV stage*

Stage	TLC value	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	False negative rate (%)	False positive rate (%)
I	1250	50.72 *38.41,62.98*	89.77 *87.06,92.48*	41.67 *31.00,52.94*	92.67 *90.30,95.04*	49.28 *37.02,61.59*	10.23 *7.52,12.94*
II	1250	50.59 *39.52,61.61*	88.03 *84.63,91.43*	50.59 *39.52,61.61*	88.03 *84.63,91.43*	49.41 *38.39,60.48*	11.97 *8.57,15.37*
III	1250	63.17 *58.14,68.20*	85.88 *82.25,89.51*	81.68 *77.09,86.27*	70.05 *65.74,74.36*	36.83 *31.80,41.86*	14.12 *10.49,17.75*
IV	1250	72.47 *67.30,77.64*	82.67 *72.19,90.43*	94.12 *91.02,97.22*	43.97 *35.78,52.16*	27.53 *22.36,32.70*	17.33 *8.76,25.90*

* 95 % confidence intervals are shown in *italics*.

Table 2.14 Sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) of estimated TLC values for detecting CD4 counts < 50 or > 50 in HIV stage III and IV*

Stage	TLC value	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	False negative rate (%)	False positive rate (%)
III	467	29.81 *21.02,38.60*	98.18 *97.11,99.25*	73.81 *57.96,86.14*	89.02 *86.64,91.40*	70.19 *61.40,78.98*	1.82 *0.08,2.89*
IV	659	63.27 *55.48,71.06*	87.91 *83.55,92.27*	78.15 *70.73,85.57*	77.78 *72.55,83.01*	36.73 *28.94,44.52*	12.09 *7.73,16.45*

* 95 % confidence intervals are shown in *italics*.

We assume the proportion, p, is normally distributed , and then its 95 % confidence interval is given by:

$$p \pm 1.96\sqrt{\frac{pq}{n}}$$

where:

p = Sensitivity or Specificity or PPV or NPV or False negative or False positive rate

$q = 1 - p$

n = sample size

It was further assumed that the TLC determinations were independent of each other, that is, the fact that the determinations were repeated measures on the same patient was ignored.

CHAPTER 3

The relationship between the CD4 count and Viral load in HIV infected individuals

3.1 Introduction

Until recently, the CD4 count has been used as an important marker of disease progression in HIV-infected individuals. However, despite being a good prognostic marker in advanced HIV disease, it is of little value in early asymptomatic stages (Mellors, 1998 and Mellors *et al.*, 1996,). Hoover *at al.*, (1995) has asserted that some individuals develop AIDS at higher CD4 count cells than might be expected, while some individuals remain free of AIDS despite having low CD4 count. These two studies indicate that the CD4 cell count is not a perfect marker of HIV disease progression.

The ability to quantify viral load (the amount of virus in blood) has revolutionized the monitoring of HIV disease progression in patients. HIV-infected individuals experience a continual loss of CD4 cells throughout infection with a lower CD4 count indicating more severe immune deficiency and a higher risk of developing AIDS. On the other hand, HIV-infected individuals experience a gradual rise in viral load throughout the HIV infection period with higher values at much more advanced stages of the disease (Pantaleo *et al.*, 1995). Wei *at al.*, (1995) contends that viraemia in HIV infection is sustained by rapid, high-level viral replication, requiring continuous reinfection and destruction of CD4 counts.

The relationship between the CD4 cell count and viral load has been considered by several authors (Mellors, 1998; Mellors *et al.*, 1996; Saksela *et al.*, 1995; Soriano *et al.*, 1998). Although higher levels of viral load have been associated with a fall in CD4 cells, the relationship between these variables is still not well known. Because general use of viral load in clinical practice has only been recent, there are a few longitudinal studies that have considered the relationship between these two markers.

This study examined the relationship between the CD4 count and viral load in a repeated measures setting for an African population. The relationship between these two variables in an African population has not been studied in the literature.

3.2 Objectives

The goals in this analysis were to:

i) investigate the relationship between the CD4 count and viral load in a single sample from an African population.

ii) examine the relationship between the CD4 count and viral load when both markers consist of repeated measurements.

iii) estimate the variation in the CD4 count that is due to the patient (biological variation) and that, which is due to random error.

3.3 Methodology

The data used for this analysis was for HIV patients enrolled in two clinical studies at Somerset Hospital HIV Clinic, in Cape Town. The two data sets from these studies were used, seperately, to investigate the relationship between the CD4 count and viral load. The first data set consisted of single measurements of CD4 counts and viral load on 123 patients. The second data set consisted of repeated measurements of CD4 counts and viral load on 116 patients enrolled in a clinical trial.

For the single measurements of CD4 counts and viral load measurements, we used simple linear regression to perform the analysis while for the repeated measurements of the these two markers we employed the random effects model discussed earlier in Section 2.3 in Chapter 2 of this study. Both analyses were done using the GENSTAT package.

3.4 Results

3.4.1 The association between single measurements of CD4 and Viral load

3.4.1.1 Data description

Initial examination of the relationship between CD4 cell counts and viral load was conducted using single measurements of CD4 count and viral load of 123 HIV patients. Patients were not taking antiretroviral drugs, nor had received immunizations nor were suffering any opportunistic or non-opportunistic infections at the time of the study. Selected patients had CD4 counts above 200. For some patients either CD4 cell count or the viral load or both were not measured, we excluded these patients (62) from the analysis. The analysis was then based on data for 61 patients.

Table 3.1 and Table 3.2 show the summary statistics of the CD4 counts and viral load; both their untransformed measurements and their natural logarithmic transformations. Viral load measurements were found to be more variable than the CD4 cell counts, with a coefficient of variation of 1.11 compared to 0.30 for the CD4 counts.

Figure 3.1 and Figure 3.2 show the histograms of the untransformed CD4 counts and log CD4 counts, respectively. The distribution of the untransformed CD4 counts was somewhat less skewed than that of the log-transformed CD4 counts. Figure 3.3 and Figure 3.4 show the histograms of the untransformed viral load and log-transformed viral load, respectively. The distribution of the log viral load appeared skewed than that of the untransformed viral load. The distribution of the log CD4 counts and log viral load provides a simple description of the patients within the different stages of the HIV disease. For example patients in the right tail of the distribution log CD4 counts represent healthier patients and those in the left tail of the distribution represent sicker patients and are thus likely to be at advanced stages of the HIV disease. The data we used in this analysis did not contain information on HIV stage of the patients.

Table 3.1 Summary Statistics of CD4 cell counts for 61 HIV patients

Variable	N	Mean	Min.	Max.	Med.	Std. dev.	Coeff. of var.
CD4	61	356.0	202.0	634.0	356.0	106.40	0.30
Log CD4	61	5.8	5.3	6.5	5.8	0.30	0.05

Table 3.2 Summary Statistics of Viral load for 61 HIV patients

Variable	N	Mean	Min.	Max.	Med.	Std. Dev.	Coeff. of var.
Viral load (000's)	61	103.3	0.98	480.0	62.7	114.70	1.11
Log Viral load	61	10.8	6.90	13.1	11.1	1.50	0.14

Figure 3.1 Histogram of CD4 count

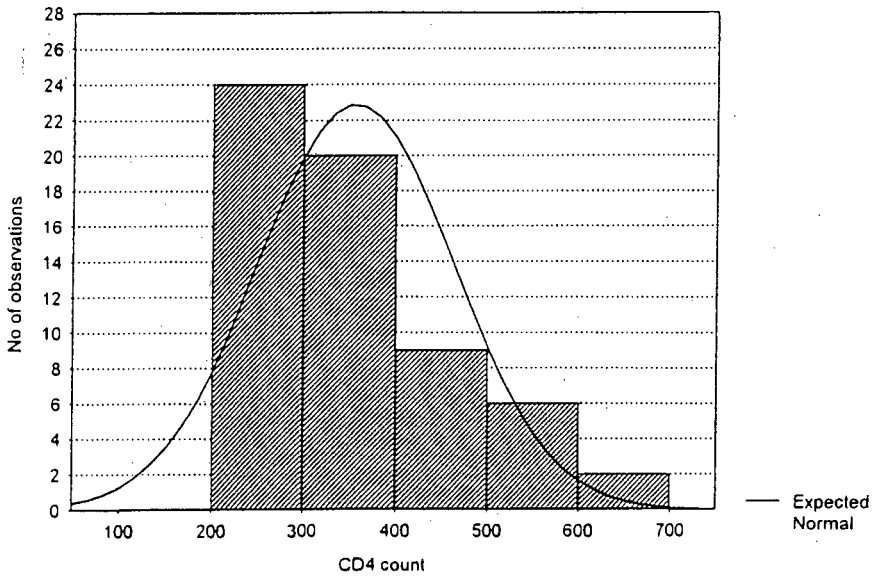


Figure 3.2. Histogram of Log CD4 count

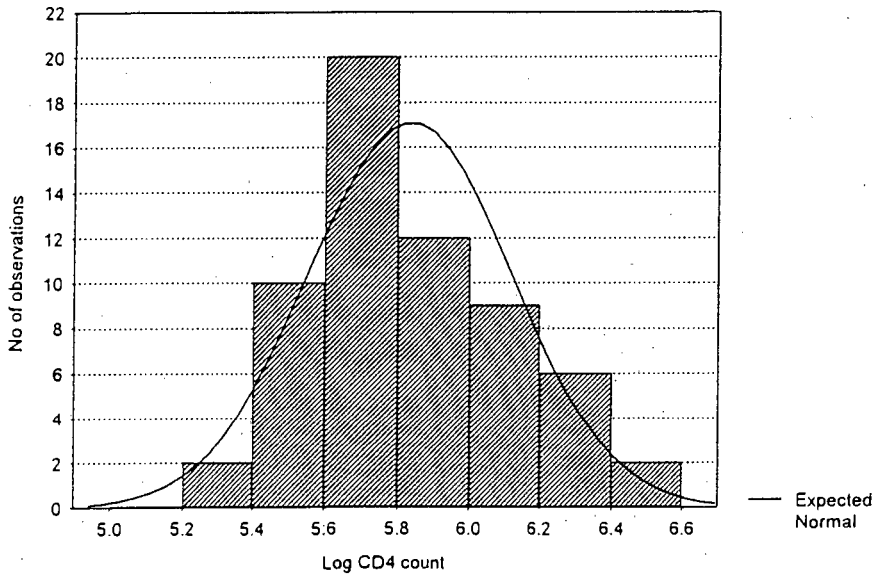


Figure 3.3. Histogram of Viral load (000's)

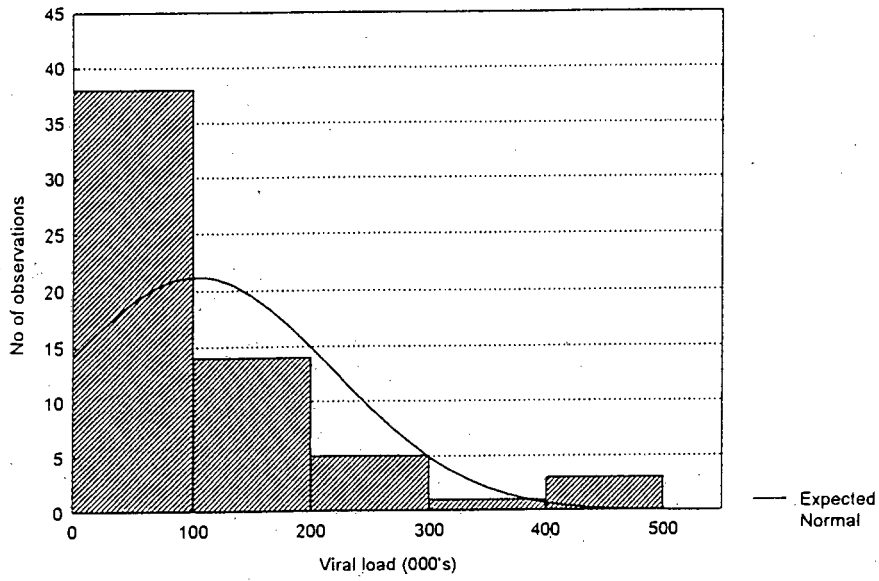


Figure 3.4. Histogram of Log Viral load

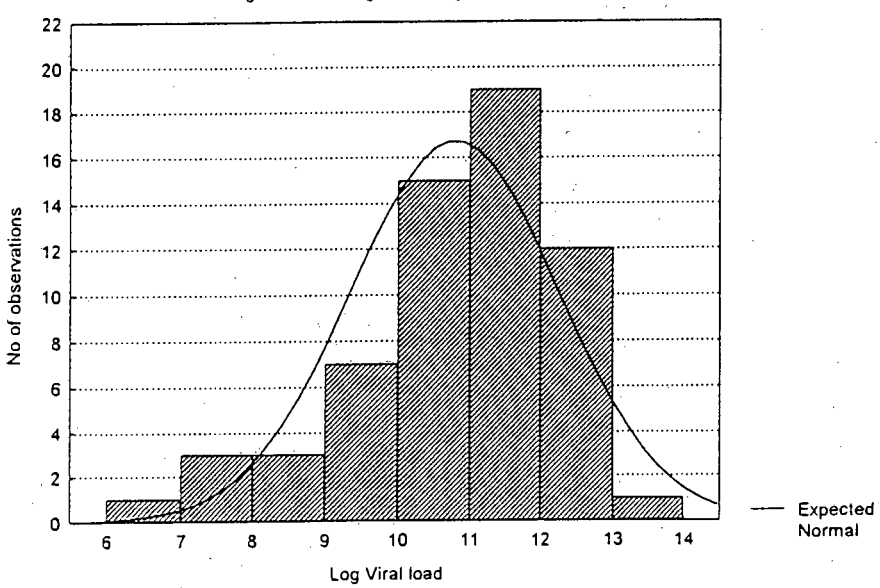


Figure 3.5. Scatterplot of CD4 count against Viral load (000's)

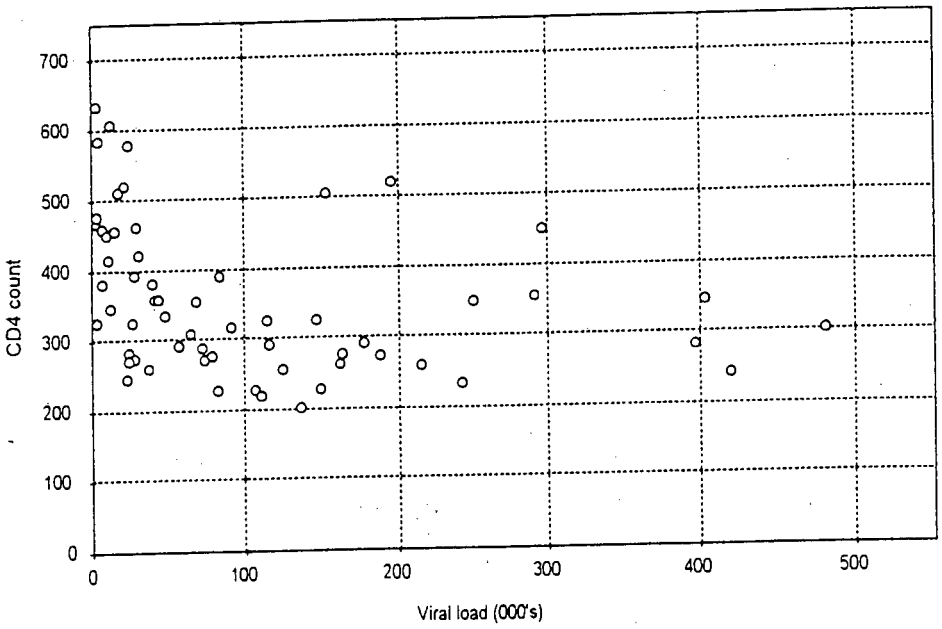
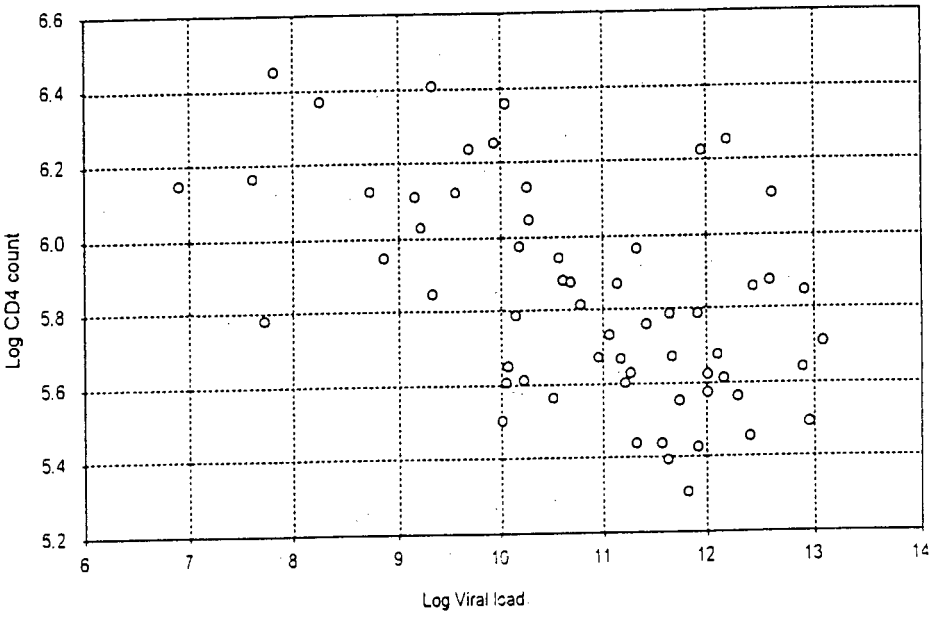


Figure 3.6. Scatterplot of Log CD4 count against Log Viral load



3.4.1.2 Model Specification

Since the range of the viral load values were large, so we fitted a linear regression model to log-transformed CD4 counts with log-transformed viral load as the only explanatory variable.

The linear regression model was parameterised as follows:

Let y_i be the CD4 cell count for patient i

x_i be the viral load measurement for patient i

Then the linear model for the i^{th} patient is given by:

$$\log y_i = \mu + \beta \log x_i + e_i \quad i = 1,...,61$$

where μ is the constant term, β is the regression coefficient for the natural logarithm of viral load and e_i is the random error.

The model makes three basic assumptions:

i) e_i is a random variable with mean zero and unknown variance σ_e^2 i.e. $E(e_i) = 0$ and $\text{Var}(e_i) = \sigma_e^2$

ii) e_i and e_j are uncorrelated , $i \neq j$, so that $\text{cov}(e_i, e_j) = 0$

iii) $e_i \sim N(0, \sigma_e^2)$

Then estimates of the regression parameters are obtained by ordinary least squares.

3.4.1.3 Model Selection

The regression results are given below (Table 3.3 and Table 3.4). Figure 3.7 plots the observed CD4 counts and fitted values on the log-scale against the log-transformed viral load.

The correlation between the log-transformed CD4 counts and log-transformed viral load was found to be -0.54 and was statistically significant. This implies that viral load explained 29 percent of the variability in CD4 counts. This estimate of the correlation coefficient was consistent with results from previous studies of the relationship between CD4 counts and viral load. Soriano *et al.*, (1998) reported a correlation coefficient of -0.61 between the CD4 count and viral load. Mellors *et al.*, (1996) found a weak association between the CD4 count and viral load with a Spearman's rank correlation of -0.27.

The negative regression coefficient estimate for log-transformed viral load confirmed that there was a negative relationship between CD4 and viral load i.e. a decrease in the log (CD4) measurement was associated with an increase in the log-transformed viral load. The regression coefficient was found to be statistically significant (Table 3.4). According to this model, the relationship between the variables could be summarized as (on the log-scale)

$$\log y_i = 6.979 - 0.106 \times \log x_i ,$$

or equivalently (on the original scale of the data) as

$$y_i = exp(6.979 - 0.106 x_i)$$

Table 3.3 Analysis of variance table for the CD4 count data from the final model

Source	d.f.*	Sum of squares	Mean square	Variance ratio
Log Viral load	1	1.421	1.421	24.490
Residual	59	3.424	0.058	
Total	60	4.845	0.081	

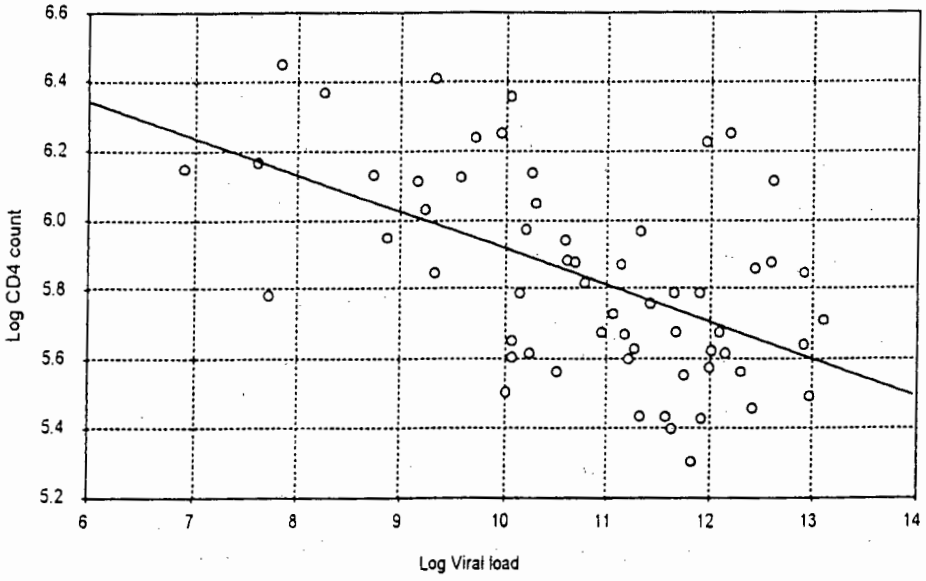
* d.f. refers to degrees of freedom

Table 3.4 Estimates of regression parameters and their standard errors of the final model for the CD4 count data

Variable	Parameter Estimate	Standard error	t-statistic	d.f.*
Constant	6.979	0.233	29.90	59
Log Viral load	-0.106	0.021	-4.95	59

* d.f. refers to degrees of freedom

Figure 3.7. Scatterplot of Log CD4 against Log Viral load and the fitted regression line



3.4.1.4 Model Checking

Figure 3.8 shows a composite plot of the residuals from the fitted model. The observations of patient 3 was found to have a relatively large log CD4 residual of 2.39 (Figure 3.9), with a corresponding CD4 count of 520 cells and a viral load of 194359.

An index plot of the leverage values from the model is given in Figure 3.10. The plot shows that the viral load for patient number 3 was not at all extreme, relative to those of other patients, but they are for patient number 13, 36, 46 and 51. In other words these patients had large leverage values; their viral load measurements were less than 3000. Although these observations were not outliers, they had high leverage values and so they may have been influential. On the other hand, although the observation of patient number 3 was an outlier it was not potentially influential.

Figure 3.11 shows the influence of the each observation on the parameters of the fitted model. Observations of patients 3, 41 and 51 were found to be influential. When these observations were excluded from the analysis, the regression coefficient estimates were changed but their standard errors were not changed (Table 3.6). The log viral load explained 40.8 percent of the variability in the log CD4; this implied a correlation coefficient instead of the –0.64. In contrast when the influential observations were included in the analysis, the log viral load only explained 29.2 percent of the variability in the log CD4 count.

The preceding analysis revealed that CD4 count was inversely related to viral load and that viral load could only explain 41% of the variability in CD4 count. This finding indicated that the relationship was not strong enough to allow for satisfactory prediction of CD4 from viral load. However, the linear regression model fitted provided a simple description of the relationship between CD4 count and viral load.

We noted that our results were more in agreement with findings by Soriano *et al.*, (1998). They found that in a simple regression model, viral load explained 37% of the variability in the CD4 count. In contrast Mellors *et al.*, (1996) found a weak association between viral load and CD4 count with a Spearman's correlation coefficient (r) of –0.27. Different methodology for calculating the correlation coefficient employed in our analysis and that used by Mellors *et al.*, (1996) may have resulted in different estimates of the correlation coefficient. Furthermore, the range of CD4 counts involved in our analysis was much lower, being 202-634 compared to 400-800 in Mellors's *et al.*, (1996) study. It must be noted that our analysis did not take into account the censoring caused by the selection criterion of CD4 counts above 200.

Table 3.5 Analysis of variance table for the CD4 count data from the final model with influential observations excluded

Source	d.f.*	Sum of squares	Mean square	Variance ratio
Log Viral load	1	1.917	1.917	40.32
Residual	56	2.663	0.048	
Total	57	4.580	0.080	

Table 3.6 Estimates of regression parameters and their standard errors from the final model with influential observations excluded

Variable	Parameter Estimate	Standard error	t-statistic	d.f.*
Constant	7.236	0.224	32.24	56
Log Viral load	-0.131	0.021	-6.35	56

* d.f. refers to degrees of freedom

Figure 3.8. Composite Plot of Log CD4 residuals

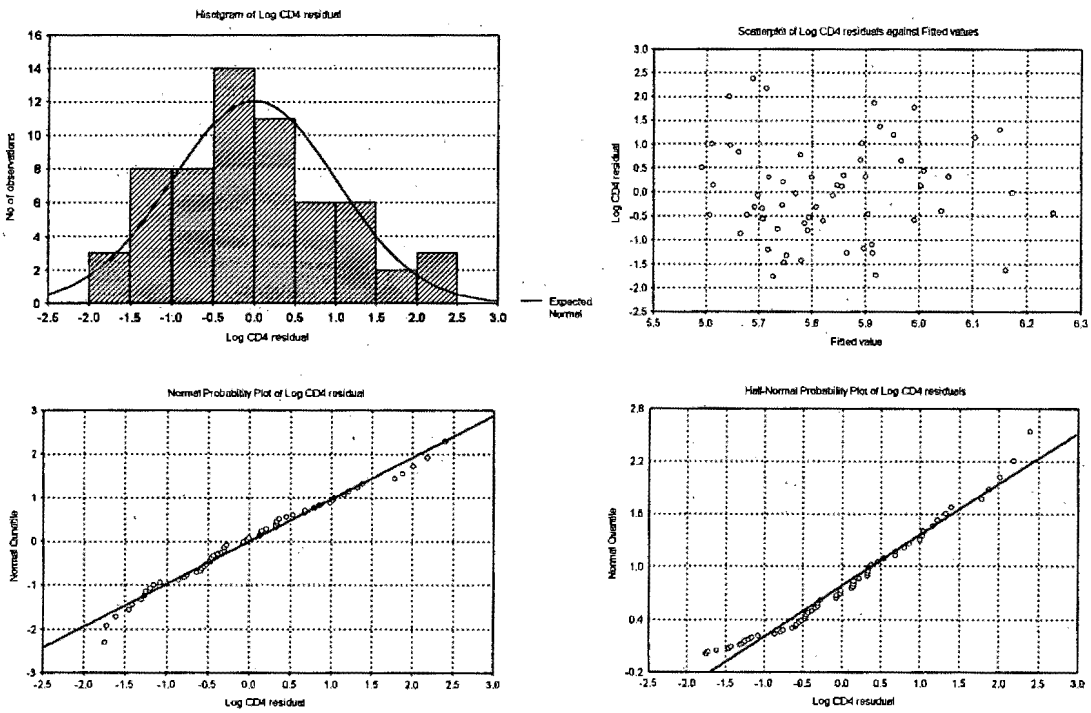


Figure 3.9. Index Plot of Log CD4 residuals

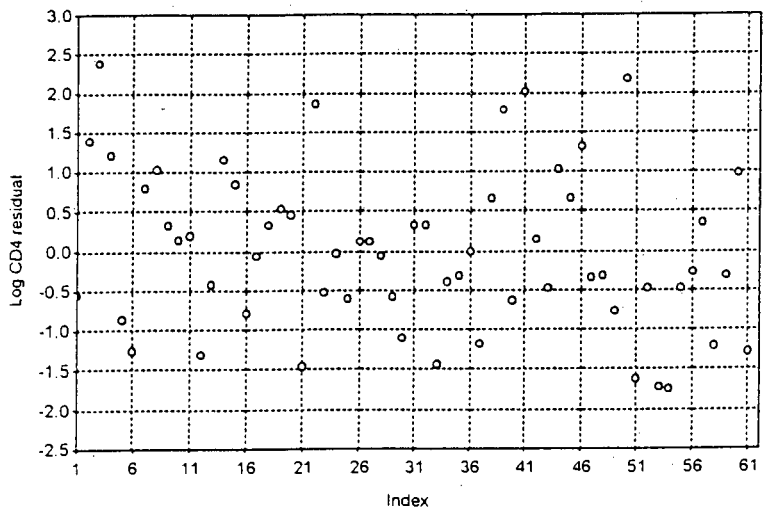


Figure 3.10. Index plot of leverage values

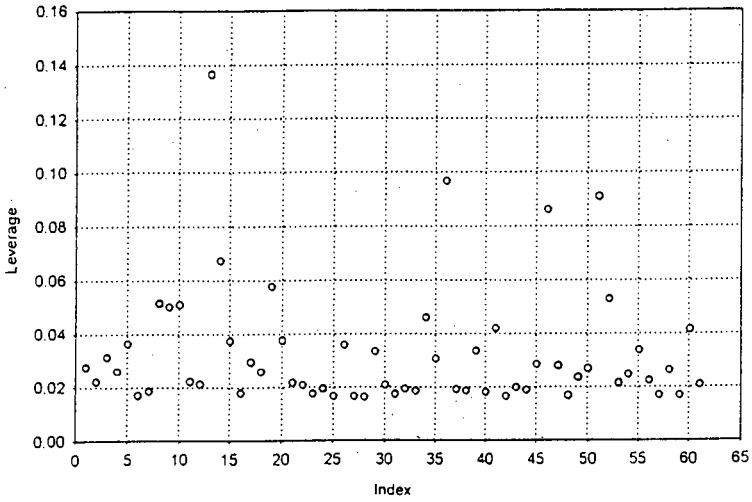
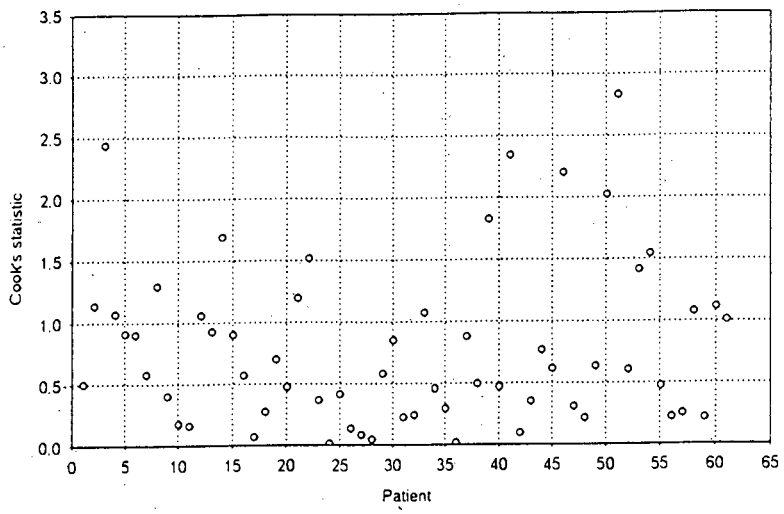


Figure 3.11. Index plot of Cook's statistics



3.4.2 The relationship between repeated measurements of CD4 and Viral load

3.4.2.1 Description of the data

It is well known that the CD4 cell count of a HIV-infected individual, at a particular time, is subject to biological variation or random error, therefore it (CD4 count) is not directly measurable (Hughes *et al* .; 1994; Taylor *et al.*, 1989 and 1994). This means that large changes in a patient's CD4 cell count may not reflect significant changes in the underlying immune function. It is therefore necessary to consider variability in the CD4 count when monitoring HIV infection. This analysis investigated aspects of variability in the CD4 count and examined the association between repeatedly measured CD4 counts and viral load.

The data set used for the analysis consisted of CD4 counts, viral load measurements on patients entered into the clinical trial. Eligible patients were not on antiretroviral therapy prior to entry to the trial and could be at any stage of the HIV disease. It was a requirement that selected patients should be independent to be included in the clinical trial, that is they should be managing their lives. In addition selected patients did not have active infections or HIV-related diseases at the time of entry into the trial. Lastly, patients had to be sixteen years or older to be included in the clinical trial. Treatment commenced on all selected patients at the 12^{th} week of the clinical trial.

The CD4 count and or viral load were not measured at some visits because patients missed their scheduled clinic visits. Reasons for missing clinic visits were mainly of a social nature and were not related to illness or disease. In addition the instrument for measuring the viral load did not measure under 400 HIV-RNA copies and thus made a further contribution to the missing viral load values in the dataset. We considered only patients with complete records (56 patients); this criterion of selection of patients resulted in a balanced structure of the data, which we then used in the analysis. The data set then consisted of measurements taken at five equally spaced intervals (an interval being 12 weeks) for 56 HIV patients.

Table 3.7 and Table 3.9 show the summary statistics of the CD4 counts and viral load, respectively, by week of measurement. CD4 counts ranged from 10 to 993 cells, while viral load ranged from 310 to 9740000. Summary statistics of the log CD4 counts and log viral load, by week of measurement, are given in Table 3.8 and Table 3.10, respectively. Figure 3.12 and Figure 3.13 show the histograms of log CD4 count and log viral load, respectively, by week. The distribution of the log-transformed viral load appeared less skewed than that of the log-transformed CD4 counts.

Figure 3.14 and Figure 3.15 show the Box –Whisker plots of log CD4 count and log-viral load, respectively. The CD4 counts appeared to be stable throughout the clinical trial but viral load seemed to increase steadily over time after the first twelve weeks. However, not all patients followed this general pattern. Figure 3.16 and 3.17 present profiles of log-transformed CD4 counts and viral load for 5 selected patients, respectively. The pattern of both CD4 and viral load over the study period seemed

very erratic. It must noted that our analysis did not take into account the censoring caused by the quantification limit of 400 HIV-RNA copies (viral load).

Table 3.7 Summary Statistics of CD4 count by week

Week	No. of patients	No. of obs.	Mean	Min.	Max.	Std. Dev.	Med.	Coeff. of var.
12	56	224	274.1	40	711	145.9	251	0.53
24	56	224	252.1	12	537	136.1	253	0.54
36	56	224	235.3	19	532	132.8	232	0.56
48	56	224	225.3	12	993	164.2	208	0.73
Total	56	224	235.8	10	993	147.4	228	0.63

Table 3.8 Summary Statistics of log CD4 count by week

Week	No. of patients	No of obs.	Mean	Min.	Max.	Std. Dev.	Med.	Coeff. of var.
12	56	224	5.45	3.69	6.57	0.62	5.52	0.11
24	56	224	5.30	2.48	6.29	0.81	5.53	0.15
36	56	224	5.24	2.94	6.28	0.78	5.45	0.15
48	56	224	5.11	2.48	6.90	0.90	5.34	0.18
Total	56	224	5.28	2.48	6.90	0.79	5.47	0.15

Table 3.9 Summary Statistics of Viral load (000's) by week

Week	No. of patients	No of obs.	Mean	Min.	Max.	Std. Dev.	Med.	Coeff. Of var.
12	56	224	149.9	0.31	1800	378.0	13	2.52
24	56	224	228.1	0.99	4160	75.2	20	0.33
36	56	224	401.3	1.53	9740	1415.2	28	3.53
48	56	224	259.1	0.57	4230	783.8	33	3.03
Total	56	224	368.7	0.31	9740	216.7	42	0.59

Table 3.10 Summary Statistics of log Viral load by week

Week	No. of patients	No of obs.	Mean	Min.	Max.	Std. Dev.	Med.	Coeff. Of var.
12	56	224	9.68	5.74	14.40	2.20	9.45	0.23
24	56	224	10.17	6.90	15.24	2.02	9.92	0.20
36	56	224	10.62	7.33	16.09	1.94	10.25	0.18
48	56	224	10.38	6.35	15.25	2.03	10.39	0.20
Total	56	224	10.21	5.74	16.09	2.06	10.13	0.20

Figure 3.12. Histogram of Log CD4 count by Week

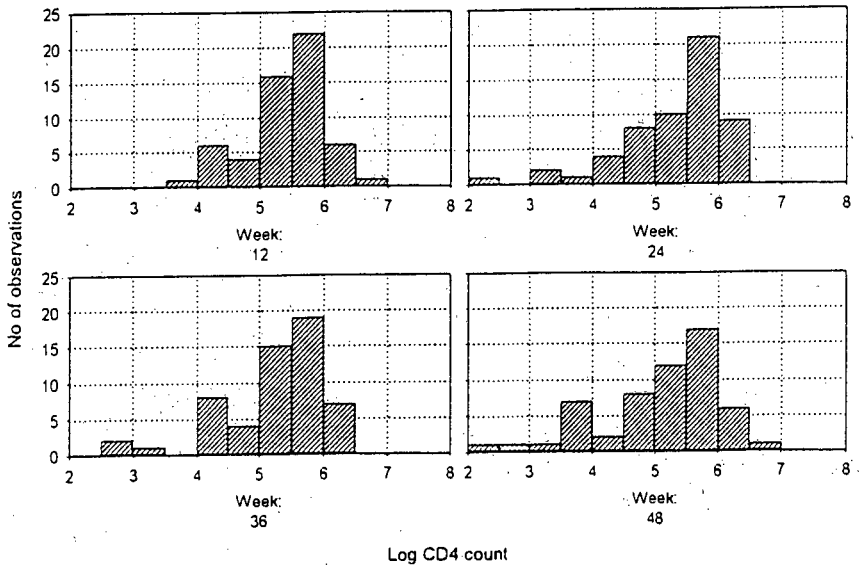


Figure 3.13. Histogram of Log Viral load by Week

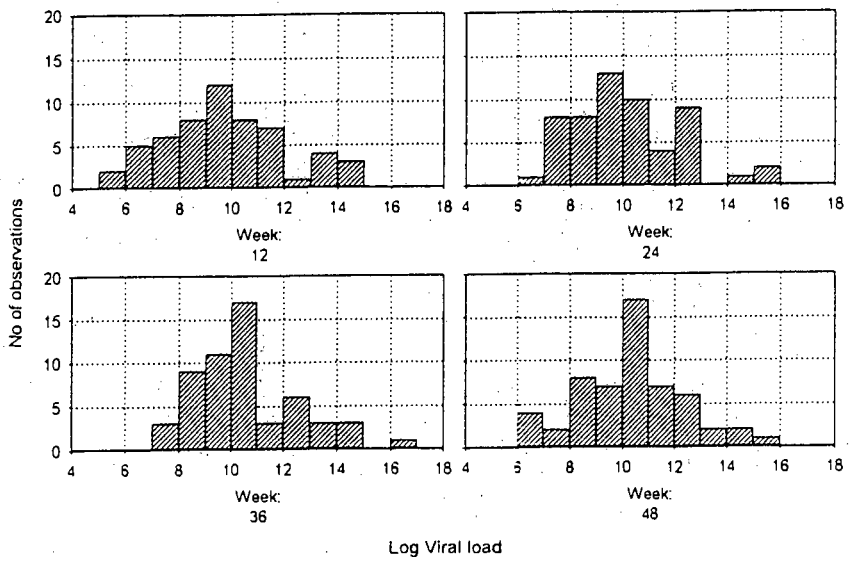


Figure 3.14. Box & Whisker Plot of Log CD4 count by Week

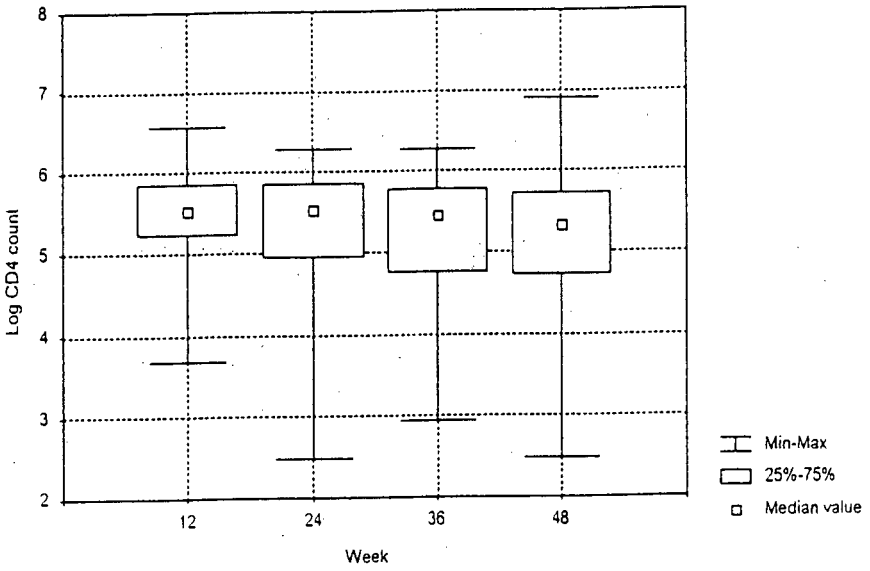


Figure 3.15. Box & Whisker Plot of Log Viral load by Week

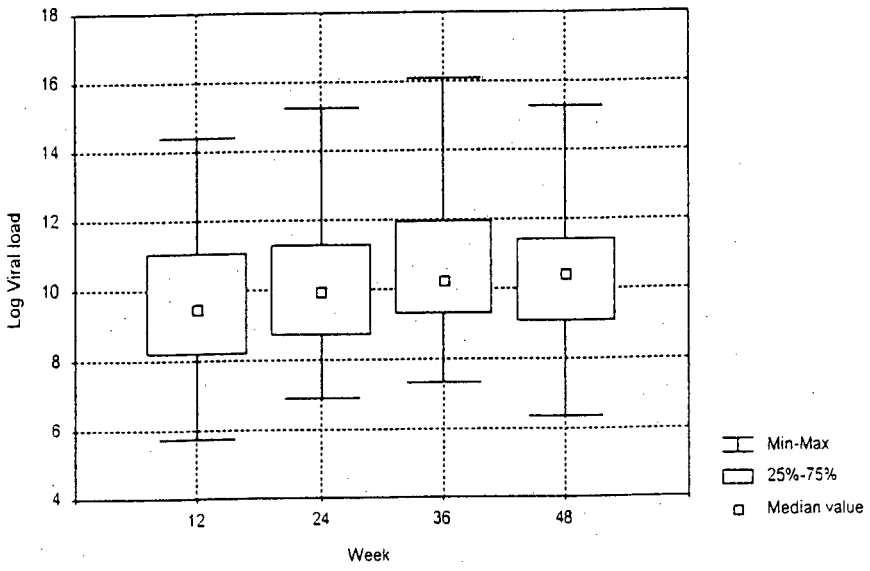


Figure 3.16. Profiles of Log CD4 counts for five selected patients

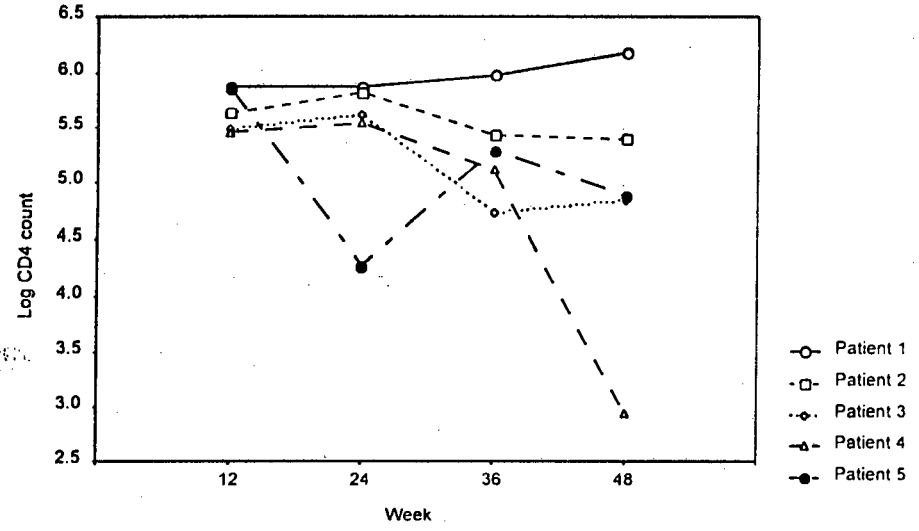


Figure 3.17. Profiles of Log Viral load for five selected patients

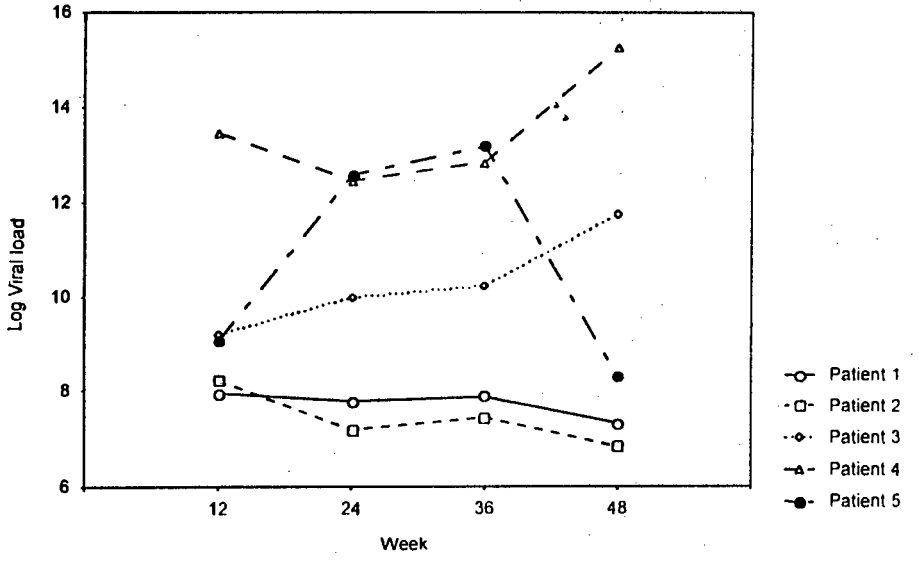


Table 3.11 presents the estimated correlation coefficients between the CD4 counts for the 56 patients by week; while Table 3.12 shows the correlations between viral load measurements by week. Both successive CD4 counts and viral load measurements were highly correlated. The correlations between successive CD4 counts were consistent with correlation estimates obtained in previous studies. Taylor *et al.*, (1989) reported a correlation of 0.64 between successive CD4 counts. The reason for the high correlations between the CD4 counts and the viral load measurements is unclear given the times between measurements (approximately 3 months for the data we have used in this analysis).

Correlation coefficients between the CD4 counts and viral load measurements for different patients for a given combination of weeks are presented in Table 3.13. The correlations point towards a weak negative relationship between the CD4 count and viral load; the correlations range between –0.33 to –0.40 (-0.27 to –0.57, on the log-scale) if the diagonal elements of the correlation matrices (Table 3.13 and Table 3.14) are inspected. However, this observation was consistent with our findings presented in Section 3.4.1 of this Chapter. The correlations indicate that the relationship between the CD4 count and viral load depends on time. Therefore the modelling of the CD4 count as a function of viral load would need to take account of this time dependence.

A scatter plot of the log-transformed data suggested a negative relationship between log CD4 count and log viral load (Figure 3.18) and that this relationship may be different for the different weeks of clinical examination (Figure 3.19). The log-transformed measurements appeared to be more variable in the 48^{th} week of measurement than in earlier weeks of measurement.

Table 3.11 Correlations between CD4 counts for 56 HIV patients

	Week 12	Week 24	Week 36	Week 48
Week 12	1.00	0.74	0.81	0.61
Week 24	0.74	1.00	0.83	0.59
Week 36	0.81	0.83	1.00	0.67
Week 48	0.61	0.59	0.67	1.00

Table 3.12 Correlations between Viral load measurements for 56 HIV patients

	Week 12	Week 24	Week 36	Week 48
Week 12	1.00	0.84	0.82	0.46
Week 24	0.84	1.00	0.81	0.42
Week 36	0.82	0.81	1.00	0.28
Week 48	0.46	0.42	0.28	1.00

Table 3.13 Correlations between CD4 counts and Viral load by week for 56 HIV patients

	Viral load			
CD4 count	Week 12	Week 24	Week 36	Week 48
Week 12	-0.33	-0.31	-0.29	-0.16
Week 24	-0.41	-0.40	-0.35	-0.18
Week 36	-0.41	-0.39	-0.33	-0.22
Week 48	-0.28	-0.23	-0.18	-0.30

Table 3.14 Correlations between log CD4 count and log Viral load by week for 56 HIV patients

	Log Viral load			
Log CD4	Week 12	Week 24	Week 36	Week 48
Week 12	-0.27	-0.44	-0.36	-0.25
Week 24	-0.35	-0.57	-0.53	-0.29
Week 36	-0.43	-0.58	-0.52	-0.33
Week 48	-0.39	-0.54	-0.49	-0.42

Figure 3.18. Scatterplot og Log Cd4 count against Log Viral load

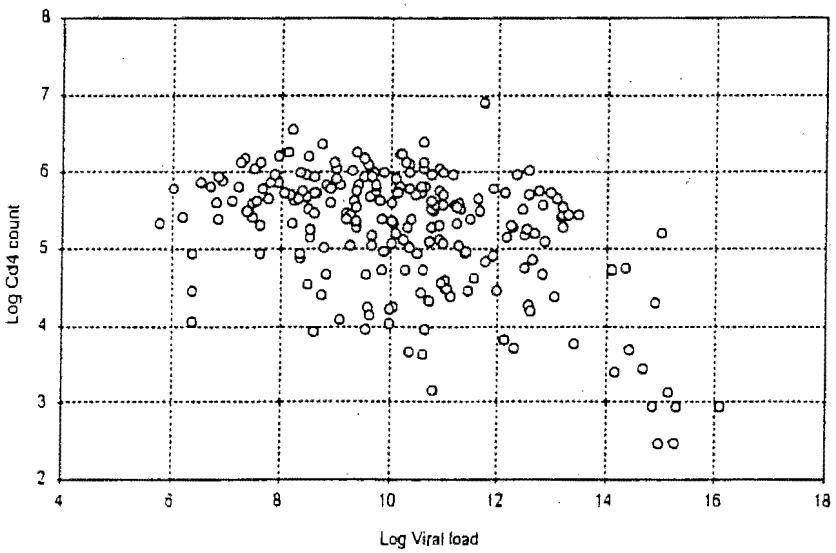
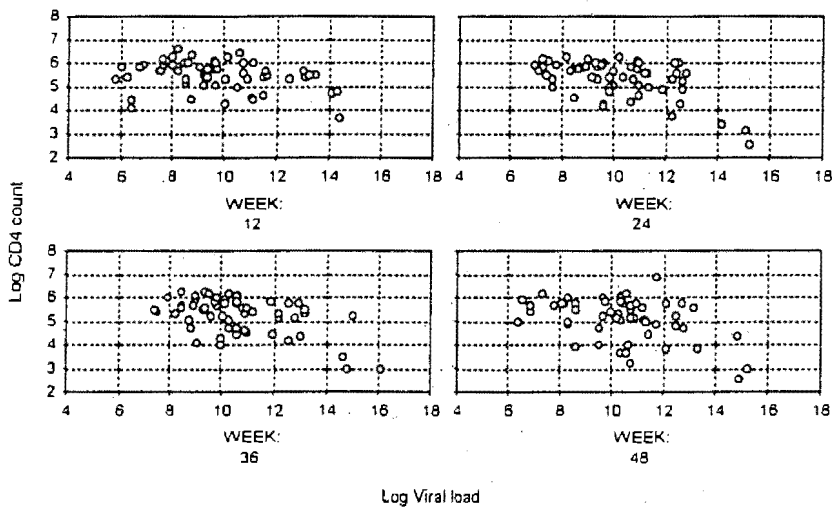


Figure 3.19. Scatterplot of Log CD4 count against Log Viral load by Week



3.4.2.2 Model Specification

A standard approach to modelling longitudinal data of this nature is to fit a random effects model, in which repeated observations for a subject are assumed to be correlated but observations from different patients are independent (Laird and Ware, 1982; Taylor *et al.*, 1994). The random effects model estimates the relationship between the response and some independent variable taking into account the dependence among repeated observations for a subject. The model also allows the partitioning of the variability of the response into two components; within and between subject variability. The model also allows us to estimate the correlation of two responses from the same patient, referred to as the within patient correlation.

In this situation, the random effects model that describes the relationship between CD4 count and viral load could be summarised as follows:

Let y_{ij} be the CD4 cell count at week j for patient i

x_{ij} be the viral load measurement at week j for patient i.

The mixed model for the i^{th} patient is given by:

$$\log y_{ij} = \mu + \beta_j \log x_{ij} + \alpha_j^{week} + u_i^{patient} + e_{ij} \text{ , for } i = 1,...,56 \text{ and } j = 1,...,4$$

where the **fixed** terms in the model are:

μ is the constant term,

β_j is the effect of the interaction between log viral load and week, for $j = 1,...,4$ with $\beta_1 = 0$

α_j is effect for week j, for $j = 1,...,4$ with $\alpha_1 = 0$

the **random** terms are:

u_i is the random effect for patient i, for $i = 1,...,56$

and e_{ij} is the random error.

u_i is assumed to be normally distributed with mean zero and variance σ_u^2, and e_{ij} is also normally distributed with mean zero and variance σ_e^2. It is further assumed that CD4 measurements taken on the same patient are correlated, but that CD4 cell counts from different patients are independent.

The variance of the CD4 count measured on patient i at week j is given by

$$\text{var}(y_{ij}) = \sigma_e^2 + \sigma_u^2,$$

Since CD4 counts from the same patient are correlated; we have

$$\text{cov}(y_{ij}, y_{i'j}) = \sigma_u^2 \quad \text{for } i \neq i'$$

In order to account for the correlation in CD4 counts from the same patient, we assume a common covariance structure on the residuals, e_i, for each patient and insist on independence between patients. This assumption can be written formally as

$$\text{var}(e_i) = C \text{ and } \text{cov}(e_i, e_{i'}) = 0 \text{ , for } i \neq i'$$

where C is the common covariance structure of the residuals in a patient. This type of covariance structure assumes that each observation (CD4 count) on a patient has the same correlation with each other observation measured on the same patient.

An alternative correlation structure, for the residuals, considered was a first-order autoregressive process structure. This model estimates the correlation, ρ, between two observations which are adjacent in time. This way of modelling the covariance structure meant that observations on the same patient that were further apart in time had smaller correlation.

3.4.2.3 Model Selection

The relationship between CD4 count and viral load was assumed to depend on time. The model selection process entailed fitting random effects models sequentially (Table 3.15 and 3.16). We examined the significance of the main effects (log viral load and week) by fitting a random effects model to the CD4 count data with patients as random effects and denoted this model as Model A. The Wald statistics for the fixed effects from this model are shown in Table 3.15; both terms were found to be statistically significant. In model B, we assessed the significance of an interaction between log viral load and week. The Wald test indicated that the interaction term was statistically significant and so it was retained in the model. This meant that the relationship between CD4 count and viral load depended on week of measurement i.e. CD4 count decline occurred with different slopes in patients with different values of viral load and also each measurement time (week) had its own intercept and slope. Figure 3.20 clearly depicts this finding.

In Model C (Table 3.16), we modelled the correlation in the observations as an autoregressive process of order 2. This assumption seemed reasonable given the high correlations between successive CD4 measurements (Table 3.11). The autoregressive process of order 2 term was found not to be statistically significant and so the model was refitted as an autoregressive process of order 1 (Model D). The correlation (ρ) between two observations in a patient, adjacent in time, was estimated to be 0.272. This model was the final model fitted to the CD4 count data. It included log viral, week and their interaction and patients were treated as random effects. We based all our inferences on this model.

The within patient correlation in the CD4 counts, was found to be 0.69 (Table 3.17). This finding was surprising given the wide intervals between CD4 count measurements. The variance ratio, the variation due to patient as a fraction of random error, was estimated to be 2.183; i.e. patient variation in the CD4 count was 2.183 times the variation due to random error.

Table 3.18 presents correlation coefficients from the fitted model by week. At week 12 the relationship between CD4 count and viral load was weak. The relationship seemed to improve with time (week) and it was strongest at week 24, with correlation coefficient of 0.463; this implies that viral load could only explain 21 percent of the variation in CD4 counts at week 24.

Table 3.15 Regression results of models fitted to the CD4 count data during the model selection process

Model A			Model B		
Fixed term	**Wald statistic**	**d.f.**	**Fixed term**	**Wald statistic**	**d.f.**
Log Viral load	17.1	1	Log Viral load	17.6	1
Week	15.9	3	Week	17.1	3
			Log Viral load × Week	15.8	3
Random term	**Variance component**	**Std. error**	**Random term**	**Variance component**	**Std. error**
Patients	0.390	0.082	Patients	0.386	0.080
Random error	0.146	0.016	Random error	0.136	0.015

Table 3.16 Autoregressive models fitted models fitted to the CD4 count data during the model selection process

Model C			Model D		
Fixed term	**Wald statistic**	**d.f.**	**Fixed term**	**Wald statistic**	**d.f.**
Log Viral load	16.0	1	Log Viral load	15.9	1
Week	13.2	3	Week	14.5	3
Log Viral load × Week	12.5	3	Log Viral load × Week	13.0	3
Random term	**Variance component**	**Std. error**	**Random term**	**Variance component**	**Std. error**
Patients	0.280	0.174	Patients	0.358	0.084
Patient × Week	0.242	0.159	Patient × Week	0.164	0.030
Autoreg. term			**Autoreg. Term**		
AR(1)	0.403	0.191	AR(1)	0.272	0.140
AR(2)	0.210	0.182			

Table 3.17 Estimates of regression parameters and their standard errors of the final model for the CD4 count data

Fixed term	**Parameter estimate**	**Std. error**	**Z-value**
Constant	5.483	0.300	18.277
Log Viral load	-0.003	0.029	0.103
Week 12	0.000		
Week 24	0.783	0.372	2.104
Week 36	0.645	0.372	1.734
Week 48	0.934	0.372	2.511
Log Viral load × Week 12	0.000		
Log Viral load × Week 24	-0.092	0.035	-2.629
Log Viral load × Week 36	-0.081	0.036	-2.250
Log Viral load × Week 48	-0.122	0.035	-3.486
Random term	**Variance Component**	**Std Error**	
Patients	0.358	0.084	
Random error	0.164	0.030	
Variance Ratio	2.183		
Within Patient Correlation	0.686		

Table 3.18 Correlation between log (CD4) and log(viral load) by Week

Week	Correlation coefficient (r)	Coefficient of determination (R^2)
12	-0.077	0.005
24	-0.463	0.214
36	-0.417	0.174
48	-0.401	0.161
All weeks	-0.418	0.175

Given a patient's viral load and week of measurement, the equations for estimating the CD4 count at each of the weeks, from the fitted model, were given by:

Let y_{ij} be the CD4 cell count at week j for patient i

x_{ij} be the viral load measurement at week j for patient i.

Week 12: $\log y_{ij} = 5.483 - 0.003 \times \log x_{ij}$

Week 24: $\log y_{ij} = 6.266 - 0.095 \times \log x_{ij}$

Week 36: $\log y_{ij} = 6.128 - 0.084 \times \log x_{ij}$

Week 48: $\log y_{ij} = 6.417 - 0.126 \times \log x_{ij}$

Since the fitted CD4 values were on the log-scale, to make them comparable to the observed CD4 counts they were back-transformed by taking exponents resulting in the following set of equations:

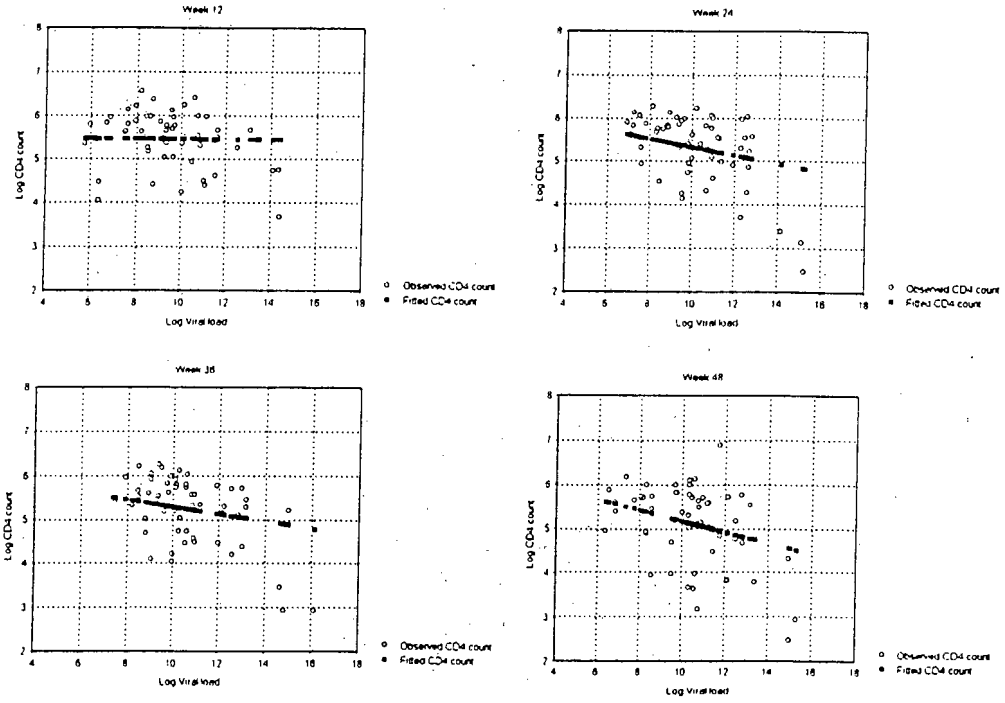
Week 12: $y_{ij} = exp(5.483 - 0.003 \times \log x_{ij})$

Week 24: $y_{ij} = exp(6.266 - 0.095 \times \log x_{ij})$

Week 36: $y_{ij} = exp(6.128 - 0.084 \times \log x_{ij})$

Week 48: $y_{ij} = exp(6.417 - 0.126 \times \log x_{ij})$

Figure 3.20. Scatterplot of Log CD4 count and Fitted value against Log Viral load by Week



3.4.2.4. Model checking

A plot of residuals against fitted values, by week of measurement, from the final model is shown in Figure 3.21. Figure 3.22 shows a plot of the residuals against week of measurement. Only two observations, observations of patients 17 and 23 had extreme residuals, their residuals being –2.35 and 1.97, respectively. The residuals did not appear to increase with time on study of patient. This finding gave an indication that the autoregressive process of order 1 correlation structure fitted to the data was adequate. Although the residuals were quite small in magnitude, the normal probability plots indicated substantial departure from normality for observations taken at week 24 and 36 (Figure 3.23).

Figure 3.21. Scatterplot of Log CD4 residual against Fitted values by Week

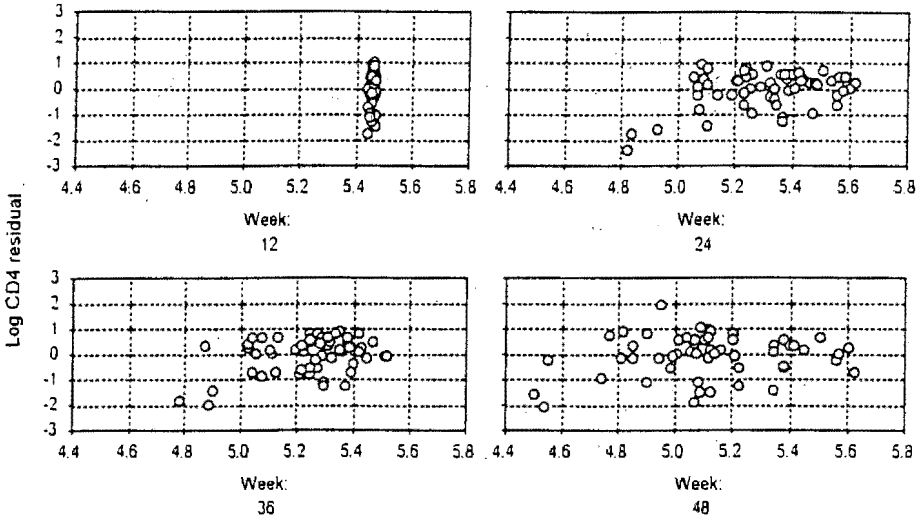


Figure 3.22. Plot of Log CD4 residuals by Week

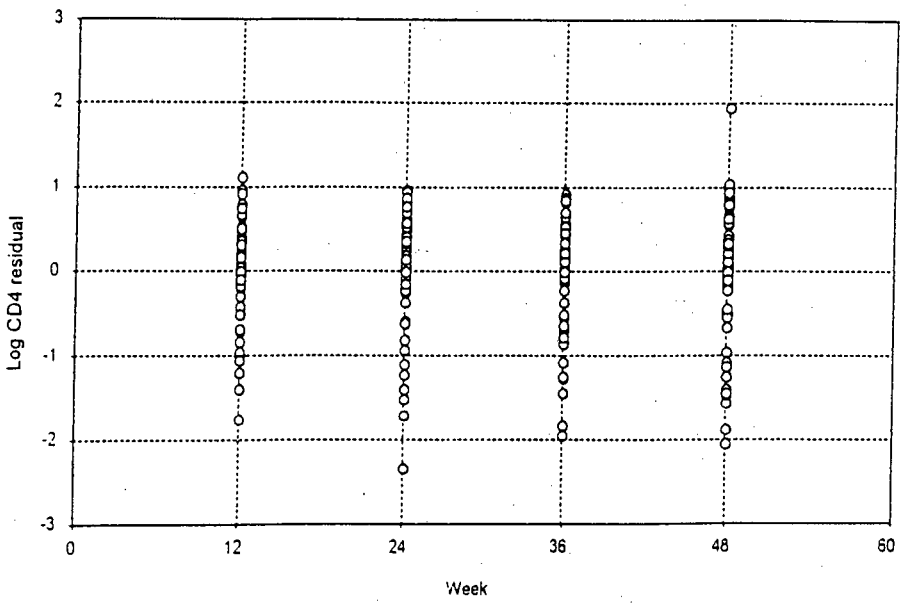
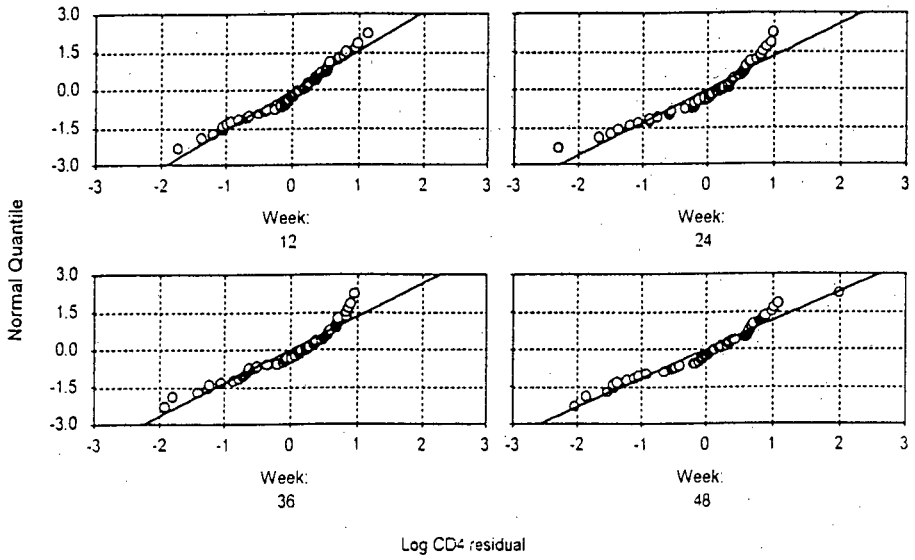


Figure 3.23. Normal Probability Plot of Log CD4 residual by Week



3.5 Discussion

The preceding analysis indicated that there was a weak relationship between the log CD4 count and log viral load. The analysis also revealed that there was considerable patient variation in the CD4 count, patient variation was 2.183 times the random variation in the CD4 counts.

The data used in this analysis came from patients in a clinical trial, and so these patients are not a random sample of HIV-infected subjects; nonetheless, the results were consistent with those reported in other studies of the relationship between the CD4 count and viral load. Previous studies of the relationship between these two markers have been conducted using data on HIV infected individuals from developed countries. In this study, we used measurements taken on African HIV infected individuals. Therefore the results showed that the relationship between the CD4 count and viral load is the same in both the HIV population from developed countries and those from Africa.

A limitation of the analysis was that other factors not considered in the model could affect relationship between the CD4 count and viral load. Among them the length of HIV infection, patient's age and HIV stage. In particular, the relationship between the two markers might not be the same at all stages of the HIV disease; it is not known whether the relationship increases or decreases with disease progression.

CHAPTER 4

Some other approaches to modelling the CD4 count

4.1. Introduction

In the preceding analyses (Chapter 2 and Chapter3) it was assumed that logarithm of the CD4 count was normally distributed. The linear mixed models fitted further assumed that the random (patient) effects were also normally distributed. Other authors have considered models for both the natural logarithm and square root transformation of the CD4 count (Faucet and Thomas, 1996; Lange *et al.*, 1992; and Self and Pawitan, 1992). In these studies both the square root of the CD4 and log CD4 were assumed to be normally distributed.

In this chapter we briefly explore some other approaches to the analysis of markers of HIV disease progression, the CD4 count in particular. We assume the CD4 counts are Poisson distributed and apply appropriate statistical methods for analysing longitudinal data when the response is non-normal. The methods we employ also allow for some flexibility in the distributional assumptions about the random effects, that is, the random effects are no longer restricted to be normal.

Generalized linear models (GLMs) (McCullagh and Nelder, 1989; Nelder and Wedderburn, 1972; Wedderburn, 1974) are standard statistical methods for modelling discrete and continuous response variables that can be assumed to be independent. In these models all the explanatory variables are assumed to be fixed and the observations come from a distribution in the exponential family. However, in some circumstances, it is useful to assume that some of the explanatory variables are randomly distributed. If the random explanatory variables are assumed to be normally distributed, the resultant models are called generalized linear mixed models (GLMMs).

Many researchers have investigated the extension of random effects to GLMs. Williams (1982) considered the beta-binomial random effects GLM. Poisson-gamma models were first studied by Breslow (1984). Stiratelli, Laird and Ware (1984) and Anderson and Aitkin (1985) considered multiple logistic regression models with normally distributed random effects using expectation maximization (EM) and Newton-Raphson algorithms, respectively. Harville and Mee (1984) investigated random effects models for ordered categorical data. For Poisson distributed data, Breslow (1984), Crowder (1985) and Tsutakawa (1985) have investigated log-linear models with random effects. Zeger and Karim (1991) considered the GLMM in the Bayesian framework. They estimated the parameters of a GLMM using a Monte Carlo method, the Gibbs sampler (Gelfand and Smith, 1990; Geman and Geman, 1984).

An approach to the analysis of correlated responses was given by Liang and Zeger (1986) and was further discussed by Zeger *et al.*, (1988) and Liang *et al.*, (1992). They adopted a quasi-likelihood approach (McCullagh and Nelder, 1989; Wedderburn, 1974) and modelled the marginal expectation of the response rather than the conditional expectation given the random (subject) effect. Within this class of

models, Zeger *et al.*, (1988) distinguished between subject-specific and population-averaged models. In the population-averaged model, the regression coefficient is interpreted as the change in the 'population-averaged' response rather than the change in a subject's expected response with the covariates. On the other hand in the subject-specific model, the regression coefficient describes how a subject's response depends on the covariates.

Lee and Nelder, (1996) consider GLMMs where the distribution of the random effects is not restricted to be normal. These models are called hierarchical generalized linear models (HGLMs) and include GLMMs. George *et al.*, (1993) considered Bayesian hierarchical models with conjugate prior distributions.

HGLMs provide a more flexible approach to parametric modelling of the distribution of independent and identically distributed random effects. The distribution of the random effects are assumed to come from an arbitrary distribution, conjugate to that of the response, y. They (HGLMs) also provide a natural framework for model diagnostics, through the checking of assumptions about the distribution of y given the random effect and that of the random effects.

4.2 Methodology

In this study we focus on HGLMs since they generalize the linear mixed models we considered in Chapter 2 and 3; and they (HGLMs) can be considered to be generalizations of GLMMs. The following is a brief description of the general form of a HGLM we employ in the analysis.

Let y be the outcome variable of scientific interest, and u be the random component. Conditional on random effect u, we have a GLM (McCullagh and Nelder, 1989) where y follows an exponential family distribution such that

(a) The conditional likelihood of y given u has the form

$$l(\theta^{'},\phi;y|u)=\{y\theta^{'}-b(\theta^{'})\}/a(\phi)+c(y,\phi)$$

where $\theta^{'}$ refers to the unknown parameters relating to the mean of the distribution of y given u. ϕ is the dispersion parameter and is associated with the variance of the distribution of y given u.

Let $E(y|u)=\mu^{'}$; then $\mu^{'}$ is related to the linear predictor by the link function g such that $\eta^{'}=g(\mu^{'})$ and we have

$$g(\mu^{'})=\eta^{'}=X\beta+v$$

where v is some function u.

(b) u has a distribution, which is not restricted to be normal.

Different assumptions about the distribution of y given u and that of v, the random component, lead to different forms of the HGLM. Some of the possible models are given below. In these models the linear predictor is given by $\eta = X\beta + v$.

i) **Normal-Normal HGLM**: In this model both the conditional distribution of y given u and the distribution of u are assumed to be normal with the identity link function. The random error has a mean of zero and variance σ_e^2 while the random effect, u_i, has a mean of zero and variance σ_u^2. The dispersion parameter, ϕ, is equal to the residual variance (random error), σ_e^2.

ii) **Poisson-Normal HGLM**: In this model, it is assumed that y given u is distributed Poisson with mean $E(y|u) = \mu u$ and in the linear predictor, η, $v = \log u$. It is further assumed that v is distributed normally and so the distribution of u is log-normal. The dispersion parameter is equal to 1. When overdispersion is present, the conditional distribution of y given u is assumed to be overdispersed Poisson such that $E(y|u) = \mu u$ and $\text{Var}(y|u) = \phi\mu u$. This model is a GLMM and could be called the Poisson-log-normal HGLM (Lee and Nelder, 1996).

iii) **Poisson-Gamma HGLM**: This model assumes that the distribution of y given u is Poisson and that of u is Gamma. In the linear predictor, η, $v = \log u$ and so v is distributed as the log-gamma. The random effect u is distributed gamma with mean 1 and shape parameter α. Again, the dispersion parameter is 1. When overdispersion is present $\text{Var}(y|u) = \phi\mu u$.

Patterson and Thompson introduced the restricted likelihood for estimation of the dispersion parameters for normal-normal models Several authors have developed methods for estimating the parameters of GLMMs (Breslow and Clayton, 1993; Engel and Keen, 1994;Gilmour *et al.*, 1985;McGilchrist, 1994; Schall, 1991).

Parameter estimation in HGLMs is achieved through maximizing the hierarchical likelihood, $\log - h - \text{likelihood}$ (Lee and Nelder, 1996), an analogue of Henderson's mixed model likelihood equations (Henderson, 1975). The $\log - h - \text{likelihood}$ is defined as

$$h = l(\theta', \phi; y|v) + l(\alpha; v)$$

where $l(\alpha; v)$ is the logarithm of the density function for v with parameter α, and $l(\theta', \phi; y|v)$ is the log-likelihood for $y|u$ with unknown parameters θ' and ϕ. The $\log - h - \text{likelihood}$ is not a joint likelihood in the orthodox sense because v or u are not observed.

The parameter estimates of a HGLM are then solutions of the equations:

$$\partial h/\partial\beta = 0, \ \partial h/\partial v = 0$$

An adjusted profile $h-$likelihood is used to estimate of the dispersion parameters, ϕ ($=\sigma_e^2$ for Normal-Normal HGLM). This is an extension of the adjusted profile likelihood of Lindsey (1994, page 112). The dispersion parameters may also be estimated by an extended quasi-h-likelihood, an analog of Wedderburn's (1974) quasi-likelihood equations. In both methods the estimation of the fixed and random effects precedes the estimation of the dispersion parameters, ϕ.

Having fitted any statistical model, it is important to ascertain the validity of its underlying assumptions. Model checking in HGLMs is performed by examination plots of residuals from the fitted model(s). Distributional assumptions about the random component v or u can be checked by fitting appropriate distributions or by graphical examination of the half-normal plot of deviance residuals from the fitted model (Lee and Nelder, 1996).

4.3 Results

To fit the above HGLMs, we used two data sets. The first data set consisted of clinical records of 376 patients presenting with HIV infection at the Somerset Hospital HIV Clinic over the period 1984 to 1997; this is the data set we used in Chapter 2 of this study to investigate the relationship the between CD4 count and TLC. An analytical challenge presented by this data was that the timing and number of measurements taken varied from patient to patient so that the data was unbalanced and unequally spaced. Furthermore, the observations came from patients for which there was no explicit criteria of selection.

The second data set consists of CD4 counts and viral load measurements taken on 56 HIV patients enrolled in a clinical trial at the Somerset Hospital HIV Clinic; this is the data set we used in Chapter 3 of this study to examine the relationship between the CD4 count and viral load. For this data set the CD4 counts and viral load measurements were taken at 12-weekly intervals such that the data was balanced and equally spaced.

We also briefly investigated a Bayesian approach to modelling the relationship between the CD4 count and viral load using the data from the clinical trial, involving 56 patients.

4.3.1 The relationship between CD4 counts and total lymphocyte counts

We consider the following hierarchical generalized linear models (HGLMs) for relationship between CD4 count and total lymphocyte count (TLC):

Let y_{ijk} be the CD4 cell count at visit j for patient i in HIV stage k

x_{ijk} be the total lymphocyte count at visit j for patient i in HIV stage k

1.) Model I: Normal-Normal HGLM

In this model the CD4 count was transformed using the natural logarithm and the log-transformed CD4 count were then assumed to be normally distributed.

This model had the linear predictor

$$\log y_{ijk} = \mu + \beta(\log x_{ijk}) + \alpha_k^{stage} + u_i^{patient} + e_{ijk} \quad \text{, for } i = 1,...,376,\ k = 1,...,4 \text{ and } j = 1,...,n_i$$

The model assumed that u_i was to be normally distributed with mean zero and variance σ_u^2, and e_{ijk} was also normally distributed with mean zero and variance σ_e^2 and $\text{Var}(y_{ijk}|u_i) = \sigma_u^2 + \sigma_e^2$. The dispersion parameter, ϕ, was given by σ_e^2 .

This was the final fitted model in Chapter 2 of this study.

ii) Model II: Poisson Generalized Linear Model

It was assumed that $y_{ijk} \sim Poisson(\mu_{ijk})$. The linear predictor was given by

$$\log \mu_{ijk} = \mu + \beta(\log x_{ijk}) + \alpha_k^{stage} \quad \text{, for } i = 1,...,376,\ k = 1,...,4 \text{ and } j = 1,...,n_i$$

The model has the properties: $E(y_{ijk}) = \mu_{ijk}$ and $\text{Var}(y_{ijk}) = \mu_{ijk}$

The model did not include a random component. The dispersion parameter, ϕ, was set equal to 1.

iii) Model III: Poisson-Normal HGLM

The model assumed $y_{ijk}|u_i \sim Poisson(\mu_{ijk}u_i)$ and $\log u_i \sim Normal(0,\sigma_u^2)$

The model has the linear predictor

$$\log \mu_{ijk} = \mu + \beta(\log x_{ijk}) + \alpha_k^{stage} + \log u_i^{patient} \quad \text{, for } i = 1,...,376,\ k = 1,...,4 \text{ and } j = 1,...,n_i$$

with the properties: $E(y_{ijk}|u_i) = \mu_{ijk}u_i$ and $\text{Var}(y_{ijk}|u_i) = \mu_{ijk}u_i$

The dispersion parameter, ϕ, taken to be 1.

Model IV: Poisson-Gamma HGLM

The model assumed $y_{ijk}|u_i \sim Poisson(\mu_{ijk}u_i)$ with the linear predictor

$$\log \mu_{ijk} = \mu + \beta(\log x_{ijk}) + \alpha_k^{stage} + \log u_i^{patient} \quad , \text{ for } i = 1,...,376, \quad k = 1,...,4 \text{ and } j = 1,...,n_i$$

with the properties: $E(y_{ijk}|u_i) = \mu_{ijk}u_i$ and $\text{Var}(y_{ijk}|u_i) = \mu_{ijk}u_i$

The model assumed that the random effect, u_i, was distributed gamma with mean 1 and shape parameter, $\alpha (= 1/\sigma_u^2)$. The dispersion parameter, ϕ, was set equal to 1.

In the above models the **fixed** terms are:

μ which represents the constant term,
β represents the effect of the logarithm of TLC
α_k represents effect of HIV stage k, for $k = 1,...,4$ with $\alpha_1 = 0$

the **random** terms are:

u_i represent the random effect for patient i, for $i = 1,...,376$

We also investigated overdispersion by considering the following models:

iv) Model V: Poisson Generalized Linear Model with overdispersion

This model was an extension of model II with the properties

$$E(y_{ijk}) = \mu_{ijk} \text{ and } \text{Var}(y_{ijk}) = \phi\mu_{ijk}$$

vi) Model VI: Poisson-Normal HGLM with overdispersion

This model was an extension of model III with properties

$$E(y_{ijk}|u_i) = \mu_{ijk}u_i \text{ and } \quad \text{Var}(y_{ijk}|u_i) = \phi\mu_{ijk}u_i$$

vii) Model VII: Poisson-Gamma HGLM with overdispersion

This model was an extension of model IV with the properties

$$E(y_{ijk}|u_i) = \mu_{ijk}u_i \text{ and } \text{Var}(y_{ijk}|u_i) = \phi\mu_{ijk}u_i$$

The underlying assumptions in each of the above models (HGLMs) have been discussed, in general, in Section 4.2. of this Chapter. Table 4.1. gives a summary of the assumptions made in each of these models.

We used the K-system and HG-system (Nelder, 1993), which are incorporated in GENSTAT, to fit the above HGLMs. The procedures used to fit the HGLMs estimate the variance components on the log-scale to achieve both faster convergence and non-negative variance estimators (Lee and Nelder, 1996). The delta- method was used to obtain appropriate variances for the random components on the scale of the data

The K-system allows for a variety of model-checking statistics, which could be used to assess the adequacy of the fitted HGLMs. These statistics are William's one-step cross-validatory residual, one-step approximation to Cook's statistic, Studentized deviance residuals, Studentized Pearson's residuals, Atkinson's deletion residuals, Atkinson's modified Cook's statistic. In this analysis, we used Studentized deviance residuals check for outliers and the Modified Cook's statistics to check for influential observations in the parameters of the models fitted.

The results from fitting the above-specified models are given in Table 4.2. (a) and Table 4.2. (b). Model II was the only model we considered that did not involve a random component in the linear predictor; the other models included random effects. The regression coefficient estimates from these models were consistent with the results from fitting random effects models in Chapter 2 of this study.

The regression coefficients of models I, III and IV, the models involving random components, had larger standard errors compared to the Poisson GLM. This was caused by the additional variance due to random effects (Lee and Nelder, 1996). We found that model III and model IV gave similar regression coefficients and standard errors but had different estimates for the variance due to the random component, the variation in the CD4 count due to patients. This finding confirms Lee and Nelder's (1996) conclusion that conjugate HGLMs and GLMMs often give similar results.

If the models, which assumed a Poisson distribution for the fixed effects (models II, III and IV), were appropriate we would expect the residual deviance to be approximately the same as the residual degrees of freedom. The degree of dispersion (deviance/degrees of freedom) was 75.875, 28.146 and 28.257 for models II, III and IV, respectively. For instance for the Poisson-Normal model (model III), the overdispersion implied that confidence intervals for the regression coefficients would be increased by a factor of $\sqrt{28.146} = 5.305$, giving an increase in width of more than five times. This indicated overdispersion in CD4 count; thus models V, VI and VII, which would take account of the overdispersion, were fitted. We observed that for the

models involving a random component in the linear predictor, the variation due to patient decreased when overdispersion was explicitly taken into account in the model. It was also was noted that standard errors of the regression coefficient estimates of the overdispersed Poisson models were increased over those of models without overdispersion.

In the models V, VI and VII, the dispersion parameter, ϕ, was assumed to be a constant. We also investigated a structured dispersion on the fixed effect for HIV stage; this implied the dispersion depended on HIV stage. The decision whether to use a structured dispersion on HIV stage or a constant dispersion seemed to matter little in the results. In general, the structured dispersion would be more appropriate if the mechanism that causes the overdispersion were known exactly (McCullagh and Nelder, 1989, page 199).

For comparative purposes the normal probability plots of deviance residuals for each of the models are shown in Figure 4.1 (a) and Figure 4.1 (b). The deviance residuals for the models without overdispersion (models II, III and IV) were much larger than those of models with overdispersion (Figure 4.1. (b)). This was due to overdispersion in the CD4 count. Models II, III and IV identified the same outliers, these were observations of patients 31, 48, 60, 78, 85 and 210. The overdispersed models also identified the same outliers, these observations belonged to patients 6, 31, 60, 85 and 210. Despite this lack of fit the overdispersed Poisson models appeared to fit the data better than the models without overdipersion.

Figure 4.2 presents modified Cook's statistics from fitting models VI and VII. Several observations taken on different patients at different stages of the HIV disease were found to be influential in each of these models. For instance, in model VI, observations belonging to patients 6, 31, 48, 60, 68, 210 and 335 were influential, while in model VII, observations of patients 60, 210, 214, 242, 333, 334 and 335 were influential. We refitted the models VI and VII with the influential observations removed. The regression coefficient estimates from the refitted models (Table 4.5) were similar to those obtained when the influential observations were included (Table 4.2. (b)).

Distributional assumptions about the random effects, u_i, for the HGLMs involving a random component, were checked by conducting goodness of fit tests and fitting the assumed distributions to the random effects estimates. Table 4.3 presents the results of the goodness of fit tests. The assumptions about the random effects were met for the models (models VI and VII) with overdispersion and did not hold for the models without overdispersion. Figure 4.3. (a) and Figure 4.3. (b) show the fitted distributions to the random effects of the models without overdispersion and those with overdispersion, respectively.

We also assessed the practical validity of models VI and VII by using the regression coefficient estimates to estimate the CD4 from TLC. Again we used observations used in Chapter 2 of this study to estimate the CD4 counts. Table 4.4. shows a distribution of the relative distances of the estimated values from the observed CD4 counts. The predictions from these models were not improved over those obtained from fitting a random effects model in Chapter. This might be due to the weak relationship between the CD4 and TLC.

Random effects and overdispersion modelling

In the specification of a random effects generalized linear model, the conditional likelihood of y given u is given by

$$l(\theta', \phi; y|u) = \{y\theta' - b(\theta')\}/a(\phi) + c(y, \phi)$$

where θ' refers to the unknown parameters relating to the mean of the distribution of y given u. ϕ is the dispersion parameter and is associated with the variance of the distribution of y given u.

When the distribution of y given u is assumed to be normally distributed, ϕ is the variance, $\text{Var}(y|u)$, $(\phi = \sigma_e^2)$, which is strictly positive. In the binomial and Poisson distribution $\phi = 1$. For the Poisson distribution we have $E(y|u) = \mu u$ and $\text{Var}(y|u) = \mu u$.

If the conditional distribution of y given u is assumed to be Poisson or binomial it may be overdispersed or underdispersed. For instance if conditional distribution of y given u is assumed to be Poisson, we have $E(y|u) = \mu u$ but $\text{Var}(y|u) = \phi \mu u$. Overdispersion or underdispersion can be interpreted as the departure from the assumed conditional distribution of y given u.

In the literature the terms dispersion or overdispersion and random effects have been used interchangeably to refer to the modelling of the random component. For instance Lee and Nelder; (1996) state that modelling of the random component in the linear predictor describes overdispersion. Overdispersion in a generalized linear model, with random effects, relates to the conditional distribution of y given u and is distinct from the need of postulated model to have an extra random effect. In the models with overdispersion we fitted; instead of estimating the overdipersion parameter, ϕ, explicitly, we could have estimated it by fitting an extra random component for the units, u_{ij}. The results would have the same interpretation according to Lee and Nelder, (1996). We consider dispersion to relate to the structure of the random effect u while overdispersion refers to variation in addition to the specified random effects u. This is useful for model interpretation since the random component represents a known source of variation and the overdispersion corresponds to unknown sources of variation.

Table 4.1 Summary of assumptions of hierarchical generalized linear models for the CD4 count data

Model	Term	Mean			Dispersion		
		Distribution	Link	Linear Predictor	Distribution	Link	Linear Predictor
I	**Fixed**	Normal	Identity	log TLC + Stage	Gamma	log	
	Random	Normal	Identity	Patient	Gamma	log	
II	**Fixed**	Poisson	Log	log TLC + Stage	Gamma	log	
	Random						
III	**Fixed**	Poisson	log	log TLC + Stage	Gamma	log	
	Random	Normal	identity	Patient	Gamma	log	
IV	**Fixed**	Poisson	log	log TLC + Stage	Gamma	log	
	Random	Gamma	log	Patient	Gamma	log	
V	**Fixed**	Poisson	log	log TLC + Stage	Gamma	log	
	Random						
VI	**Fixed**	Poisson	log	log TLC + Stage	Gamma	log	
	Random	Normal	identity	Patient	Gamma	log	
VII	**Fixed**	Poisson	log	log TLC + Stage	Gamma	log	
	Random	Gamma	log	Patient	Gamma	log	

Table 4.2. (a) Parameter estimates of the models without overdispersion for the CD4 count data

Term	Model I		Model II		Model III		Model IV	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
Fixed								
Constant	-2.583	0.310	-0.489	0.032	-0.705	0.083	-0.518	0.083
Log TLC	1.117	0.041	0.859	0.004	0.844	0.010	0.842	0.010
Stage II	-0.074	0.055	0.003	0.005	0.000	0.008	0.001	0.008
Stage III	-0.369	0.051	-0.406	0.005	-0.166	0.009	-0.165	0.009
Stage IV	-0.756	0.068	-0.684	0.007	-0.297	0.014	-0.295	0.014
Random								
Patient	0.239	0.010			0.416	0.031	0.356	0.026
Dispersion	0.176	0.021	1.000		1.000		1.000	
Deviance	675.6		70716		15851.0		15841.0	
Degrees of freedom	675.6		932		562.8		560.6	

Table 4.2. (b) Parameter estimates of the models with overdispersion for the CD4 count data

Term	Model V		Model VI		Model VII	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
Fixed						
Constant	-0.489	0.275	-1.116	0.301	-1.157	0.304
Log TLC	0.859	0.036	0.928	0.040	0.941	0.041
Stage II	0.003	0.040	-0.039	0.039	-0.034	0.038
Stage III	-0.406	0.043	-0.273	0.041	-0.262	0.041
Stage IV	-0.684	0.061	-0.576	0.059	-0.565	0.059
Random						
Patient			0.169	0.015	0.168	0.015
Dispersion	76.000		3.410	0.054	3.407	0.054
Deviance	70716		682.1		676.9	
Degrees of freedom	932		682.1		676.9	

Table 4.3 Goodness of fit tests for random effects estimates of each model

Model*	H_0	χ^2 statistic	Degree of freedom
I	$u \sim$ Normal	38.019	15
III	$u \sim Log-$normal	75.409	21
IV	$u \sim$ Gamma	37.301	19
VI (with overdispersion)	$u \sim$ Log - normal	9.363	13
VII (with overdispersion)	$u \sim$ Gamma	7.353	11

* Model II and V not shown because they did not involve random effects.

Table 4.4 Distribution of relative distances of estimated CD4 counts to observed CD4 count

Within k % of observed CD4 count	Model V		Model VI		Model VII	
	Freq.	Percent	Freq.	Percent	Freq.	Percent
< 10	1364	66.5	1367	66.6	1379	67.2
10-20	352	17.1	368	17.9	334	16.3
20-30	107	5.2	97	4.7	110	5.4
> 30	230	11.2	221	10.8	230	11.2
Total	2053	100.0	2053	100.0	2053	100.0

Figure 4.1. (a). Normal probability plots of deviance residuals for models without overdispersion

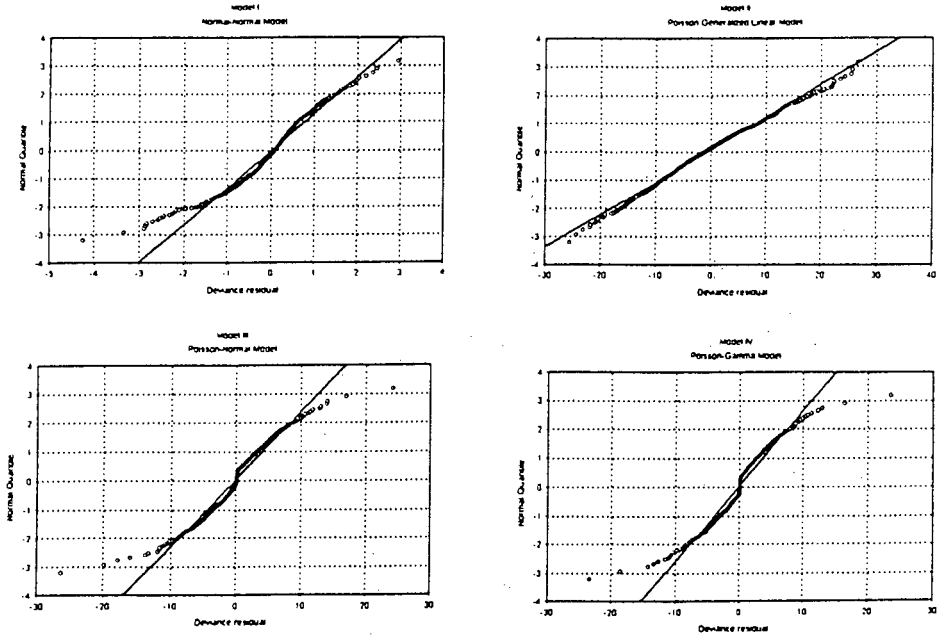


Figure 4.1. (b). Normal probability plot of deviance residuals for models with overdispersion

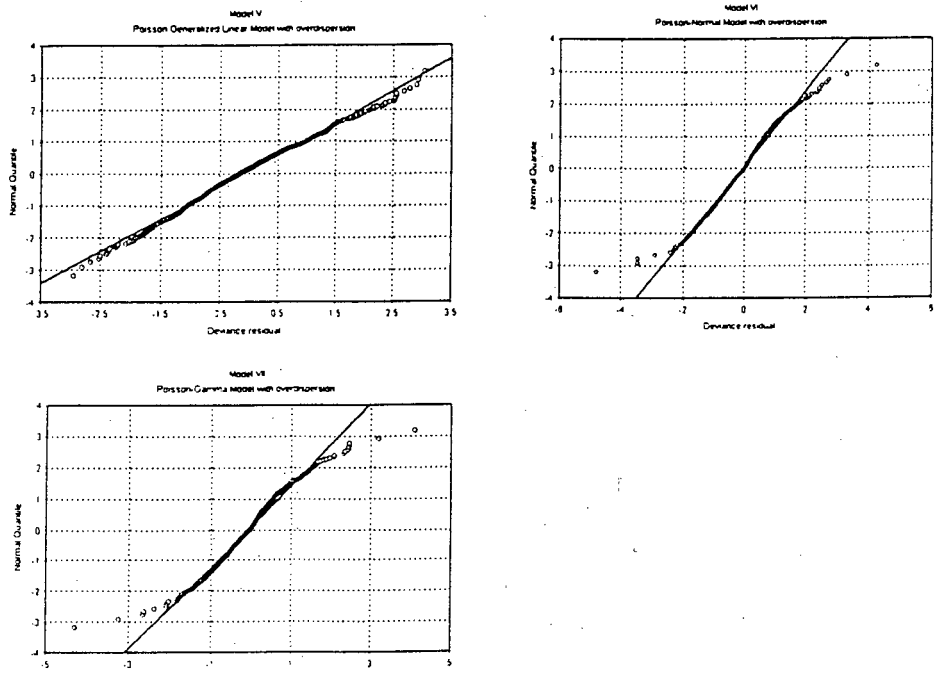


Figure 4.2. Index plots of Modified Cook's statistics
Model VI and Model VII

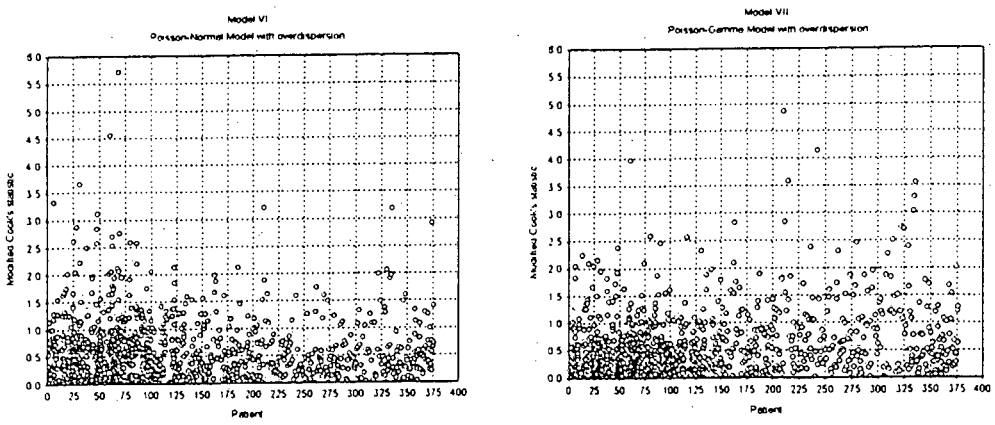


Figure 4.3. (a) Distribution of Random effects for models without overdispersion

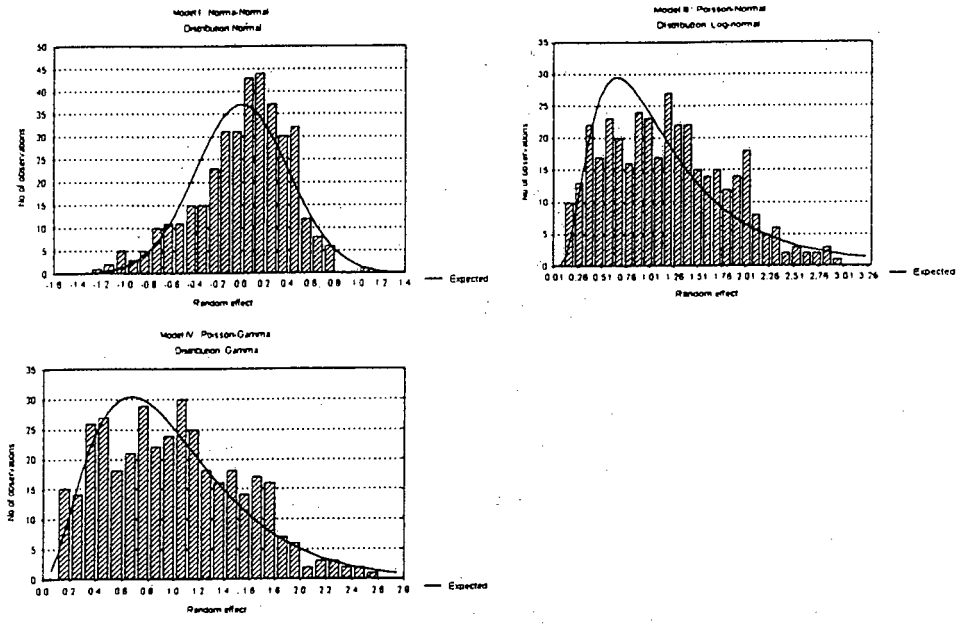


Figure 4.3. (b). Distribution of Random effects for models with overdispersion

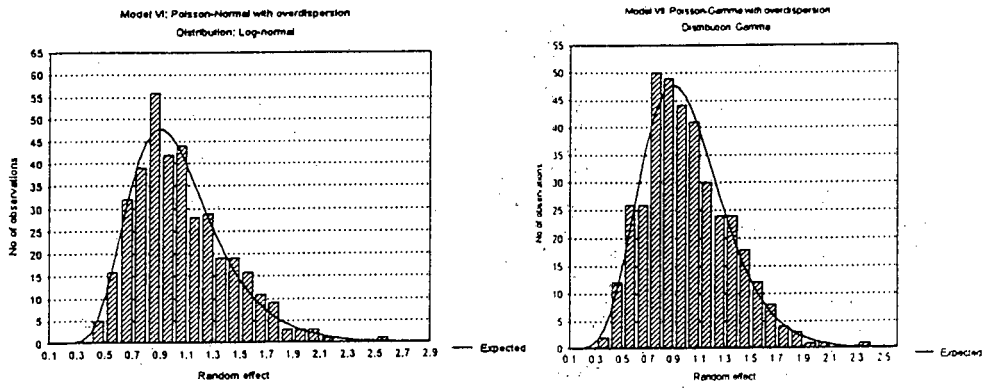


Table 4.5 Parameter estimates from the models VI and VII with influential observations excluded

Term	Model VI		Model VII	
	Estimate	Std. Error	Estimate	Std. Error
Fixed				
Constant	-1.247	0.290	-1.228	0.295
Log TLC	0.944	0.039	0.952	0.039
Stage II	-0.037	0.036	-0.040	0.037
Stage III	-0.259	0.039	-0.265	0.039
Stage IV	-0.549	0.056	-0.564	0.057
Random				
Patient	0.178	0.015	0.166	0.015
Dispersion	3.243	0.055	3.407	0.054
Deviance	662.1		666.9	
Degrees of freedom	662.1		666.9	

4.3.2 The relationship between CD4 counts and viral load

4.3.2.1 Modelling the CD4 count using HGLMs

We refitted the models considered in Section 4.3.1 of this Chapter using data for 56 HIV patients. The data consisted of repeated measurements of CD4 and viral for each patient.

The assumptions and notation for the HGLMs given in Section 4.3.1 were retained; only in the linear predictor were they changed since a different data set was used. For instance the logarithm of viral load and week of measurement represented the explanatory variables, instead of log TLC and HIV stage, in the models we considered.

The following hierarchical generalized linear models (HGLMs) describing the relationship between CD4 count and viral load were considered:

Let y_{ij} be the CD4 cell count at week j for patient i

x_{ij} be the viral load measurement at week j for patient i.

i) Model I: Normal-Normal HGLM

This model has the linear predictor

$$\log y_{ij} = \mu + \beta_j \log x_{ij} + \alpha_j^{week} + u_i^{patient} + e_{ij} \text{ , for } i = 1,...,56 \text{ and } j = 1,...,4$$

ii) Model II: Poisson Generalized Linear Model

The linear predictor in this model is given by

$$\log \mu_{ij} = \mu + \beta_j \log x_{ij} + \alpha_j^{week} \text{ , for } i = 1,...,56 \text{ and } j = 1,...,4$$

iii) Model III: Poisson-Normal HGLM

The linear predictor in this model is given by

$$\log \mu_{ij} = \mu + \beta_j \log x_{ij} + \alpha_j^{week} + \log u_i^{patient} \text{ , for } i = 1,...,56 \text{ and } j = 1,...,4$$

iv) Model IV: Poisson-Gamma HGLM

The linear predictor in this model is given by

$$\log \mu_{ij} = \mu + \beta_j \log x_{ij} + \alpha_j^{week} + \log u_i^{patient} \text{ , for } i = 1,...,56 \text{ and } j = 1,...,4$$

In these above model the **fixed** terms are:

μ which represents the constant term,

β_j represents the effect of the interaction between log viral load and week, for $j = 1,...,4$ with $\beta_1 = 0$

α_j represents effect for week j, for $j = 1,...,4$ with $\alpha_1 = 0$

and the **random** term is:

u_i represents the random effect for patient i, for $i = 1,...,56$

and e_{ij} represents the random error in Model I.

Table 4.6 gives a summary of the assumptions made in each of these models.

We also investigated overdispersion for the models that assumed a Poisson distribution for the fixed effects. These models were extensions of models II, III, and IV and were denoted as model V, VI and VII, respectively.

The results from fitting the above-specified models are given in Table 4.7. (a) and Table 4.7 (b). Model II was the only model we considered that did not involve a random component in the linear predictor; the other models included random effects. The regression coefficient estimates from these models were consistent with the results obtained from fitting random effects models in Chapter 3 of this study. However, when overdispersion was considered (models V, VI and VII), the interaction terms between the week of measurement and the log-transformed viral load were not statistically significant. For the Poisson generalized linear model, with overdispersion (model V), the relationship between the CD4 count did not depend on time. The estimates of the regression coefficients from fitting this model were similar to those obtained from the analysis of single measurements in section 3.4.1 of Chapter 3, in this study. This implies that the overdispersion, which explains all unknown sources of variation, has taken account of the time dependence.

Figure 4.4. (a) and Figure 4.4. (b) show normal probability plots of deviance residuals for each of the models. The deviance residuals for the models without overdispersion (models II, III and IV) were much larger than those of models with overdispersion (Figure 4.4. (b)). This was due to overdispersion in the CD4 count. Models I, II, III and IV identified the same outliers, these were observations of patients 9 and 17. However the models with overdispersion identified the observation of patient 17, taken at week 48, as the only outlier. From the data, the CD4 count corresponding to this observation was 993 and was larger compared to other observations at week 48, although it was a genuine observation.

Figure 4.5 presents modified Cook's statistics from fitting models VI and VII. In both models, observations of patients 4 and 17 were found to be influential.

Goodness of fit tests for the null distributions of the random effects, u_i, in the HGLMs involving a random component, revealed that the assumptions about the random effects were only valid for the models VI and VII but did not hold for the models without overdispersion (Table 4.8). Figure 4.6. (a) and Figure 4.6. (b) show the fitted distributions to the random effects of the models without overdispersion and those with overdispersion, respectively.

Table 4.6 Summary of assumptions of hierarchical generalized linear models for CD4 count data

Model	Term	Mean			Dispersion		
		Distribution	Link	Linear Predictor	Distribution	Link	Linear Predictor
I	**Fixed** **Random**	Normal Normal	identity identity	log VLD× Week Patient	Gamma Gamma	log log	
II	**Fixed** **Random**	Poisson	log	log VLD× Week	Gamma	log	
III	**Fixed** **Random**	Poisson Normal	log identity	log VLD× Week Patient	Gamma Gamma	log log	
IV	**Fixed** **Random**	Poisson Gamma	log log	log VLD× Week Patient	Gamma Gamma	log log	
V	**Fixed** **Random**	Poisson	log	log VLD	Gamma	log	
VI	**Fixed** **Random**	Poisson Normal	log identity	log VLD+ Week Patient	Gamma Gamma	log log	
VII	**Fixed** **Random**	Poisson Gamma	log log	log VLD+ Week Patient	Gamma Gamma	log log	

* VLD refers to the viral load measurement.

Table 4.7. (a) Parameter estimates of the models without overdispersion for CD4 count data

Term*	Model I		Model II		Model III		Model IV	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
Fixed								
Constant	5.410	0.292	6.232	0.037	5.074	0.104	5.252	0.096
X	0.004	0.028	-0.065	0.004	0.039	0.006	0.039	0.005
Week 24	0.853	0.355	0.607	0.057	0.065	0.065	0.065	0.065
Week 36	0.786	0.376	0.659	0.064	0.239	0.074	0.238	0.074
Week 48	1.009	0.362	0.125	0.059	0.278	0.068	0.278	0.068
X × Week 24	-0.099	0.035	-0.067	0.006	-0.017	0.007	-0.017	0.007
X × Week 36	-0.095	0.036	-0.073	0.006	-0.042	0.007	-0.042	0.007
X × Week 48	-0.130	0.035	-0.027	0.006	-0.050	0.007	-0.050	0.007
Random								
Patient	0.386	0.077			0.442	0.084	0.368	0.070
Dispersion	0.136	0.014	1.000		1.000		1.000	
Deviance	165.6		16947		3615.0		3615.0	
Degrees of freedom	165.6		216		161.2		161.2	

*X refers to the logarithm of viral load measurement

Table 4.7. (b) Parameter estimates from the models with overdispersion for the CD4 count data

Term*	Model V		Model VI		Model VII	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
Fixed						
Constant	6.554	0.190	5.611	0.199	5.717	0.193
X	-0.105	0.019	-0.014	0.019	-0.011	0.019
Week 24			-0.079	0.057	-0.080	0.056
Week 36			-0.140	0.060	-0.143	0.060
Week 48			-0.184	0.060	-0.187	0.060
Random						
Patient			0.333	0.067	0.305	0.060
Dispersion	78.000		3.143	0.109	3.131	0.109
Deviance	17142		168.8		168.0	
Degrees of freedom	219		168.8		168.0	

*X refers to the logarithm of viral load measurement

Table 4.8 Goodness of fit tests for random effects estimates for each model

Model*	H_0	χ^2 statistic	Degree of freedom
I	$u \sim$ Normal	18.258	5
III	$u \sim$ Log - normal	15.307	5
IV	$u \sim$ Gamma	20.209	5
VI (with overdispersion)	$u \sim$ Log - normal	5.991	3
VII (with overdsipersion)	$u \sim$ Gamma	7.213	4

* Model II and V not shown because they did not involve random effects.

Figure 4.4. (a). Normal probability plots of deviance residuals for models without overdispersion

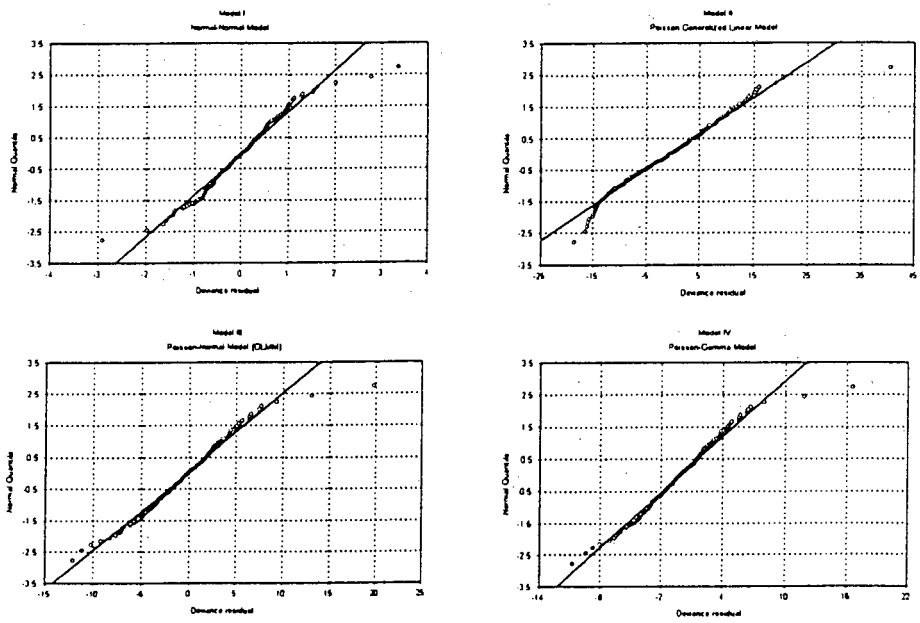
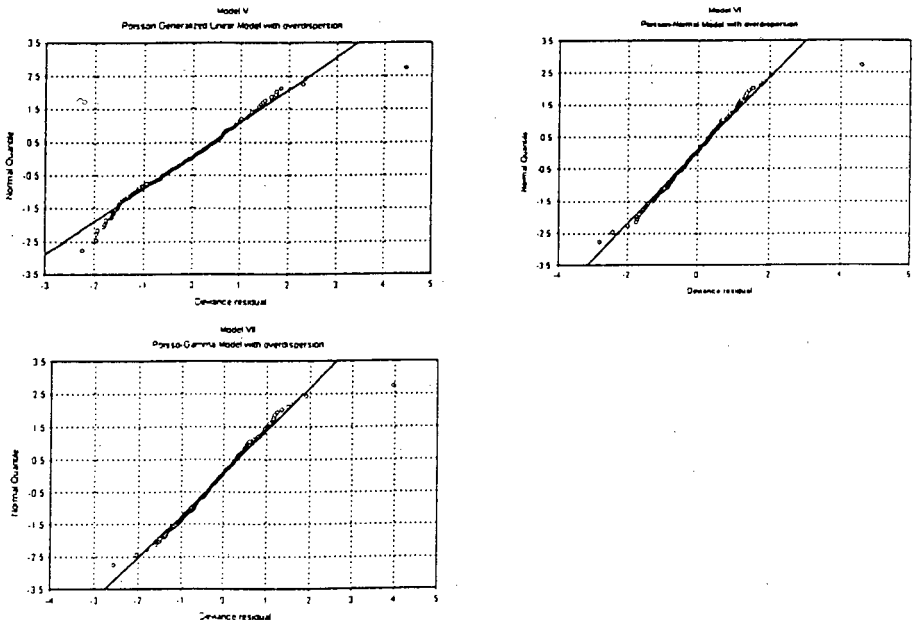
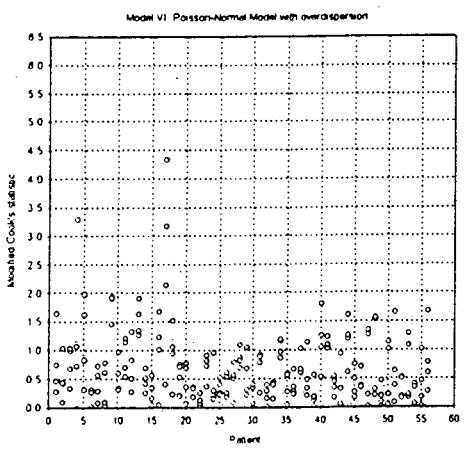


Figure 4.4. (b). Normal probability plots of deviance residuals for models with overdispersion



Note: The observation of patient 17 taken at the 48th week was not well fitted by all the models. The CD4 count of this patient was 993 with a corresponding viral load of 121000.

Figure 4.5. Index plots of Modified Cook's statistics
Model VI and Model VII



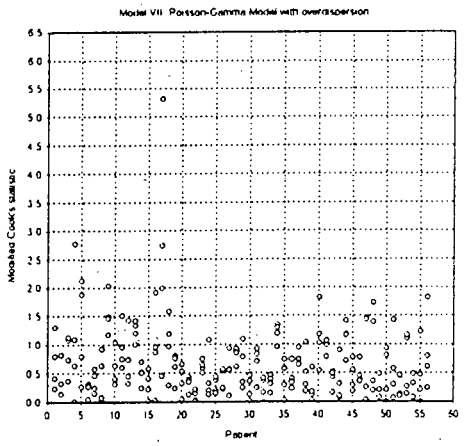


Figure 4.6. (a). Distribution of Random effects for models without overdispersion

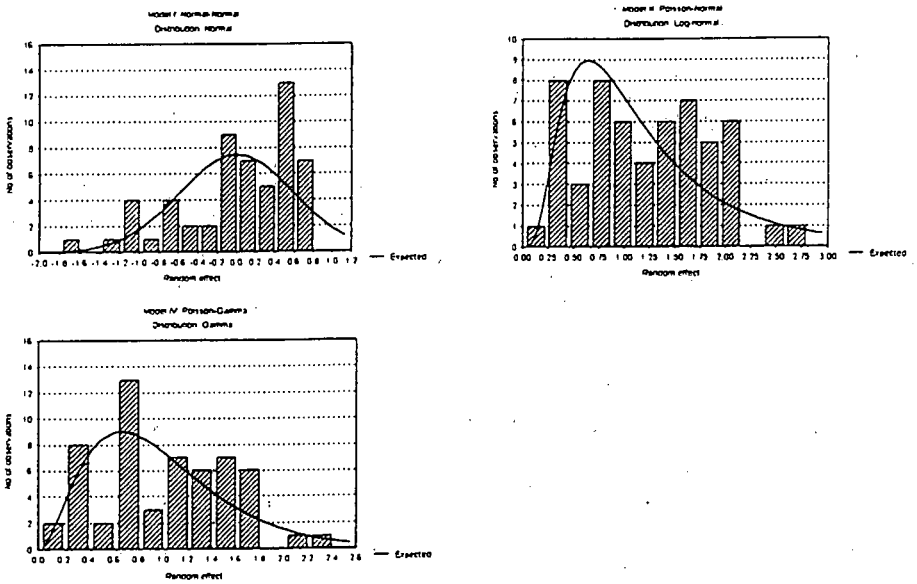
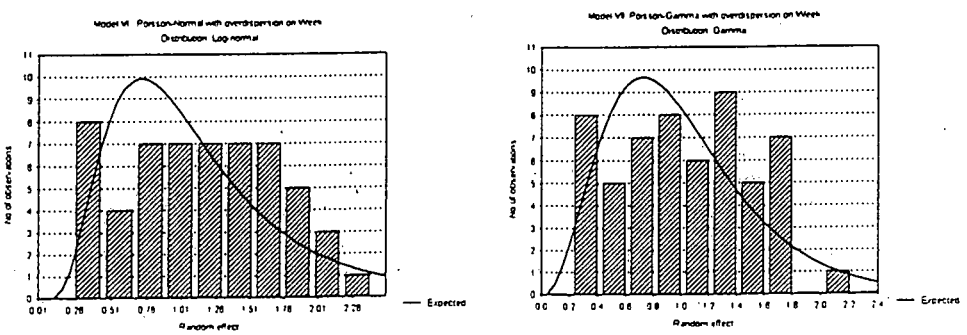


Figure 4.6. (b). Distribution of Random effects for the models with overdispersion



4.3.2.2 A Bayesian approach to modelling the relationship between the CD4 count and Viral load

We employed the Gibbs sampling technique in modelling the relationship between the CD4 count and viral load in the Bayesian framework. The Gibbs sampler is a statistical technique for simulating samples from the joint posterior distribution of the unknown quantities in a statistical model. The basic idea behind the method is that it generates random variables from a marginal distribution indirectly, without having to calculate the density. Casella and George, (1992) gave an exposition of the fundamental ideas behind the Gibbs sampler.

Gibbs sampling can be implemented using the BUGS software (Gilks *et al.*, 1994), a generic program that carries out Bayesian inference. The software can handle an extensive class of models that have used the Gibbs sampler, including cluster analysis (Gilks *et al.*, 1989), survival analysis models (Clayton, 1991), econometric models (Blattberg and George, 1990) and random effects generalized linear models (Dellaportas and Smith, 1993; Zeger and Karim, 1991).

The BUGS software requires that prior distributions for the parameters of interest be specified. It is hoped that these priors would have minimal influence on the results, that is, they are non-informative in the Bayesian sense. Bayesian inference then proceeds by calculating posterior distributions of the parameters of interest by taking the specified full joint probability distribution and conditioning on the observed data. Essentially, the analysis is one of simulation. The method for conducting this simulations are known as Markov Chain Monte Carlo methods (Gilks *et al.*, 1996), because the simulated values of the unknown parameters of the model follow a Markov chain whose stationary distribution is the required posterior distribution. Inferences concerning the unknown parameters in the statistical model are then based on summaries of the sampled values. Bayesian inference using BUGS have been discussed by several authors (Best and Spiegelhalter, 1996; Spiegelhalter, 1998).

Model fitting using BUGS

In this Section, we considered the Poisson GLM, Normal-Normal model, Poisson-Normal model and the Poisson-Gamma model; which were denoted as Model I, II, III and IV, respectively, in Section 4.3.2.1 of this chapter.

The fixed effects parameters assumed the same prior distribution in all the models, a normal prior with mean 0 and variance 10 000, which was intended to be proper but locally uniform. We used the dispersion parameter estimates obtained in the previous analysis (Section 4.3.2.1) to set-up prior distributions for the dispersion parameters.

The BUGS software parameterises the normal distribution in terms of the mean and precision $\tau = 1/\sigma^2$ and the gamma distributions in terms of its shape and scale parameters α and β. We gave prior distributions for τ in Model I and III and priors α and β were only specified for Model IV, since it was the only model considered with gamma distributed random effects. George *et al.*, (1993) considered a conjugate Poisson-Gamma hierarchical model and assumed the random effects followed a gamma distribution with parameter α and β. They assumed the prior distribution for

α was exponential while β was given a gamma prior. In contrast, Lee and Nelder, (1996) in their HGLMs assume the random effects follow a gamma distribution with mean 1 with the variance of the random effect being the reciprocal of the shape parameter α. We used the latter parametrization to set-up a prior distribution for the random effects in model IV.

The BUGS software also requires initial values to run iterations of the Gibbs sampler. The regression coefficients for the fixed effects were all given initial values of 0 while the precision parameters assumed the value 1.

The results were based on 10 000 iterations of the Gibbs sampler, after having 1000 iterations to set-up the monitoring of sampled values. The models were reasonably well behaved and convergence appeared to be rapid. The results for the models I, II, III and IV (Table 4.9 (a)) were consistent with the results obtained earlier in Section 4.3.2.1 of this Chapter. Figure 4.7 shows kernel density plots of the sampled values for the fixed effects parameters for Model I and indicates little evidence of departure from the prior assumptions for the interactions terms in the model.

We also investigated overdispersion in the Bayesian models. In BUGS, overdispersion in the distribution of the response variable is handled by adding a random effect for the units, u_{ij}. Breslow and Clayton, (1993), Lee and Nelder, (1996) used this parametrization in their re-analysis of data on epileptics (Thall and Vail, 1990). The models that assumed overdispersion were denoted as model V, VI and VII, as in the previous analysis (Section 4.3.2.1). The results from fitting these models are presented in Table 4.9 (b). For Poisson GLM, it was noted that the relationship between the CD4 count and log viral load depended on time. All the terms in models VI and VII were not statistically significant. This finding might reflect the weak relationship between the CD4 count and log viral load, however, it is not known to what extent is it due to inappropriate parametrizations for these models.

The Bayesian approach to inference in random effect models could be useful when there are few observations available to estimate the parameters of the statistical model. We mention below some of the limitations of using the BUGS software to fit generalized linear models with random effects.

Limitations of the Bayesian approach to fitting random effects models

i) The assumed prior distributions, required in the Bayesian approach could be influential in the results and this may lead to incorrect inferences.

ii) The Gibbs sampler relies on convergence of a simulation. However, there appears to be no method of checking whether the simulations are from the desired distribution.

iii) Parameterization chosen for a given model can be influential in the efficiency and behaviour of the simulation.

iv) While procedures for checking the convergence of the Gibbs sampler are available, procedures for checking model adequacy are yet to be developed.

Table 4.9. (a) Bayesian estimates from the models without overdispersion for CD4 count data

Term	Model I		Model II		Model III		Model IV	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
Fixed								
Constant	5.401	0.285	6.229	0.036	4.977	0.114	3.909	0.252
X	0.006	0.028	-0.065	0.004	0.041	0.005	0.029	0.006
Week 24	0.846	0.330	0.607	0.054	0.089	0.062	0.084	0.060
Week 36	0.767	0.353	0.661	0.062	0.263	0.069	0.268	0.079
Week 48	1.025	0.351	0.131	0.056	0.296	0.067	0.266	0.063
X × Week 24	-0.098	0.032	-0.067	0.006	-0.019	0.006	-0.018	0.006
X × Week 36	-0.093	0.034	-0.073	0.006	-0.044	0.007	-0.044	0.008
X × Week 48	-0.132	0.034	-0.028	0.006	-0.052	0.007	-0.048	0.006
Random								
Patient	0.402	0.090			0.442	0.084	0.368	0.070
Dispersion	0.137	0.016						

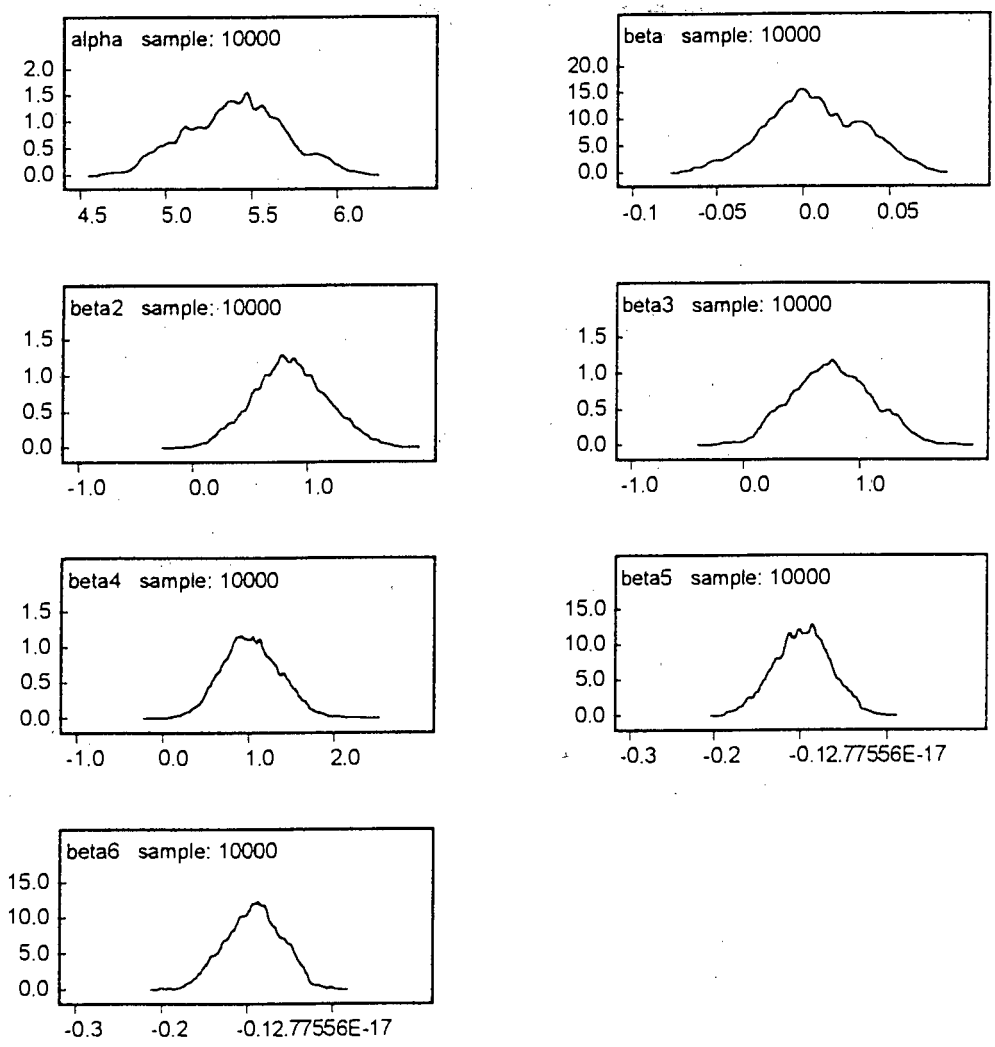
*X refers to the logarithm of viral load measurement

Table 4.9. (b) Bayesian estimates from the models with overdispersion for the CD4 count data

Term	Model V		Model VI		Model VII	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
Fixed						
Constant	4.602	0.461	3.209	1.735	3.098	1.404
X	0.090	0.044	0.121	0.103	0.128	0.076
Week 24	2.907	0.374	0.751	1.477	0.734	1.161
Week 36	1.985	0.662	0.588	1.378	0.675	1.339
Week 48	2.213	0.387	0.487	1.412	0.581	1.230
Random						
Patient			0.442	0.084	0.368	0.070
Dispersion	0.516	0.073	0.957	1.405	1.268	0.995

*X refers to the logarithm of viral load measurement

Figure 4.7. Kernel density plots of the sampled values for the regression coefficients
Model I: Normal-Normal Model



Notes:

alpha = constant term
beta = coefficient for log TLC
beta2 = coefficient for week 24
beta3 = coeffcient for week 36
beta4 = coefficient for week 48
beta5 = coefficient for week 24 and log TLC interaction
beta6 = coefficient for week 36 and log TLC interaction
beta7 = coefficient for week 48 and log TLC interaction

CHAPTER 5

Conclusions

5.1. The relationship between the CD4 count and TLC

The analysis revealed that there was a weak relationship between the CD4 count and the TLC. The relationship was not strong enough to allow for satisfactory prediction of the CD4 count from the TLC. As result of the weak relationship between the CD4 count and TLC, predictions of the CD4 counts from the TLC were poor. Variation in the CD4 counts was largely due to patient variation; with patient variation being 1.358 times random variation.

The analysis suggested different thresholds (cut-offs) of the TLC, for each HIV stage, for determining commencement of treatment. The TLC cut-offs were higher in advanced stages of the disease. This was due to high values of the TLC of some patients with correspondingly low CD4 counts, in advanced stages of the HIV disease. There is need for further examination of this group of patients in studying the relationship between the CD4 count and TLC.

The inclusion of HIV patients, who had previous exposure to antiretroviral drugs in the study instead of naïve HIV patients, may have biased the results because of the potent effect the drugs could have on the CD4 counts and TLC measurements. In general, however, it may be difficult to find patients who have never been exposed to antiretroviral drugs.

5.2. The relationship between the CD4 count and viral load

The relationship between the CD4 count and viral load was found to be weak. Variation in the CD4 counts was mostly due to patient variation. Patient variation was found to be 2.183 times the variation due to random variation. In the models considered in this study, the relationship between the CD4 count and viral load seemed to depend on time. However further modelling of this relationship (Chapter 4, fitting a Poisson GLM with overdispersion, Model V) revealed that the time dependence result depended on model choice.

The results from the analysis were in agreement with those from other similar studies conducted in developed countries. This implies that the relationship between the CD4 count and total lymphocyte count was the same in both the African population and HIV infected individuals in developed countries. This indicated that the pathogenesis of HIV was the same in both African patients and those in developed countries.

In this study we analysed data that came from HIV infected individuals who were given antiretroviral drugs. This could have influenced their CD4 counts and viral load measurements. There is a need therefore for prospective studies that will follow naïve HIV patients and measure their CD4 count and viral load measurements and other important covariate information such as the HIV stage or the duration of HIV infection. This data could then be used to study

the relationship between the CD4 count and viral load. Furthermore, the longitudinal data could also be used to describe the natural course of viral load throughout HIV infection in African patients.

5.3. Statistical methods

For the relationship between the CD4 count and TLC, we investigated models that assume normality and Poisson assumptions about the CD4 counts. Since the CD4 count is a proportion of the TLC, another approach to the analysis would be to use logistic regression to model this proportion. We did not investigate this approach in this study.

In this study, HGLMs provided reasonable models for the CD4 counts, in particular the Poisson-Gamma model. The latter model accommodated outlying observations. This was because a skewed distribution, the gamma distribution, was assumed for the random effects, u.

HGLMs adopt a parametric approach to random effects modelling and therefore provide a natural framework for model checking, while model checking procedures for Bayesian Markov Chain Monte Carlo (MCMC) methods are yet to be developed. However, a few caveats are in order. Firstly, HGLMs rely on the profile likelihood, which can underpropagate uncertainty about inference parameters in small samples. Secondly, they are limited to conjugate families, while Bayesian MCMC methods do not have this restriction. Thirdly, in HGLMs, parameter estimation is achieved through maximization of the $h-$likelihood which avoids the integration needed to calculate the usual marginal likelihood. With the advent of modern MCMC methods for high dimensional integrals the latter motivation may be weaker. Given the competing computational features and inferential properties between HGLMs and fully Bayesian methods, it would seem easier to use HGLMs for problems that may be too large to handle with Bayesian methods and to use Bayesian methods for data from small samples.

Statistical properties of diagnostics for generalized linear models are well established (McCullagh and Nelder, 1989), while regression diagnostics for HGLMs have not been well studied. There is therefore as need for further research into construction and statistical properties of diagnostics for HGLMs. In this study, we only assessed the distributional properties of the random effects, u.

There is a need in generalized linear models with random effects literature to distinguish between dispersion parameters and random effects. The random effects could be considered to represent known sources of variation while the dispersion represent unknown sources of variation.

References

Alcabes, P., Schoenbaum, E.E. and Klein, R.S. (1993). Correlates of the rate of decline of CD4+ lymphocytes among injection drug users infected with the Human Immunodeficiency Virus. *American Journal of Epidemiology*, **137**, 989-1000.

Altman, D.G. (1994). *Practical Statistics for Medical Research*. Chapman & Hall, London.

Anderson, D.A. and Aitkin, M. (1985). Variance component models with binary responses: Interviewer variability. *Journal of the Royal Statistical Society, Series B*, **47**, 203-210.

Anderson, R.M. and Darby, S.C. (1998). The role of statistics in human immunodeficiency virus reasearch. *Journal of the Royal Statistical Society, Series A*, **161**, 161-166.

Barlett, J.G. and Moore, R.D. (1998). Improving HIV Therapy. *Scientific American*, **279**,64-67.

Bachetti, P., Moss, A.R., Andrews, J.C. and Jacobson, M.A., (1992). Early predictors of survival in symptomatic HIV infected persons treated with high-dose zidovudine. *Journal of Acquired Immune Deficiency Syndromes*, **5**, 732-736.

Beckman, R.J., Nachttsheim, C.J. and Cook, R.D. (1987). Diagnostics for mixed-model analysis of variance. *Technometrics*, **29**, 413-426.

Berman, S.M. and Dubin, N. (1992). Is earlier better for AZT therapy in HIV infection? A mathematical model. In: Jewell, N.P., Dietz, K. and Farewell, V.T. (eds)., *AIDS Epidemiology: Methodological Issues*, Birkhauser, Boston.

Best, N. and Spiegelhalter, D.J. (1996). Modelling complexity using BUGS. In: Forcina, A., Marchetti, G.M., Hatzinger, R. and Galmacci, G. (eds). *Statistical Modelling*, Proceedings of the 11th International workshop on Statistical Modelling, 13-22.

Blattberg, R. and George, E.I. (1991). Shrinkage estimation of price and promotional elasticities:seemingly unrelated equations. *Journal of the American Statistical Association*, **86**, 304-315.

Breslow, N.E. (1984). Extra-Binomial Variation in Log-Linear Models. *Applied Statistics*, **33**, 38-44.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9-25.

Casella, G. and Berger, R.L. (1990). *Statistical Inference*. Wadsworth and Brooks Cole, Pacific Grove, California.

Casella, G. and George, E.I. (1992). Explaining the Gibbs Sampler. *The American Statistician*, **46**, 167-174.

Christensen, R., Pearson, L.M., and Johnson, W. (1992). Case-deletion diagnostics for mixed models. *Technometrics*, **34**, 38-45.

Clayton, D.G. (1991). A Monte Carlo method for Bayesian inference in frailty models, *Biometrics*, **47**, 467-485.

Coates, R.A., Farewell, V.T., Raboud, J., Read, S.E., Klein, M., MacFadden, D..K., Calzavara, L.M., Johnson, J.K., Fanning, M.M. and Shepherd, F.A. (1992). Using serial observations to identify predictors of progression to AIDS in the Toronto Sexual Contact Study. *Journal of Clinical Epidemiology*, **45**, 76-83.

Crowder, M.J. (1985). Gaussian estimation for correlated binary data, *Journal of the Royal Statistical Society, Series A*, **47**, 229-237.

De Cock, K.M., Soro, B., Coulibaly, I.M., Lucas, S.B. (1992). Tuberculosisand HIV infection in sub-Saharan Africa. *Journal of the American Medical Association*, **268**, 1581-1587.

DeGruttola, V., Lange, N. and Dafni, U. (1991). Modelling the Progression of HIV Infection. *Journal of the American Statistical Association*, **86**, 569-577.

DeGruttola, V. and Tu Ming, Xin (1992). Modelling the Relationship Between Progression of CD4-Lymphocyte Count and Survival Time. In: Jewell, N.P., Dietz, K. and Farewell, V.T. (eds)., *AIDS Epidemiology: Methodological Issues*, Birkhauser, Boston.

Dellaportas, P. and Smith, A.F.M. (1990). Bayesian inference for generalized linear models. Sampling based methods. *Technical Report BU-1138-M*, Cornell University, Biometrics Unit.

Diggle, P., Liang, K-Y. and Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Clarendon Press, Oxford.

Draper, N. and Smith, H. (1981). *Applied Regression Analysis*. 2nd edition, John Wiley & Sons, New York.

Engel, B. and Keen, A. (1994). A simple approach for the analysis of generalized linear mixed models, *Statistica Neerlandica*, **48**, 1-22.

Faucett, C.L. and Thomas, D.C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine*, **15**, 1663-1685.

Gelfand, A.E. and Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **90**, 97-985.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian Restoration images. *IEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721-741.

GENSTAT 5 Committee. (1993). Genstat 5 Release 3 Reference Manual. Oxford University Press, Oxford, United Kingdom.

George, E.I., Makov, U.E. and Smith, A.F.M. (1993). Conjugate likelihood distributions. *Scandanavian Journal of Statistics*, **20**, 147-156.

Gilks, W.R., Clayton, D.G., Spiegelhalter, D.J., Best, N.G., McNeil, A.J., Sharples, L.D. and Kirby, A.J. (1993). Modelling complexity: application of Gibbs sampling in medicine, *Journal of the Royal Statistical Society, Series B*, **58**, 619-678.

Gilks, W.R., Oldfield, L. and Rutherford, A. (1989). Statistical Analysis. *In Leukocycte Typing IV: While Cell Differentiation antigens*, Knapp, W., Dorken, B., Gilks, W.R., Rieber, E.P, Schmidt, R.E., Stein, H. and Dorken, B. (eds), 6-12, Oxford University Press, Oxford.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J.(eds) (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall: New York.

Gilks, W.R., Thomas, A. and Spiegelhalter, D.J. (1994). A language and program for complex Bayesian modelling. *The Statistician*, **43**, 169-177.

Gilmour, A.R., Anderson, R.D. and Rae, A.L. (1985). The analysis of binomial data by a generalized linear mixed model. *Biometrika*, **72**, 593-599.

Gilmour, A.R., Thompson, R. and Cullis, B.R. (1995). Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. *Biometrics*, **51**, 1440-1450.

Graham, N.M.H., Piantadosi, S., Park, L.P., Phair, J.P., Rinaldo, C.R. and Fahey, J.L. (1993). CD4+ lymphocyte response to zidovudine as a predictor of AIDS-free time and survival time. *Journal of Acquired Immune Deficiency Syndromes*, **6**, 1258-1266.

Hand, D. and Crowder, M. (1996). *Practical Longitudinal Data Analysis*. Chapman and Hall, London.

Harries, A.D. (1990). Tuberculosis and human immunodeficiency virus infection in developing countries. *Lancet*, **335**, 387-390.

Harville, D.A. (1974). Bayesian iference for variance components using only error contrasts. *Biometrika*, **61**, 383-5.

Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, **72**, 320-40.

Harville, D.A. and Mee, R.W. (1984). A Mixed-Model procedure for analyzing ordered catergorical data. *Biometrics*, **40**, 393-408.

Henderson, C.R. (1950) Estimation of genetic parameters. *Annals of Mathematical Statistics*, **21**, 309-310.

Henderson, C.R. (1973). Sire evaluation and genetic trends. *In Proceedings of the Animal Breeding and Genetics Symposium in Honour of Dr. Jay. Lush*, 10-41.

Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, **31**, 423-447.

Hoover, D.R., Graham, N.M.H., Chen, B., Taylor, J.M.G., Phair, J., Zhou, S.Y.J. and Munoz, A. (1992). Effect of CD4+ cell count measurement variability on staging HIV-1 infection. *Journal of Acquired Immune Deficiency Syndromes*, **5**, 794-802.

Hoover, D.R., Rinaldo, C., He, Y., Phair, J.P., Fahey, J.L. and Graham, N.M.H. (1995). Long-term survival without clinical AIDS after CD4+ cell counts fall below 200 x 10^6/l. *AIDS*, **9**, 145-52.

Hughes, M.D., Stein, D.S., Gundacker H.M., Valentine, F.T., Phair, J.P. and Volberding, P.A., (1994). Within-Subect variation in CD4 Lymphocyte Count in Asymptomatic Human Immunodeficiency Virus Infection: Implications for Patient Monitoring. *Journal of Infectious diseases*, **169**, 28-36.

Judge, G.G, Griffiths, W.E., Hill, R.C. and Lee, T-C. (1980). *The Theory and practice of econometrics*. J. Wiley & Sons, New York.

Khuri, A.I. and Sahai, H. (1985). Variance components analysis: A selective literature. *International Statistical Review*, **53**, 279-300.

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

Lange, N., Carlin, B. and Gelfand, A.E. (1989). Hierarchical Bayes models for progression of HIV infection using longitudinal CD4 T-cell numbers, *Journal of Acquired Immune Deficiency Syndromes*, **2**, 63-69.

Lange, N., Carlin, B. and Gelfand, A.E. (1992). Hierarchical Bayes models for progression of HIV infection using longitudinal CD4 T-cell numbers (with discussion), *Journal of the American Statistical Association*, **87**, 615-626.

Lee, Y. and Nelder, J.A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, Series B*, **58**, 619-678.

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

Liang, K-Y. and Zeger, S.L. and Qaqish, B. (1992). Multivariate regression analysis for categorical data (with discussion). *Journal of the Royal Statistical Society, Series B*, **54**, 3-40.

Lindsey, J.K. (1994). *Parametric Statistical Inference*. Clarendon Press, Oxford.

Lindstrom, M.J., and Bates, D.M. (1990). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated measures data. *Journal of the American Statistical Association*, **84**, 1014-22.

Longford, N.T. (1993). *Random Coefficient Models*. Clarendon Press, Oxford.

Maartens, G., Wood, R., O'Keefe, E. and Byrne C. (1997). Independent Epidemics of Heterosexual and Homosexual Infection in South Africa-Survival Differences. *Quarterly Journal Medicine*, **90**, 449-454.

Malone, J.L., Simms, T.E., Gray, G.C., Wagner, K.F., Burge, J.R. and Burke D.S. (1990). Sources of variability in repeated T-helper lymphocyte counts from human immunodeficiency virus type 1-infected patients: total lymphocyte count fluctuations and diurnal cycles are important. *Journal of Acquired Immune Deficiency Syndromes*, **3**, 144-51.

Mann, J.M., and Tarantola, D.J.M. (1998). HIV 1998: The Global Picture. *Scientific American*. **279**, 62-63.

McGilchrist, C.A. (1994). Estimation in generalized mixed models. *Journal of the Royal Statistical Society, Series B*, **56**, 61-69.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. 2nd edn. Chapman and Hall, London.

Mellors, J.W. (1998). Viral load test provide valuable answers. *Scientific American*, **279**,70-70.

Mellors, J.W., Rinaldo, Jr., C.R., Gupta, P., White, R.M., Todd, J.A. and Kinsgley, L.A. (1996). Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. *Science*, **272**, 1167-1170.

Montaner, J.S.G., Le, T.N. Le, N., Craib, K.J.P. and Schechter, M.T. (1992). Application of the World Health Organization system for HIV infection in a cohort of homosexual men in developing a prognostically meaningful staging system. *AIDS*, **6**, 719-24.

Moss, A.R. and Bachetti, P. (1989). Natural history of HIV infection. *AIDS*, **3**, 55-61.

Moss, A.R., Bachetti, P. and Osmond, D. (1988). Seropositivity for HIV and the development of AIDS and AIDS related condition: three year follow up of the San Fransisco general hospital cohort. *British Medical Journal*, **296**, 745-750.

Munoz, A., Cary, V., Saah, A.J., Phair, J.P., Kingsley, L.A., Fahey, J.L.,Ginzburg, H.M. and Polk, B.F. (1988). Predictors of decline in CD4 lymphocytes in a cohort of homsexual men infected with human immunodeficiency virus. *Journal of Acquired Immune Deficiency Syndromes*, **1**, 396-404.

Nelder, J.A. (1993). The K system for GLMs in Genstat. *Technical Report TRI/93*. Numerical Algorithms Group, Oxford.

Nelder, J.A., and Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika*, **74**, 221-231.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, **135**, 370-384.

Oman, S.D. (1995). Checking the assumptions in mixed-model analysis of variance: a residual analysis approach. *Computational Statistics and Data Analysis*, **20**, 309-330.

Pantaleo, G., Menzo, S., Vaccareza, M., Graziosi, C., Cohen, O.J., Demarest, J.F., Montefiori, D., Orenstein, J.M., Fox, C., Schrager, L.K., Margolick, J.B., Buchbinder, S., Giorgi, J.V. and Fauci, A.S. (1995). Studies in subjects with long-term nonprogressive human immunodeficiency virus infection. *New England Journal of Medicine*, **332**, 209-216.

Patterson, H.D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545-54.

Patterson, H.D. and Thompson, R. (1974). Maximum likelihood estimation of components of variance. *Proceedings of the eighth International Biometrics Conference*, 197-209.

Post, F.A., Wood, R., and Maartens, G. (1996). CD4 and Total lymphocyte Counts as predictors of HIV Disease Progression. *Quartely Journal of Medicine*, **89**, 505-508.

Post, F.A., Wood, R., and Pillay G.P. (1995). Pulmonary tuberculosis in HIV infection: radiographic appearance is related to CD4+ T-lymphocyte count. *Tubercle and Lung disease*, **76**, 518-521.

Raboud, J.M., Coates, R.A., and Farewell V.T. (1993) Estimating Risks of Progression to AIDS when covariates are measured with error. *Journal of the Royal Statistical Society, Series A*, **156**, 393-406.

Robinson D.L. (1987). Estimation and the use of variance components. *The Statistician*, **36**, 3-14.

Robinson, G.K. (1991). That BLUP Is a Good Thing: The Estimation of Random Effects. *Statistical Science*, **6**, 15-51.

Sabin, C.A. (1995). The follow-up of a cohort of anti-HIV seropositive haemophiliacs for up to 15 years from seroconversion. *PhD thesis*, Royal Free Hospital School of Medicine, Univerisity of London, London.

Sabin, C.A., Mocroft, A., Lepri, A.C., and Phillips, A.N. (1998). Cofactors of Markers of desease progression in human immunodeficiency virus infection. *Journal of the Royal Statistical Society Series A*, **161**,2, 177-189.

Saskela, K., Stevens, C.E., Rubinstein, P., Taylor, P.E. and Baltimore, D. (1995). HIV-1 messenger RNA in pripheral blood mononuclear cells as an early marker of risk for progression to AIDS. *Annals of Internal Medicine*, **123**, 641-648.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, **78**, 719-27.

Searle, S.R., Casella, G. and McCulloch, C.E. (1992). *Variance Components*. Wiley, New York.

Segal, M.R., James, J.R., French, M.A., and Mallal, S.A. (1994). Statistical issues in the evaluation of markers of HIV progression. *International Statistical Review*, 63, 179-197.

Self, S. and Pawitan, Y. (1992). Modelling a Marker of Disease Progression and Onset of Disease. In: Jewell, N.P., Dietz, K. and Farewell, V.T. (eds)., *AIDS Epidemiology: Methodological Issues*, Birkhauser, Boston.

Sloan, P., Carlin, J. and Crowe, S.M. (1991). Total lymphocyte count (TLC) can predict low CD4: a useful substitute for T subset analysis. *VII International conference on AIDS.* Florence.

Soriano, V., Castilla, J., Gomez-Cano, M., Holguin, A., Villalba, N., Mas, A. and Gonzalez-Lahoz, J. (1998). The Decline in CD4+ T lymphocytes as a Function of the Duration of HIV Infection, Age at Seroconversion, and Viral Load. *Journal of Infection*, **36**, 307-311.

Spiegelhalter, D.J. (1998). Bayesian graphical modelling: a case-study in monitoring health outcomes, *Applied Statistics*, **47**, 115-133.

Spiegelhalter, D.J., Thomas, Best, N.G. and Gilks, W.M. (1995). *BUGS Manual: version 0.50.* Medical Research Council Biostatistics Unit, Cambridge.

Stiratelli, R., Laird, N. and Ware, J.H. (1984). Random-effects Models for serial observations with binary responses. Biometrics, **40**, 961-971.

Swamy, P.A.V.B. (1971). *Statistical Inference in Random Coefficient Regression Models.* Springer-Verlag, New York.

Taylor, J.M.G., Fahey, J.L., Detels, R. and Giorgi J.V. (1989). CD4 percentage, CD4 number and CD4:CD8 ratio in HIV infection: which to choose and how to use. *Journal of Acquired Immune Deficiency Syndromes*, **2**, 114-24.

Taylor, J.M.G., Cumberland, W.G. and Sy, J.P. (1994). A stochastic model for analysis of longitudinal AIDS data. *Journal of the American Statistical Association*, **89**, 727-736.

Thall, P.F. and Vail, S.C., (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics*, **46**, 657-671.

Thompson, R. (1979). Estimation of variance and covariance components with an application when records are subject to culling. *Biometrics*, **29**, 527-50.

Tsiatis, A., DeGruttola, V., Strawderman, R. L., Dafni, U., Propert, K.L., Wulfsohn, M. (1992) The Relationship of CD4 Counts Over Time to Survival in Patients with AIDS: Is CD4 a Good Surrogate Marker? In: Jewell, N.P., Dietz, K. and Farewell, V.T. (eds)., *AIDS Epidemiology: Methodological Issues*, Birkhauser, Boston.

Tsutakawa, R.K. (1988). Mixed models for analyzing geographic variability in mortality rates. *Journal of the American Statistical Association*, **83**, 37-42.

Vittinghoff, E., Malami, H.M. and Jewell, N.P. (1994). Estimating patterns of CD4+ lymphocyte decline from a prevalent cohort of HIV-infected individuals. *Statistics in Medicine*, **113**, 1101-1118.

Waclawiw, M.A. and Liang, K-Y. (1993). Prediction of Random Effects in the Generalized Linear Model. *Journal of the Amearican Statistical Association*, 88, 171-178.

Waddington, D., Welham, S.J., Gilmour, A.R. and Thompson, R. (1994). Comparisons of some GLMM estimators for a simple binomial model, *Genstat Newsletter*, **30**, 13-24.

Wedderburn, R.W.M. (1974). Quasi-likelihood Functions, Generalized Linear Models, and the Gauss-Newton method. *Biometrika*, **61**, 439-447.

Wei, X., Ghosh, S.K., Taylor, M.E., Johnson, V.A., Emini, E.A., Deutsch, P., Lifson, J.D., Bonhoeffer, S., Nowak, M.A., Hahn, B.H., Saag, M.S. and Shaw, G.M. (1995). Viral dynamics in human immunodeficiency virus type 1 infection. *Nature*, **373**, 117-126.

Williams, D.A. (1982). Extra-Binomial Variation in Logistic Linear Models. *Applied Statistics*, **31**, 144-148.

Wood, R., (1999). *Personal communication*, Somerset Hospital HIV Clinic, Cape Town South Africa.

Wood, R., O'Keefe and Maartens, G. (1996). The changing pattern of transmission and clinical presentation of HIV infection in the Western Cape region of South Africa (1984-95). *South African Journal of Epidemiology and Infection*, **11**, 96-98.

WHO global programme on AIDS Proposed World Health Organization staging system for HIV infection and disease. (1993). *AIDS*, 7, 711-18.

Zeger, S.L. and Liang, K-Y. (1986). Longitudinal data Analysis for Discrete and Continuous Outcomes, *Biometrics*, **42**, 121-130.

Zeger, S.L., Liang, K-Y. and Albert, P.S. (1988). Models for Longitudinal Data: A Generalized Estimating Equation Approach. *Biometrics*, **44**, 1049-1060.

Zeger, S.L. and Karim M.R. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association*, **86**, 79-86.

APPENDICES

Appendix A Crosstabulations showing cell counts for calculating sensitivity, specificity, positive predictive value, negative predictive value, false negative rate and false positive rate

Table 2.13 Relation between estimated TLC values (cut-offs) and CD4 counts for HIV Stage I patients

	Observed CD4 count		
Estimated TLC value		> 200	**Total**
	29	38	67
>1159	40	441	481
Total	69	479	548

Table 2.14 Relation between estimated TLC values (cut-offs) and CD4 counts for HIV Stage II patients

	Observed CD4 count		
Estimated TLC value		> 200	**Total**
	41	40	81
>1240	44	311	355
Total	85	351	436

Table 2.15 Relation between estimated TLC values (cut-offs) and CD4 counts for HIV Stage III patients

	Observed CD4 count		
Estimated TLC value		> 200	**Total**
	278	112	390
>1615	75	242	317
Total	353	354	707

Table 2.13 Relation between estimated TLC values (cut-offs) and CD4 counts for HIV Stage IV patients

	Observed CD4 count		
Estimated TLC value		> 200	**Total**
	275	49	324
> 2281	12	26	38
Total	287	75	362

Table 2.17 Relation between a fixed TLC value (cut-off) of 1250 and CD4 counts for HIV Stage I patients

	Observed CD4 count		
Fixed TLC value		> 200	**Total**
	35	49	84
>1250	34	430	464
Total	69	479	548

Table 2.18 Relation between fixed TLC value (cut-off) of 1250 and CD4 counts for HIV Stage II patients

	Observed CD4 count		
Fixed TLC value		> 200	**Total**
	43	42	85
>1250	42	309	351
Total	85	351	436

Table 2.19 Relation between a fixed TLC value (cut-off) of 1250 and CD4 counts for HIV Stage III patients

	Observed CD4 count		
Fixed TLC value		> 200	**Total**
	223	50	273
>1250	130	304	434
Total	353	354	707

Table 2.20 Relation between a fixed TLC value (cut-off) of 1250 and CD4 counts for HIV Stage IV patients

	Observed CD4 count		
Fixed TLC value		> 200	**Total**
	208	13	221
>1250	79	62	141
Total	287	75	362

Appendix B GENSTAT programs for Random effects models for the CD4, TLC and Viral load

1. MODELS FOR THE RELATIONSHIP BETWEEN LOG CD4 COUNT AND LOG TOTAL LYMPHOCYTE COUNT FOR 376 HIV PATIENTS

```
FACTOR [LEVEL=376]PATNO
FACTOR [LEVEL=4]STAGE
FACTOR [LEVEL=2]TBSTATUS
FACTOR [LEVEL=16]VISIT

OPEN 'C:\\TTERM\\CD4\\CD4.PRN'; CHANNEL=3; FILETYPE=INPUT
READ[CH=3]PATNO,STAGE,CD4,TLC,VISIT,AGE,TBSTATUS,TIME

CALCULATE Y=LOG(CD4)
CALCULATE X=LOG(TLC)
CALCULATE X2=X*X

DESCRIBE [SELECTION=NOBS,NMV,MEAN,MEDIAN,MIN,MAX,Q1,Q3] Y,X

TABULATE [PRINT=MEAN;CLASSIFICATION=STAGE]CD4,TLC,AGE

"Model A: Main effects model"

VCOMPONENTS [FIXED=X+X2+STAGE+AGE+TBSTATUS+TIME;CADJUST=none;\
 CONSTANT=ESTIMATE]  RANDOM=PATNO; CONSTRAIN=POSITIVE
REML [PRINT=MODEL,COMPONENTS,WALD; PSE=DIFFERENCES; \
     MVINCLUDE=*; METHOD=FISHER]Y
VDISPLAY [PRINT=MODEL,COMPONENTS,EFFECTS,MEANS,STRATUMVAR,\
 VCOVARIANCE,DEVIANCE,WALD,MISSINGVALUES,MONITORING; PSE=DIFFERENCES]

VKEEP[RESIDUALS=RESID; FITTED=FIT]


"Model B: Checking for interactions between Main effects"

VCOMPONENTS [FIXED=X*STAGE;CADJUST=none;CONSTANT=ESTIMATE]
RANDOM=PATNO;\
      CONSTRAIN=POSITIVE
 REML [PRINT=MODEL,COMPONENTS,WALD; PSE=DIFFERENCES;\
      MVINCLUDE=*; METHOD=FISHER] Y
VDISPLAY [PRINT=MODEL,COMPONENTS,EFFECTS,MEANS,STRATUMVAR,\
 VCOVARIANCE,DEVIANCE,WALD,MISSINGVALUES,MONITORING; PSE=DIFFERENCES]

VKEEP[RESIDUALS=RESID; FITTED=FIT]


"Model C: Final model fitted to the CD4 data"

VCOMPONENTS [FIXED=X+STAGE;CADJUST=none;CONSTANT=ESTIMATE] \
    RANDOM=PATNO;CONSTRAIN=POSITIVE
REML [PRINT=MODEL,COMPONENTS,WALD; PSE=DIFFERENCES; \
     MVINCLUDE=*; METHOD=FISHER] Y
VDISPLAY [PRINT=MODEL,COMPONENTS,EFFECTS,MEANS,STRATUMVAR,\
 VCOVARIANCE,DEVIANCE,WALD,MISSINGVALUES,MONITORING; PSE=DIFFERENCES]

VKEEP[RESIDUALS=RESID; FITTED=FIT]
```

2. MODELS FOR THE RELATIONSHIP BETWEEN REPEATED MEASUREMENTS OF CD4 COUNT AND VIRAL LOAD FOR 56 HIV PATIENTS

```
FACTOR [LEVEL=56]PATNO
FACTOR [LEVEL=4] WEEK

OPEN 'C:\\TTERM\\CD4\\CD4VLD.PRN'; CHANNEL=3; FILETYPE=INPUT
READ[CH=3]IDNO,PATNO,CD4,VLOAD,WEEK

CALCULATE Y=LOG(CD4)
CALCULATE X=LOG(VLOAD)

DESCRIBE [SELECTION=NOBS,NMV,MEAN,MEDIAN,MIN,MAX,Q1,Q3] Y,X

TABULATE [PRINT=MEAN;CLASSIFICATION=CLINSTG]CD4,VLD,Y,X


"Model A: Main effects model"

VCOMPONENTS [FIXED=X+X2+STAGE+AGE+TBSTATUS+TIME;CADJUST=none;\
 CONSTANT=ESTIMATE]  RANDOM=PATNO; CONSTRAIN=POSITIVE
REML [PRINT=MODEL,COMPONENTS,WALD; PSE=DIFFERENCES; \
     MVINCLUDE=*; METHOD=FISHER]Y
VDISPLAY [PRINT=MODEL,COMPONENTS,EFFECTS,MEANS,STRATUMVAR,\
 VCOVARIANCE,DEVIANCE,WALD,MISSINGVALUES,MONITORING; PSE=DIFFERENCES]

VKEEP[RESIDUALS=RESID; FITTED=FIT]

"Model C: Final model fitted to the CD4 data"

VCOMPONENTS [FIXED=X+WEEK;CADJUST=none;CONSTANT=ESTIMATE] \
    RANDOM=PATNO;CONSTRAIN=POSITIVE
REML [PRINT=MODEL,COMPONENTS,WALD; PSE=DIFFERENCES; \
     MVINCLUDE=*; METHOD=FISHER] Y
VDISPLAY [PRINT=MODEL,COMPONENTS,EFFECTS,MEANS,STRATUMVAR,\
 VCOVARIANCE,DEVIANCE,WALD,MISSINGVALUES,MONITORING; PSE=DIFFERENCES]

VKEEP[RESIDUALS=RESID; FITTED=FIT]

"Model B: Checking for interactions between Main effects"

VCOMPONENTS [FIXED=X*WEEK;CADJUST=none;CONSTANT=ESTIMATE]
RANDOM=PATNO;\
      CONSTRAIN=POSITIVE
 REML [PRINT=MODEL,COMPONENTS,WALD; PSE=DIFFERENCES;\
      MVINCLUDE=*; METHOD=FISHER] Y
VDISPLAY [PRINT=MODEL,COMPONENTS,EFFECTS,MEANS,STRATUMVAR,\
 VCOVARIANCE,DEVIANCE,WALD,MISSINGVALUES,MONITORING; PSE=DIFFERENCES]

VKEEP[RESIDUALS=RESID; FITTED=FIT]


"Model C: Fitting AR(2) covariance structure"

VCOMPONENTS [FIXED=X*WEEK;CADJUST=none;CONSTANT=ESTIMATE]\
  RANDOM=PATNO.WEEK; CONSTRAIN=POSITIVE
VSTRUC [PATNO.WEEK] MODEL=AR; ORDER=2;FACTOR=WEEK
REML [PRINT=MODEL,COMPONENTS,WALD; PSE=DIFFERENCES; \
```

```
 MVINCLUDE=*; METHOD=FISHER]Y
VDISPLAY [PRINT=MODEL,COMPONENTS,EFFECTS,MEANS,
 VCOVARIANCE,DEVIANCE,WALD,MISSINGVALUES,MONITORING; PSE=DIFFERENCES]

VKEEP[RESIDUALS=RESID; FITTED=FIT]


"Model D: Fitting AR(1) covariance structure; Final model fitted to
the CD4 count and viral load data"

VCOMPONENTS [FIXED=X*WEEK;CADJUST=none;CONSTANT=ESTIMATE]\
  RANDOM=PATNO.WEEK; CONSTRAIN=POSITIVE
VSTRUC [PATNO.WEEK] MODEL=AR; ORDER=1;FACTOR=WEEK
REML [PRINT=MODEL,COMPONENTS,WALD; PSE=DIFFERENCES; \
 MVINCLUDE=*; METHOD=FISHER]Y
VDISPLAY [PRINT=MODEL,COMPONENTS,EFFECTS,MEANS,
 VCOVARIANCE,DEVIANCE,WALD,MISSINGVALUES,MONITORING; PSE=DIFFERENCES]

VKEEP[RESIDUALS=RESID; FITTED=FIT]
PRINT CD4, Y,X,FIT, RESID
```

Appendix C. GENSTAT programs for hierarchical generalized linear models for the relationship between CD4 count and TLC

```
Model I : Normal-Normal Model

FACTOR [LEVEL=376]PATNO
FACTOR [LEVEL=4]STAGE
FACTOR [LEVEL=2]TBSTATUS
FACTOR [LEVEL=16]VISIT

OPEN 'C:\\TTERM\\CD4\\CD4.PRN'; CHANNEL=3; FILETYPE=INPUT
READ[CH=3]PATNO,STAGE,CD4,TLC,VISIT,AGE,TBSTATUS,TIME
calculate y=log(CD4)
calculate x=log(TLC)

"Load K system"

k
kun 937
yvar y
err n
kfit x+CLINSTG

"Load HG system for fitting HGLMs"

load 'sys.hg'
openhg
vershg
suhg[n;i;x+CLINSTG;PATNO]n;i
fithg
dehg[m]
dehg[d]
aplhg
modhg
"Model diagnostics"
map[m]0
kmcs
print PATNO,CLINSTG,%rw,%cst,%rds,%rps,%ra,%csta
kmcd

"Print BLUPs"
map[m]1...nrc
kde

"Index plot of std. leverages"
klvp
"Plot of Log CD4 against Log TLC"
ksm y;x
```

Model II : Poisson Generalized Linear Model

```
FACTOR [LEVEL=376]PATNO
FACTOR [LEVEL=4]STAGE
FACTOR [LEVEL=2]TBSTATUS
FACTOR [LEVEL=16]VISIT

OPEN 'C:\\TTERM\\CD4\\CD4.PRN'; CHANNEL=3; FILETYPE=INPUT
READ[CH=3]PATNO,STAGE,CD4,TLC,VISIT,AGE,TBSTATUS,TIME
calculate y=log(CD4)
calculate x=log(TLC)

"Load K system"

k
kun 937
yvar CD4
err p
kfit x+CLINSTG
"Print parameter estimates"
kde

"Model diagnostics"
map[m]0
kmcs
print PATNO,CLINSTG,%rw,%cst,%rds,%rps,%ra,%csta
kmcd

"Index plot of std. leverages"
klvp
"Plot of Log CD4 against Log TLC"
ksm y;x
```

Model III : Poisson-Normal Model

```
FACTOR [LEVEL=376]PATNO
FACTOR [LEVEL=4]STAGE
FACTOR [LEVEL=2]TBSTATUS
FACTOR [LEVEL=16]VISIT

OPEN 'C:\\TTERM\\CD4\\CD4.PRN'; CHANNEL=3; FILETYPE=INPUT
READ[CH=3]PATNO,STAGE,CD4,TLC,VISIT,AGE,TBSTATUS,TIME
calculate y=log(CD4)
calculate x=log(TLC)

"Load K system"

k
kun 937
yvar CD4
err p
kfit x+CLINSTG


"Load HG system for fitting HGLMs"

load 'sys.hg'
openhg
vershg
suhg[p;l;x+CLINSTG;PATNO]n;i
fithg
dehg[m]
dehg[d]
aplhg
modhg
"Model diagnostics"
map[m]0
kmcs
print PATNO,CLINSTG,%rw,%cst,%rds,%rps,%ra,%csta
kmcd

"Print BLUPs"
map[m]1...nrc
kde
```

Model IV : Poisson-Gamma Model

```
FACTOR [LEVEL=376]PATNO
FACTOR [LEVEL=4]STAGE
FACTOR [LEVEL=2]TBSTATUS
FACTOR [LEVEL=16]VISIT

OPEN 'C:\\TTERM\\CD4\\CD4.PRN'; CHANNEL=3; FILETYPE=INPUT
READ[CH=3]PATNO,STAGE,CD4,TLC,VISIT,AGE,TBSTATUS,TIME
calculate y=log(CD4)
calculate x=log(TLC)

"Load K system"

k
kun 937
yvar CD4
err p
kfit x+CLINSTG


"Load HG system for fitting HGLMs"

load 'sys.hg'
openhg
vershg
suhg[p;1;x+CLINSTG;PATNO]
fithg
dehg[m]
dehg[d]
aplhg
modhg
"Model diagnostics"
map[m]0
kmcs
print PATNO,CLINSTG,%rw,%cst,%rds,%rps,%ra,%csta
kmcd

"Print BLUPs"
map[m]1...nrc
kde
```

Model VI : Poisson-Normal Model
(with constant overdispersion)

```
FACTOR [LEVEL=376]PATNO
FACTOR [LEVEL=4]STAGE
FACTOR [LEVEL=2]TBSTATUS
FACTOR [LEVEL=16]VISIT

OPEN 'C:\\TTERM\\CD4\\CD4.PRN'; CHANNEL=3; FILETYPE=INPUT
READ[CH=3]PATNO,STAGE,CD4,TLC,VISIT,AGE,TBSTATUS,TIME
calculate y=log(CD4)
calculate x=log(TLC)

"Load K system"

k
kun 937
yvar CD4
err p
kfit x+CLINSTG


"Load HG system for fitting HGLMs"

load 'sys.hg'
openhg
vershg
suhg[p;l;x+CLINSTG;PATNO;;y;e]n;i
fithg
dehg[m]
dehg[d]
aplhg
modhg
"Model diagnostics"
map[m]0
kmcs
print PATNO,CLINSTG,%rw,%cst,%rds,%rps,%ra,%csta
kmcd

"Print BLUPs"
map[m]1...nrc
kde
```

Model VII : Poisson-Gamma Model
(with constant overdispersion)

```
FACTOR [LEVEL=376]PATNO
FACTOR [LEVEL=4]STAGE
FACTOR [LEVEL=2]TBSTATUS
FACTOR [LEVEL=16]VISIT

OPEN 'C:\\TTERM\\CD4\\CD4.PRN'; CHANNEL=3; FILETYPE=INPUT
READ[CH=3]PATNO,STAGE,CD4,TLC,VISIT,AGE,TBSTATUS,TIME
calculate y=log(CD4)
calculate x=log(TLC)

"Load K system"

k
kun 937
yvar CD4
err p
kfit x+CLINSTG


"Load HG system for fitting HGLMs"

load 'sys.hg'
openhg
vershg
suhg[p;1;x+CLINSTG;PATNO;;y;e]
fithg
dehg[m]
dehg[d]
aplhg
modhg
"Model diagnostics"
map[m]0
kmcs
print PATNO,CLINSTG,%rw,%cst,%rds,%rps,%ra,%csta
kmcd

"Print BLUPs"
map[m]1...nrc
kde
```

Appendix D BUGS programs of Bayes hierarchical models for the relationship between the CD4 count and Viral load

Model I: Normal-Normal Model

```
model
        {for(j in 1 : N) {      # loop over patients
                        for(k in 1 : T) {  # loop over weeks
                                mu[j, k] <- alpha + beta* (X[j,k]) + beta2*(t2[k])+beta3*(t3[k])
                                        + beta4*(t4[k])
                                        + beta5*(t2[k])*(X[j,k])+beta6*(t3[k]) *(X[j,k])
                                        + beta7*(t4[k])*(X[j,k])
                                + b1[j]         # regression model (linear predictor)

                                y1[j,k] ~ dnorm(mu[j,k], tau.b)  # distribution of the response
                                y1[j,k]<- log(y[j,k])                # log (CD4 count)
                                X[j,k]<- log(v[j,k]) ;               # log(Viral load)

                                }

        b1[j]  ~ dnorm(0.0, tau.b1)      # subject random effects


                }

        ########################################
        #  prior distributions of model parameters          #
        ########################################

                alpha ~ dnorm(0.0,1.0E-4)
                beta ~ dnorm(0.0,1.0E-4)
                beta2  ~ dnorm(0.0,1.0E-4)
                beta3  ~ dnorm(0.0,1.0E-4)
                beta4  ~ dnorm(0.0,1.0E-4)
                beta5  ~ dnorm(0.0,1.0E-4)
                beta6  ~ dnorm(0.0,1.0E-4)
                beta7  ~ dnorm(0.0,1.0E-4);
                tau.b    ~ dgamma(1.0E-3,1.0E-3); sigm2.b <- 1.0 /(tau.b)
                tau.b1     ~ dgamma(1.0E-3,1.0E-3); sigm2.b1 <- 1.0 /(tau.b1)
}

# Reading the data:
# y = CD4 count , V= Viral load

Data  list(N = 56, T = 4,
        y = structure(.Data = c(357,357,395,486,
279,338,228,219,
242,276,114,126,
234,254,166,19,
351,72,198,132,
176,201,153,155,
82,94,60,52,
229,159,211,244,
40,23,19,177,
220,143,179,88,
612,367,334,449,
393,205,208,201,
114,12,19,12,
155,201,174,107,
255,252,205,271,
194,315,376,360,
386,281,427,993,
155,197,67,39,
391,428,281,409,
210,140,109,142,
298,337,316,275,
289,264,236,256,
58,41,89,38,
219,115,183,170,
```

70,71,68,53,
455,434,432,311,
246,348,269,310,
328,462,293,409,
290,421,328,266,
203,222,154,177,
87,63,86,139,
321,321,340,305,
276,319,255,285,
519,257,266,462,
282,343,346,283,
210,146,114,167,
319,414,277,408,
102,129,80,44,
139,136,88,108,
466,290,532,238,
117,30,32,24,
193,222,236,153,
90,101,97,46,
339,537,510,302,
394,397,261,326,
504,382,467,340,
400,516,377,215,
711,462,489,305,
390,394,308,321,
228,245,247,311,
232,185,185,75,
81,76,57,53,
280,223,315,139,
213,162,203,170,
595,407,405,343,
328,328,311,117), .Dim = c(56,4)),
(N= 56, T=4,
v =structure(.Data =c(2820,2400,2620,1510,
3690,1320,1700,930,
9650,21800,28300,127000,
694000,252000,372000,4230000,
8520,285000,523000,4080,
4950,11600,6380,75700,
6230,4770,8450,5280,
511000,21800,72200,47400,
1800000,3670000,9740000,259900,
94900,19000,15100,92000,
39000,990,25400,31100,
4550,1970,3530,29300,
1290000,4160000,2720000,3130000,
15500,211000,190000,366000,
46300,78300,52100,74800,
4900,5570,8130,677,
930,1200,39700,121000,
10300,45900,295000,30400,
5430,52600,17700,30300,
310,1990,6590,570,
14600,41300,38700,910,
470000,357000,516000,516000,
580,218000,57900,38900,
10400,18300,23500,52300,
22900,14200,21100,13800,
14000,1770,8110,5280,
112000,11900,48600,181800,
2030,1400,4570,15800,
110000,277000,147000,71800,
52000,32300,30000,15300,
580,14600,38500,86000,
16500,53500,38300,34600,
1740,4250,4710,2280,
25300,74200,55800,39500,
10900,6720,16700,46200,
22800,87700,40000,44300,
11500,10500,7290,4080,
102000,304000,452000,639000,
36100,147000,160000,13600,
2020,4010,11400,5280,
1660000,1390000,2310000,47400,
267000,18500,10400,31100,