# MODELING OF CD4 COUNTS CHANGE AND SURVIVAL TIMES OF HIV NAIVE PATIENTS ON ARVS

by

**CHARLES KIMANI MAINA**

**I/56/P/7094/2005**

A project report submitted to the School of Mathematics in partial fulfillment requirements for degree of Master of Science in Biometry, University of Nairobi

3rd August 2007

# DECLARATION

This dissertation is my own work carried out at the University of Nairobi during the 2007 academic year and has not been presented for award of any other degree.

Signature: .....

Charles Kimani Maina

This dissertation has been submitted for examination with the approval of:

Signature and date

Dr. Rosemary Nguti

School of Mathematics

University of Nairobi

UNIVERSITY OF NAIROBI
P.O. Box 30197
NAIROBI
14/8/2007
SCHOOL OF MATHEMATICS

.....

# DEDICATION

To

**My wife, Felistus Njoki and my parents who have been very instrumental through out this degree**

# ACKNOWLEDGMENTS

I would like to acknowledge the generous help of all the people who contributed to the effectiveness of this study:

Dr. Rosemary Nguti for guidance and support, assistance in data analysis and interpretation of results

Prof. Grace John Stewart and Dr. Michael Chung of the University of Washington and Nadia Kist of Coptic Hope Center for Infectious Diseases for providing me with the data used in this dissertation

The 400 dedicated participants, whose information has this dissertation successful

Finally I would like to express my gratitude to all my lecturers, classmates and the other School of Mathematics staff for all the support they provided during my study. They include Prof. Owino, Prof. Manene, Prof. Ottieno, Dr. Nguti, Dr. Gacii, Dr. Weke, Dr. Achia, Mr. Obudho, Mr. Mwangi (of KARI), my classmates Julie, Mutua, Ochiel, Ng'etich, Achien'g, Mweru, Thuranira and Waithaka.

# CONTENTS

# ABBREVIATIONS

| | |
|---|---|
| WHO | - World Health Organization |
| ART | - Anti-Retroviral Therapy |
| ART naïve | - Patients with the HIV virus and without previous antiretroviral therapy |
| HIV | - Human Immunodeficiency Virus, the virus that causes AIDS |
| AIDS | - Acquired Immunodeficiency Syndrome |
| PEPFAR | - Presidential Emergency Plan For AIDS Relief |
| USG | - United States Government |
| ARV | - Anti Retroviral |
| HAART | - Highly Active Anti Retroviral Therapy |
| NGO | - Non Governmental Organization |
| QOL | - Quality Of Life |
| OS | - Overall Survival |
| DOT | - Direct observation Therapy |
| TB | - Tuberculosis |
| LME | - Linear Mixed Effects |

# SUMMARY

This study was a longitudinal study on 400 ARV naive HIV positive patients started on ARVS. The study was aimed at modeling of cd4 counts change and survival times of these patients. 66.7% of the study participants were female, 62.8% had education above primary level with 43% earning below KSH 2,000. The mean baseline CD4 T cell count and weight is 134.1 and 62.5 kgs respectively.

Linear mixed model was used in the modeling of CD4 T cells change before and after start of ARVS. CD4 T cell at the time of enrollment was found to be the only covariate influencing change in CD4 T cell before start of ARVS, whereas time since start of ARVS, patient age and gender were the covariates influencing the CD4 count change after start of taking ARVS.

Classification (regression) trees were used to model TB incidence after start of ARVS. Patients' baseline weight was found to influence the incidence of TB. Hence the patients were classified to either having low weights (below 60 kgs) or normal weights (60kgs and above). Incidence was found to be greater for those with low baseline weight.

# Chapter 1

# INTRODUCTION

## 1.1 Background Information

By the end of 2005, it was estimated that, globally, 40.3 million adults and children (range, 36.7–45.3 million) were living with HIV infection, 25.8 million of whom were in sub-Saharan Africa. On World AIDS Day (1 December 2003), the World Health Organization/Joint United Nations Programme on HIV/AIDS (WHO/UNAIDS) initiative "3 by 5" was launched. Its ambitious objective was to deliver antiretroviral therapy (ART) to 3 million people in low- and middle-income countries, from a baseline of about 400,000 by the end of 2005. However, the effort fell short of its target. Treatment coverage in certain middle-income countries exceeded 80%, but in the poorer countries of Latin America, the Caribbean, Eastern Europe, and most of Asia and Africa, there is a long way to go. Similarly, by the end of 2005, only ~810,000 people in sub-Saharan Africa were receiving ART; 17% of those in need of therapy (i.e., those with WHO stage 4 disease or with CD4 cell counts <= 200 cells/ ml).

Almost two decades after the first AIDS case was described in Kenya, HIV/AIDS still remains the biggest social, economic and development challenge. This is in spite of the fact that the Government together with NGOs and development partners have continued to provide resources to stymie the epidemic and reducing its impact on the Kenyan society. UNAIDS estimates that 2 million of Kenya's 29.5 million people are

currently infected with HIV and that 1.5 million have succumbed to the disease, resulting in an overall decline in the life expectancy of 13 years. This trend is reflected in most sub-Saharan Africa.

Highly active antiretroviral therapy (HAART) has dramatically reduced morbidity and mortality among individuals with AIDS, yet until recently it has been out of reach for many HIV infected individuals living in resource-poor countries. The number of AIDS cases and deaths in Kenya has been rising since Antiretroviral (ART) drugs are only accessible to a very small proportion of Kenyans. The major implication is that people dying from the disease are in the age bracket 15-49 years – people of reproductive age and economically most productive. Falling antiretroviral prices and an influx of global financial support have increased antiretroviral access for these indigent populations. HAART has been proven to decrease HIV viral load, increase CD4 cell counts and stem the progression to AIDS and death in economically developed countries.

In evaluating the efficacy of new treatment for AIDS patient, survival and development of opportunistic infections (OI's) are generally the most important clinical endpoints. Measurement of CD4 cell count is a clinically relevant predictor of survival and of short-term overall risk of developing AIDS among HIV-infected patients. In addition, the initiation of primary prophylactic therapies for HIV-infected patients is based chiefly on the absolute CD4number. For example, a large cohort study demonstrated that those patients with CD4 counts below 200 cells / $mm^3$ had a significantly elevated risk of developing *Pneumocystis Carinii Pneumonia* (PCP) [Phair J et al.]. A study on the efficacy of Antiretroviral Therapy in 1004 patients with AIDS and without previous antiretroviral therapy in Haiti was done in 2003 showed that

Antiretroviral Therapy does improve the life of AIDS patients. During a 14-month period, three-drug antiretroviral therapy was initiated in 1004 patients, including children below 13 years of age. At enrollment, the median CD4 T-cell count in adults and adolescents was 131 per cubic millimeter (inter-quartile range, 55 to 221 per cubic millimeter); in children, a median of 13 percent of T cells were CD4-positive (inter-quartile range, 8 to 20 percent). According to a Kaplan Meier survival analysis, 87 percent of adults and adolescents and 98 percent of children were alive one year after beginning of treatment. In a subgroup of 100 adults and adolescent patients who were followed for 48 to 56 weeks, 76 patients had fewer than 400 copies of human immunodeficiency virus RNA per milliliter. In adults and adolescents, the median increase in the CD4 T cells from baseline to 12 months was 163 per cubic millimeter (inter-quartile range, 77 to 251 per cubic millimeter). In children the median percentage of CD4 T cells rose from 13 percent at baseline to 26 percent (inter-quartile range, 22 to 36 percent) at 12 month. Treatment-limited toxic effects occurred in 102 of the 910 adults and adolescents (11 percent) and 5 of the 94 children (5 percent).

We aim in this study to analyze the changes in the CD4 counts among a cohort of 400 ARV naïve patients who were started on ARVS and followed for about 1 year and model these changes. We also wish to study the weights of these patients over the same period and their survival times.

## 1.2 Motivating Study

We considered data from a cohort of 400 patients who are participating in an adherence clinical trial at the Coptic HIV clinic by Dr. Michael Chung of University of Washington. The clinical study was a randomized control trial, double-blinded block design, comparing three interventions plus a control group, with equal number of participants in each arm.

After enrollment and before HAART administration, study participants were randomized to one of four arms. Patients were randomized into either of the four arms A, B, C and D. Participants in arm A receive educational counseling sessions, those in B were given a pocket-size medication alarm device called the Med Reminder PC100 made by ALRT Technologies, those in C received both educational counseling and the Med Reminder PC100 alarm device, and those in arm D did not receive any adherence intervention strategies. Participants in arm D were the control group.

Enrollment to the study occurred within 1 year. Most interventions took 6 months to complete and participants will be followed for 1 ½ years after randomization. The first participant was enrolled in May 2006 whereas the last was enrolled in March 2007. This research paper is being written in the middle of the research. The participants are currently being followed and the last one will be followed till September 2008.

Men and women attending the Coptic HIV clinic were tested for HIV and examined for signs and symptoms associated with AIDS. If they had WHO Clinical Stage IV disease or a CD4 count $\leq 200$ then they were recommended to begin antiretroviral treatment and offered free HAART. Individuals, who agreed to HAART, were above 18 years of age,

were HAART treatment-naïve, and planned to live in Nairobi for at least two years, were eligible and invited to enroll in the proposed randomized clinical trial.

Prior to enrollment, all participants were asked to sign a written informed consent. At enrollment, a complete physical examination including their weight was performed and a standardized questionnaire was administered to assess sexual, medical, psychological, and nutritional history. Blood specimens were obtained for CD4 count, forming the baseline information. TB was also checked for by the clinician in every visit that the participant came to the clinic.

### 1.3 Objectives

Our overall objective is to study the improvement and or deterioration of patients Quality of Life (QOL) before and after start of ARVS.

To fulfill this overall objective, we will study the following specific objectives;

1. Model the rate of CD4 change before and after start of use of ARVS and the factors associated with rate of decline before and after use of ARVS
2. Modeling of the survival times of patients and times to event (death and first TB incidence)
3. Cluster patient characteristics and compare their TB status

# Chapter 2

## LITERATURE REVIEW

Often in clinical trials where the primary endpoint is time to an event, patients are also monitored longitudinary with respect to one or more biological endpoints throughout the follow-up period. This may be done by taking immunological or virological measures in the case of infectious diseases or perhaps with a questionnaire assessing the quality of life (QOL) after receiving a particular treatment. These longitudinal measures are often incomplete after a long follow-up period, and may be prone to measurement error. They are important because they may be predictive of survival. Therefore, methods which can model both the longitudinal and survival components are becoming increasingly essential in most AIDS clinical trials.

In the statistical literature, joint models for longitudinal and time-to-event data have been examined by several authors. Tsiatis, DeGruttola, and Wulfson (1995) proposed a two-stage procedure by plugging the estimates for modeling the longitudinal data into a Cox proportional hazards model. Yueh-Yun and Ibrahim (2006) proposed a joint model which was primarily motivated by a clinical trial conducted by the International Breast Cancer Group (IBCSG). They directed a large clinical trial, IBCSG Trial VI, in pre-menopausal women with node-positive breast cancer to study both the duration of adjuvant chemotherapy and reintroduction of delayed chemotherapy. In the

study, in addition to the adjuvant treatment effects, patients' QOL was also hypothesized to carry prognostic information and to be predictive of breast cancer progression. Four indicators of QOL were assessed repeatedly over time with a self-assessment QOL questionnaire, which was subject to measurement error and missing information. In the trial, cancer progression was monitored in terms of two failure time random variables, disease-free survival (DFS) and overall survival (OS). Yueh-Yun and Ibrahim (2006) proposed a novel multivariate survival model to capture these features in the data set. Specifically they proposed a joint likelihood approach to jointly model multi-dimensional QOL and bivariate failure time random variables DFS and OS. They assumed that, given the latent QOL trajectory functions, QOL and bivariate failure times (DFS, OS) are conditionally independent. For the longitudinal component of the joint model, a multivariate mixed effects model was presented to explicitly capture different sources of dependence among repeated multidimensional QOL measures.

For the survival component of the joint, they introduced a shared frailty, which was assumed to have a positive stable distribution, to induce correlation between failure times. Given the common frailty, all survival responses of interest were assumed to be independent. The marginal univariate survival model, which accommodates both zero and nonzero cure fractions for the time to event, was then applied to each marginal survival fractions.

In this study, we propose to model CD4 change using the linear mixed effects model, for a group of patients started on ARVS to explicitly capture different sources of dependence among repeated multidimensional QOL measures, which include age, gender, and baseline weight among other.

## 2.1 Problem Justification

Close to 1 million people may have started receiving ART in sub-Saharan Africa, but it has been estimated that there were 3.2 million (range, 2.8–3.9 million) new infections in 2005. Evaluation of programs and more operational and fundamental research are needed to find the best models that can integrate HIV prevention and ART delivery. Candidate models will span the spectrum of care facilities, from regional and district hospitals, to the various levels of health centers, and finally to the community and to the home. *As usual, there will be no "one size fits all" solution.*

Lawn et al. studied mortality in a treatment-naive cohort that received ART over 3 years in a community health center in Cape Town, South Africa. A number of studies, including their own, have demonstrated that the majority of deaths that occur in the first year or so of treatment take place in the first few months after initiation of ART. This simply reflects the advanced state of immunodeficiency of most patients when therapy is started. However, Lawn and colleagues have now examined the "queue"—that is, the period between identifying a patient as needing treatment and enrolling him or her in the treatment program and actually initiating drug therapy.

They observed mortality rates of 33.3, 19.1, and 2.9 deaths/100 person-years in the pretreatment interval, the first 4 months of ART (early treatment), and after 4 months (late deaths), respectively. Pretreatment and early treatment deaths accounted for 87% of mortality. Late mortality was low, and patient retention within the cohort was very good. The median duration of the pretreatment period was 34 days (inter-quartile range, 28–50 days); thus, the time spent in this period by some patients was substantial. As with

mortality in the early treatment phase, there was a strong association with advanced immunosuppression.

This study proposes to study the mortality in a treatment naïve cohort of 400 HIV positive patients receiving ART for about one (1) year at the Coptic Hope Center, Nairobi. Our overall objective is to study the improvement/ deterioration of patients Quality of Life (QOL). We propose to study these changes in QOL by modeling changes in CD4 at baseline and at time $t$ among HIV patients, study their survival rates by modeling their survival functions using Kaplan Meier survival graphs and Cox proportional hazard models for the patients. We also propose to study the odds of a patient getting a TB infection within a year after starting ARVS. Lastly we will analyze the baseline profiles of the study cohort by age group, gender, education, marital status and other socio demographic and economic factors.

# Chapter 3

# RESEARCH DESIGN AND METHODOLOGY

## Introduction

In this study we propose to accomplish the objectives in a number of ways. Comparative summary tables will be provided to summarize the patient characteristics. Linear mixed models will be used to Model the rate of CD4 change before and after start of use of ARVS and the factors associated with rate of decline before and after use of ARVS. Both Kaplan Meier and Cox survival models will be used to model the time to event (death and first TB infection after start of ARVS) of the patients; and regression trees will be used in the clustering of patient characteristics to compare their TB incidence.

Each of these methods has been described below.

### 3.1. Data Summaries and Trellis Plots

Trellis plots are a powerful tool to display relationships for large datasets compactly on one page while distinguishing between different groups, i.e. to look at aspects of the data structure within each **panel**. With a trellis plot you can see how the relationship between two variables changes with variations in one or more conditioning variables. Trellis plots can be a quick decision making tool at the exploratory stage for

the approach to take when analyzing *repeated measurements*, i.e. measurements taken on the same experimental unit over time.

Means, medians and inter-quartile ranges will be used to describe patients' baseline characteristics whereas trellis plots will be used to check on the relationship between CD4 change and time after start of ARVS.

Charts and plots will be used to show the change in CD4 cell count after start of ARVS.

### 3.2. Linear Mixed Models

The normal linear model,

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi} + \varepsilon_i$$
$$\varepsilon_i \sim NID(0, \sigma_\varepsilon^2 I_n)$$

has one random effect, the error term $\varepsilon_i$. The parameters of the model are the regression coefficients, $\beta_1, \beta_2, ..., \beta_p$, and the error variance, $\sigma^2$. Usually, $x_{1i} = 1$, and so $\beta_1$ is a constant or intercept.

For comparison with the linear mixed model of the next section, we rewrite the linear model in matrix form,

$$y = X\beta + \varepsilon$$

$$\varepsilon \sim N_n(0, \sigma^2 I_n)$$

where $y = (y_1, y_2, ..., y_n)'$ is the response vector; $X$ is the model matrix, with typical row $x_i' = x_{1i}, x_{2i}, ..., x_{pi}$; $\beta = (\beta_1, \beta_2, ..., \beta_p)'$ is the vector of regression coefficients; $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)'$ is the vector of errors; $N_n$ represents the n-variable multivariate-normal distribution; 0 is an n x 1 vector of zeroes; and $I_n$ is the order- n identity matrix.

So-called mixed-effect models (or just mixed models) include additional random-effect terms, and are often appropriate for representing clustered, and therefore dependent, data – arising, for example, when data are collected hierarchically, when observations are taken on related individuals (such as siblings), or when data are gathered over time on the same individuals.

Linear mixed models may be expressed in different but equivalent forms. In the social and behavioral sciences, it is common to express such models in hierarchical form. The lme (linear mixed effects) function in the nlme library, however, employs the Laird-Ware form of the linear mixed model (after a seminal paper on the topic published by Laird and Ware, 1982):

$$y_{ij} = \beta_1 x_{1ij} + \ldots + \beta_p x_{pij} + b_{i1} z_{1ij} + \ldots + b_{iq} z_{qij} + \varepsilon_{ij}$$
$$b_{ik} \sim N(0, \psi_k^2 + Cov(b_k, b_{k'}) = \psi_{kk'}$$
$$\varepsilon_{ij} \sim N(0, \sigma_2 \lambda_{ij}), Cov(\varepsilon_{ij}, \varepsilon_{ij'}) = \sigma^2 \lambda_{ijj'}$$

Where

- $y_{ij}$ is the value of the response variable for the $j^{th}$ of $n_i$ observations in the $i^{th}$ of $M$ groups or clusters.
- $\beta_1, \ldots, \beta_p$ are the fixed-effect coefficients, which are identical for all groups.
- $x_{1ij}, \ldots, x_{pij}$ are the fixed-effect regressors for observation $j$ in group $i$; the first regressor is usually for the constant, $x_{1ij} = 1$.
- $b_{i1}, \ldots, b_{iq}$ are the random-effect coefficients for group $i$, assumed to be multivariate normally distributed. The random effects, therefore, vary by group. The $b_{ik}$ are thought of as random variables, not as parameters, and are similar in this respect to the errors $\varepsilon_{ij}$.
- $z_{1ij}, \ldots, z_{qij}$ are the random-effect regressors.

- $\psi_k^2$ are the variances and $\psi_{kk'}$ the covariances among the random effects, assumed to be constant across groups. In some applications, the $\psi$'s are parameterized in terms of a relatively small number of fundamental parameters.
- $\varepsilon_{ij}$ is the error for observation $j$ in group $i$. The errors for group $i$ are assumed to be multivariately normally distributed.
- $\sigma^2\lambda_{ijj'}$ are the covariances between errors in group $i$. Generally, the $\lambda_{ijj'}$ are parameterized in terms of a few basic parameters, and their specific form depends upon context. For example, when observations are sampled independently within groups and are assumed to have constant error variance (as in the application developed in the following section), $\lambda_{ijj} = \sigma^2$, $\lambda_{ijj'} = 0$ (for $j \neq j'$), and thus the only free parameter to estimate is the common error variance, $\sigma^2$. Similarly, if the observations in a "group" represent longitudinal data on a single individual, then the structure of the $\lambda's$ may be specified to capture autocorrelation among the errors, as is common in observations collected over time.

Alternatively but equivalently, in matrix form,

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i$$
$$b_i \sim N_q(0, \psi)$$
$$\varepsilon_i \sim N_m(0, \sigma^2\Lambda_i)$$

Where

- $y_i$ is the $n_i \times 1$ response vector for observations in the *ith* group.
- $X_i$ is the $n_i \times p$ model matrix for the fixed effects for observations in group $i$.
- $\beta$ is the $p \times 1$ vector of fixed-effect coefficients.
- $Z_i$ is the $n_i \times q$ model matrix for the random effects for observations in group $i$.

- $b_i$ is the $q \times 1$ vector of random-effect coefficients for group $i$.
- $\varepsilon_i$ is the $n_i \times 1$ vector of errors for observations in group $i$.
- $\psi$ is the $q \times q$ covariance matrix for the random effects.
- $\sigma^2\Lambda_i$ is the $n_i \times n_i$ covariance matrix for the errors in group $i$.

### 3.3. Time to Event (Survival)

In many medical studies an outcome of interest is the time to an event. Such events may be adverse, such as death or recurrence of a tumour. The distinguishing feature of survival data is that at the end of the follow up period, the event will probably not have occurred for all the patients. For these patients the survival time is said to be censored, indicating that the observation period was cut off before the event occurred. We do not know when (or, indeed, whether) the patient will experience the event, only that he or she has not done so by the end of the observation.

The Kaplan-Meier procedure is a method of estimating time-to-event models in the presence of censored cases. The Kaplan-Meier model is based on estimating conditional probabilities at each time point when an event occurs and taking the product limit of those probabilities to estimate the survival rate at each point in time.

**Example**: Does a new treatment for AIDS have any therapeutic benefit in extending life? You could conduct a study using two groups of AIDS patients, one receiving traditional therapy and the other receiving the experimental treatment. Constructing a Kaplan-Meier model from the data would allow you to compare overall survival rates between the two groups to determine whether the experimental treatment is an improvement over the traditional therapy. You can also plot the survival or hazard functions and compare them visually for more detailed information.

Cox Regression is a method for modeling time-to-event data in the presence of censored cases. However, Cox Regression allows you to include predictor variables (covariates) in your models. For example, you could construct a model of length of employment based on educational level and job category. Cox Regression will handle the censored cases correctly, and it will provide estimated coefficients for each of the covariates, allowing you to assess the impact of multiple covariates in the same model. You can also use Cox Regression to examine the effect of continuous covariates.

**Example**: Do men and women have different risks of developing lung cancer based on cigarette smoking? By constructing a Cox Regression model, with cigarette usage (cigarettes smoked per day) and gender entered as covariates, you can test hypotheses regarding the effects of gender and cigarette usage on time-to-onset for lung cancer.

We therefore obtain the Kaplan Meier estimate for the time to death of the patients started on ARVS. This estimate is given by,

$$S(t) = \prod_{t_j \le t} \left( \frac{n_j - d_j}{n_j} \right)$$

In the above equation, $n_j$ is the number of individuals at risk at $t_j^{th}$ infection time, $d_j$ is the number of deaths at this time.

For the Cox regression, we will model survival using the covariates of sex, age, baseline CD4, baseline weight, marital status, occupation, earnings and education of the patient using the following hazard model;

$$h_i(t) = \exp(\beta_1 sex_i + \beta_2 age_i + \beta_3 basecd4_i + \beta_4 basewt_i + \beta_5 marital_i + \beta_6 occupat_i + \beta_7 earnings_i + \beta_8 educ_i) h_0(t),$$

where the subscript $i$ on an explanatory variable denotes the value of that variable for the $i^{th}$ individual. The baseline hazard function is the function for an individual for whom the values of all the variables are zero. This function corresponds to a male child aged 0, who has zero values for baseline cd4 and baseline weight, separated, with no education and earning between 2000 and 5000.

### 3.3.1. The statistic -2log L

In order to compare the alternative models fitted to an observed set of survival data, a statistic which measures the extent to which the data are fitted by a particular model is required. -2log $L$ will always be positive, and for a given data set, the smaller the -2log L the better the model.

### 3.3.2. Model selection

An initial step in model selection process is to identify a set of explanatory variables that have the potential for being included in the linear component of a proportional hazards model. This set will contain those variates and factors which have been recorded for each individual, but additionally terms corresponding to the interactions between factors or between variates and factors may also be required.
Once a set of potential explanatory variables has been isolated, the combination of variables which are to be used in modeling the hazard function will not depend on a unique combination of variables. Instead, there are likely to be a number of equally good models, rather than a single 'best' model.

### 3.3.3. Variable selection procedure

We first consider the situation where all explanatory variables are on an equal footing, and the aim is to identify subset of variables upon which the hazard function

depends. Alternative nested models can be compared by examining the change in the value of -2log $L$ on adding terms in the model or deleting terms from a model. Comparison between a number of possible models, which need not necessarily be nested, can also be made on the basis of the statistic

$$AIC = -2\log L + AIC = -2\log \hat{L} + \alpha q,$$

in which q is the number of unknown $\beta$- parameters in the model and $\alpha$ is a pre-determined constant. The value of $\alpha$ is usually taken to be between 2 and 6. This statistic is known as *Akaike's information criterion*; the smaller the value of this statistic, the better the model.

We propose to model time to death using a Cox proportional hazard model for the $i^{th}$ individual as

$$h_i(t)=exp(\beta_1 Age_i + \beta_2 Sex_i + \beta_3 Arm_i + \beta_4 Weight_i + \beta_5 CD4_i + \beta_6 Education_i + \beta_7 WHO_i)h_0(t)$$

where the subscript $i$ on an explanatory variable denotes the value of the variable for the $i^{th}$ individual. The baseline hazard function $h_0(t)$, is the hazard function for an individual for whom the values of all six of these variables are zero.

We will also model the time to first TB infection after start of ARVS.

### 3.4. Classification tree

The Classification Tree procedure creates a tree-based classification model. It classifies cases into groups or predicts values of a dependent (target) variable based on values of independent (predictor) variables. The procedure provides validation tools for

exploratory and confirmatory classification analysis. The procedure can be used for Segmentation, in which it identifies persons who are likely to be members of a particular group; Stratification, in assigning cases into one of several categories, such as high-, medium-, and low-risk groups; Prediction, where it creates rules and uses them to predict future events, such as the likelihood that someone will default on a loan or the potential resale value of a vehicle or home; Data reduction and variable screening; Interaction identification.; Category merging and discretizing continuous variables; and Recoding group predictor categories and continuous variables with minimal loss of information.

The tree building process starts by partitioning a sample or the "root node" into binary nodes based upon a very simple question of the form: is $X \leq d$? where X is a variable in the data set, and d is a real number. Initially, all observations are placed at the root node. This node is impure or heterogenous since it contains observations of, say both food-secure and food-insecure localities. The goal is to devise a rule that will initially break up these observations and create groups or binary nodes that are internally more homogenous than the root node.

**Example**: A bank wants to categorize credit applicants according to whether or not they represent a reasonable credit risk. Based on various factors, including the known credit ratings of past customers, you can build a model to predict if future customers are likely to default on their loans.

A tree-based analysis provides some attractive features:

- It allows you to identify homogeneous groups with high or low risk.
- It makes it easy to construct rules for making predictions about individual cases.

TB and no TB were estimated for each category of the independent variables, socio-

demographic and clinical. Comparisons of all independent variables were made using regression tree (classification tree). The dependent variable was TB versus no-TB. The independent variables that were being considered were gender, age, baseline weight (BMI), baseline CD4, previous TB history, baseline WHO clinical staging and level of education.

# Chapter 4

# RESULTS

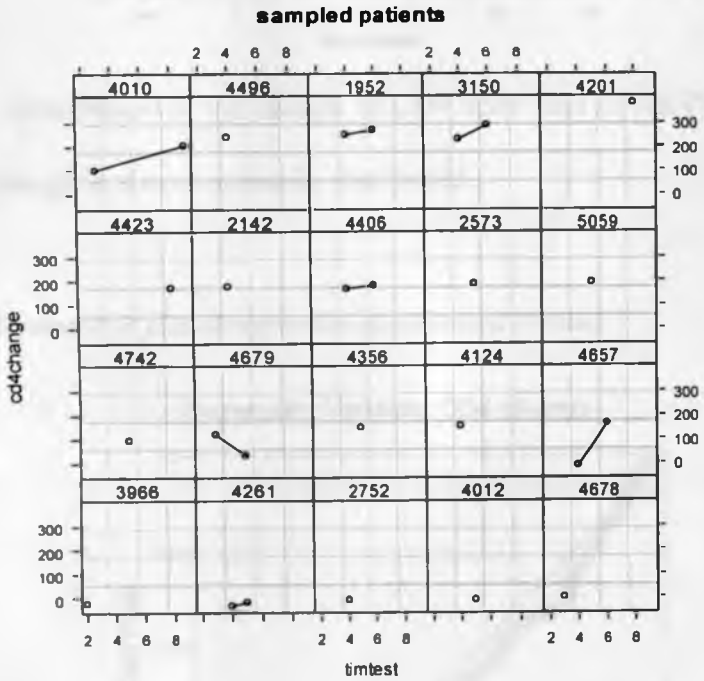
## Data Summaries

**Table 1**: *Characteristics of the patients (N = 400)*

| Characteristics | | |
|---|---|---|
| Gender | Male N (%) | 134 (33.5%) |
| | Female N (%) | 266 (66.5%) |
| Age (years) | Mean (inter-quartile range) | 36.33 (30-42) |
| Education level | No Education | 8 (2.0%) |
| | Lower primary education | 8 (2.0%) |
| | Five to eight years of primary education | 108 (27.0%) |
| | Beyond primary education | 251 (62.8%) |
| Earnings | < 2,000 | 172 (43.0%) |
| | 2,001 – 5000 | 57 (14.3%) |
| | 5,001 - 10,000 | 68 (17.0%) |
| | 10,001 - 20,000 | 48 (12.0%) |
| | 20,001 - 30,000 | 17 (4.3%) |
| | 30,001 - 50,000 | 9 (2.3%) |
| | > 50,001 | 4 (1.0%) |
| Marital status | Separated | 56 (14.0%) |
| | Married-monogamous | 156 (39.0%) |
| | Married-Polygamous | 20 (5.0%) |
| | Cohabiting | 2 (.5%) |
| | Divorced | 2 (.5%) |
| | Widowed | 65 (16.3%) |
| | Single | 74 (18.5%) |
| Baseline CD4 | Mean (inter-quartile range) | 134.10 (65.00 – 188.00) |
| | ≤ 200 | 281 (80.1%) |
| | > 200 | 70 (19.9%) |
| Occupation | Unemployed | 121 (30.3%) |
| | Employed | 112 (28.0%) |
| | Self-Employed | 93 (23.3%) |
| | Farmer | 8 (2.0%) |
| | Housewife | 10 (2.5%) |
| | Casual labourer | 31 (7.8%) |
| Baseline weight (kgs) | Mean (inter-quartile range) | 62.455 (53.0 – 70.0) |
| Time since start of ARVS (months) | Mean (range) | 4.57 (0 – 8) |

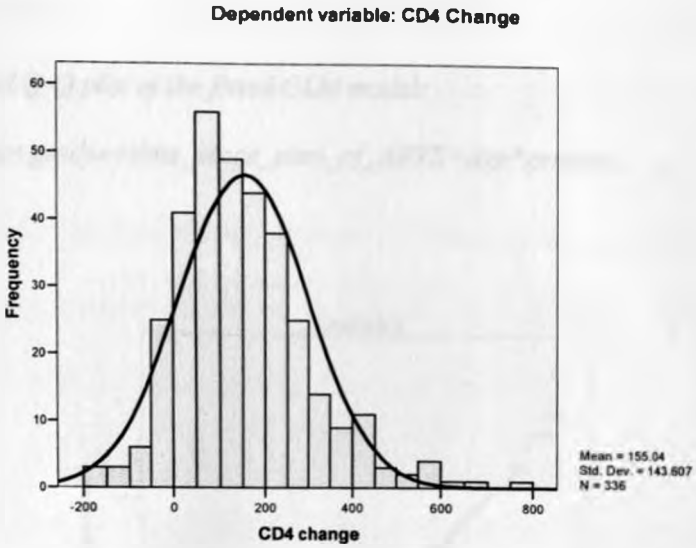
Among the 400 patients, 66.5% were female, 62.8% had a beyond primary education, 43% were earning less than KSH 2000 per month, 39% were in a monogamous marriage and 30.3% were unemployed. The mean baseline weight was 62.5 kgs, mean baseline CD4 count is 134. The mean time since start of ARVS is 4.57 months. A summary table of the baseline characteristics of the patients in the study is provided in table 1. 351 patients had a repeat CD4 after the start of the ARVS.

Figure 1.1: Trellis plot for CD4's for 20 sampled patients; CD4 change by time since start of ARVS for 20 randomly selected patients



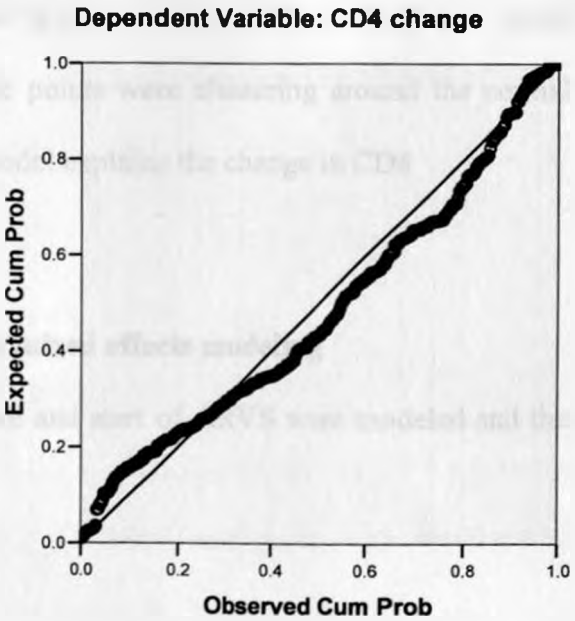

From the trellis chart, there is a positive change in CD4 counts over time for most of the patients where CD4 count change did take place; but most of the sampled patients had only one CD4 count taken after start of ARVS, hence only one CD4 change from the baseline, thus no strong within-subject patterns.

**Chart 1**: *histogram of the frequencies of change in CD4 T cells with a normal line superimposed*




From the distribution of the change in CD4 after start of ARVS in chart 1, it is evident that the change is almost normally distributed.

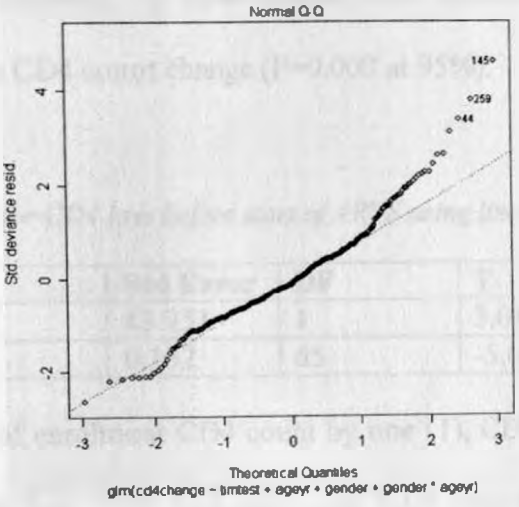
**Chart 2:** *Normal P-P plot of regression standardized residual*



From the Normal P-P chart 2 plotting the cumulative proportion of the change in CD4 cells against expected cumulative probability shows the sample is from a normal distribution.

**Chart 3**: *Normal Q-Q plot of the fitted GLM model:*

*(cd4change~age+gender+time_since_start_of_ARVS+age*gender)*




From the Normal Q-Q chart 3, the fitted GLM was plotted against standard deviance residual and the points were clustering around the normal straight line. This indicates that the fitted model explains the change in CD4

### 4.1 CD4 Change, Linear mixed effects modeling

CD4 change before and start of ARVS were modeled and the results of each are described below:

**Linear Mixed Effects**

### 4.1.1 before start of ARVS

The results for the fitted linear model for the CD4 count change before start of ARVS are given in *Table 2*. As stated above the factors considered were baseline age, CD4 count at enrolment, time between enrolment and start of ARVS, gender, and social economic factors (marital status and income). Of these factors, only enrolment CD4 count was significantly associated with CD4 count change (P=0.000 at 95%).

**Table 2**: *Parameter estimates for CD4 loss before start of ARVS using linear model*

| Variable | Coeff | Std Error | DF | T | p-value |
|---|---|---|---|---|---|
| (Intercept) | 132.557 | 43.951 | 1 | 3.016 | 0.004 |
| Enrolment CD4 | -0.859 | 0.152 | 65 | -5.664 | 0.000 |

For every increase of enrolment CD4 count by one (1), CD4 count loss reduced by 0.859. That means that those with high enrolment CD4 counts have a lower CD4 count loss, compared to those who have low CD4 count at enrolment.

The model for CD4 count loss before start of ARVS is:

$y_i = 132.557(\pm 43.951) - 0.859(\pm 0.152) x Enrolment\_CD4 + \varepsilon_i$ , where:

$y_i$ = CD4 count loss $\varepsilon_i$ = Random Error

### 4.1.2 LME after start of ARVS

The results for the fitted linear model for the CD4 count change before start of ARVS are given in *Table 3*. We considered gender, age, baseline CD4 count, time since beginning of ARVS to CD4 test-date (in months) and the social economic status. Only

time since start of ARVS, age and gender were significant in this model. But time for start of ARVS is the random effect since some patients had more than one CD4 count done on them after the start of ARVS. The full results of this analysis are given in *Table 3.*

| | **StdDev** | **Corr** |
|---|---|---|
| (Intercept) | 108.13143 | (Intr) |
| Time since start of ARVS | 10.15265 | -0.563 |
| Residual | 101.02324 | |

The value of $|\rho| = 0.563$ which shows the intra-class correlation coefficient is large enough for us to use the linear mixed effects model.

**Table 3**: *Parameter estimates for CD4 change after start of ARVS using linear mixed model*

| **Variable** | **Coeff** | **Std.Error** | **DF** | **t-value** | **p-value** |
|---|---|---|---|---|---|
| ***Fixed*** | | | | | |
| (Intercept) | -58.45677 | 74.43075 | 246 | -0.785385 | 0.4330 |
| Time since start of ARV | 18.55559 | 4.95564 | 85 | 3.744337 | 0.0003 |
| Patient age | 2.40284 | 1.81147 | 246 | 1.326457 | 0.1859na |
| Gender: Male | *Reference* | | | | |
| Female | 286.36908 | 80.66163 | 246 | 3.550251 | 0.0005 |
| Patient age: Gender Female | -6.50355 | 2.11374 | 246 | -3.076799 | 0.0023 |
| ***Random*** | | | | | |
| (Intercept) | | | | | |
| Time since start of ARV | | | | | |

From the results in Table 3, CD4 count increases by 18.6 for every month the patient is on ARVS. The interaction between gender and baseline age of a patient is significance, in that for every increase in age by a year for a female patient, CD4 reduces by 6.5 times.

A final reduced model that contains only estimable terms that contributed significantly to the response and corresponded to the best outcome of the test statistics was adopted.

## 4.2 Modeling of the Time to Event of patients

### 4.2.1 Survival curves

The Kaplan-Meier estimate of survival curves for the for the time-to-death for ARV naïve HIV patients started on ARVS is shown in Fig. 3 and Fig. 4. The first plot shows that survival is decreasing uniformly whereas the second plot shows that men were dying faster that females in the first 6 months, but thereafter time-to-death in the two groups was almost similar.

**Fig 3**: *Overall survival*

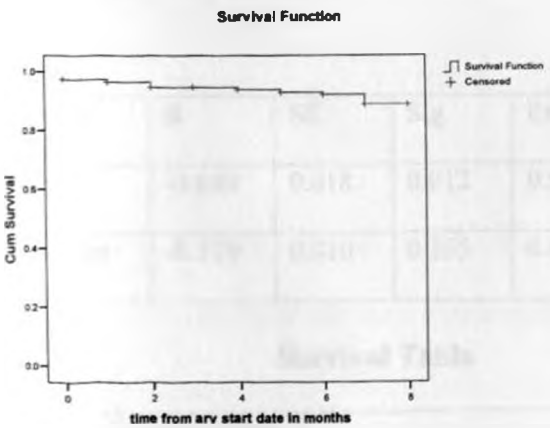
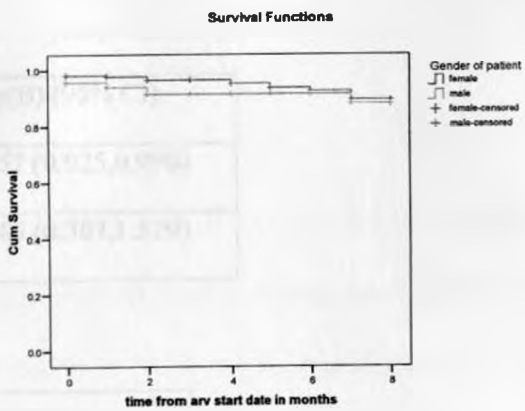



**Fig 4**: *Classified by gender*



### 4.2.2 Time to event using Cox regression

#### 4.2.2.1 Time to death

Nothing was found from the variables of interest (age, sex, previous TB incidence, socio-economic factors and previous health status), that would explain time to death. This is because the number of events (deaths) was too few less than 15, to explain survival. But since the study is ongoing and by end of the study, each of the patients will have had enough follow-up period, then Time to Death can be further investigated.

#### 4.2.2.2 Time to TB incidence after start of ARVS

The results for the fitted Cox PH model for the time-to-first TB incidence after start of ARVS, is given in table 5.

**Table 5**: *Parameter estimates for the time-to-first TB incidence using a Cox Proportional hazard model*

| Variables | B | SE | Sig. | Exp(B) (95% CI) |
|---|---|---|---|---|
| Weight | -0.044 | 0.018 | 0.012 | 0.957 (0.925,0.990) |
| TB in past | -0.379 | 0.410 | 0.355 | 0.685 (0.307,1.529) |

**Survival Table**

| Time | Baseline Cum Hazard | At mean of covariates | | |
|---|---|---|---|---|
| | | Survival | SE | Cum Hazard |
| 0 | .289 | .982 | .007 | .018 |
| 1 | .745 | .954 | .013 | .047 |
| 2 | 2.340 | .863 | .035 | .148 |
| 3 | 3.224 | .816 | .048 | .204 |
| 4 | 5.787 | .694 | .077 | .366 |
| 5 | 7.360 | .628 | .094 | .465 |
| 6 | 9.859 | .536 | .117 | .623 |

Patients' baseline weight is an important indicator of TB incidence after start of ARVS. This means that incidence of TB is reduced by 100%-(100%x0.957) = 4.3% for every kilogram of weight a patient has at baseline. The significance level of TB in the past in explaining Time to first TB incidence after start of ARVS is greater than 5% (sig. = 0.355), hence it does not explain the Time to TB after start of ARVS.

### 4.2.3 Regression (classification) Tree

**Figure 6**: *Classification of TB incidence by patient characteristics*

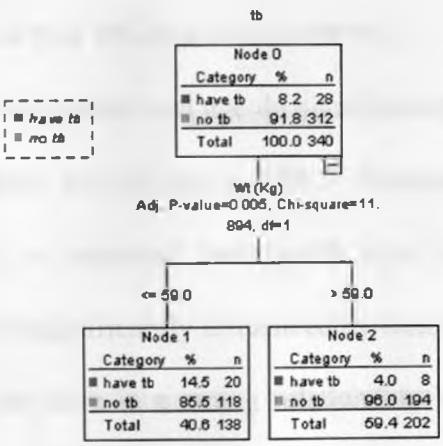



Figure 6 is an analysis of the data for incidence of TB among the patients after start of ARVS by their baseline characteristics. Patient with low baseline body weight had a significantly higher TB incidence compared to those whose baseline weight was high (above 59 KGs). Since body weight is an indicator of a persons' well being, it therefore means that patient who are more sick at baseline, hence the low body weight have a higher chance at 14.5% of developing TB compared to those who are stronger at 4%.

# Chapter 5

## DISCUSSIONS

HIV/AIDS pandemic has emerged as one of the leading challenges to global public health and development. Our overall objective was to study the improvement / deterioration of patients Quality of Life (QOL) after start of ARVS. To achieve this major objective, we sort to model the changes in CD4 T cell count before and after start of ARVS; to model the time to death and time to first TB incidence after start of ARVS; and to predict the incidence of first TB after start of ARVS.

This study revealed that gender and age do significantly influence the increase of CD4 T cells in the future, after start of taking ARVS. Besides this, we saw that once a patient start taking ARVS, as expected their health does improve over time, since increase in CD4 T cells is also significantly influenced by time since start of ARVS.

This study shows that there is a strong relationship between patients' baseline weight and TB incidence. Since patients' weight is an indicator of the patients' well being, it then means the weaker a patient is at the time of initiation to ARVS the higher the chance that they will develop TB in the first 6 months as compared to those who are not so weak. A previous study done [A. Wanchu et. al.], showed that as a patients weight and CD4 T cells was increased via taking of ARVS and treating of opportunistic infections (OI's), had a positive influence on the patient's quality of life. Increase in CD4 T cells are indicative of the likelihood of the development of OI's in the near future.

# Chapter 6

## CONCLUSSIONS

Studies of HIV dynamics in AIDS research are very important for understanding pathogenesis of HIV infection and for assessing the potency of antiviral therapies. Quality of life can be measured in a number of ways which include counting the number of CD4 T cell, weight and opportunistic infections. From the study we have seen that CD4 change is explained by the baseline patient characteristics (age, gender and occupation) at the time of starting ARVS, and the time since start of ARVS. Hence it is important to start following patients early enough before their immune system gets compromised.

TB is one of the killer opportunistic infections (OI) especially for HIV-positive clients, hence the management of this OI is important. Weight is associated with incidence of TB after start of ARVS. But weight is an indicator of the QOL of an individual and hence from the results, it can be seen the weaker a patient is, the higher the chance they have of developing TB within at least the first 6 (six) months of starting ARVS. Patients can then be classified using their weight and the weaker ones can be started on anti-TB prophylaxis.

From the Kaplan Meier graphs the survival times for HIV positive on ARVS men and female is not different over long time even though that of men is lower in the first 6 months.

In conclusion, as the study is still ongoing, a lot more will be discovered as the follow-up period increases.

# REFERENCES


1. National AIDS and STD Control Programme (NASCOP). Sentinel Surveillance of HIV and STDs in Kenya. Ministry of Health Kenya: 2005
2. UNAIDS. Report on the Global HIV/AIDS Epidemic: Fourth Global Report. Geneva: World Health Organization; 2004.
3. Steinbrook R. The AIDS epidemic in 2004. N Engl J Med 2004; 351: 115-117.
4. AVANTI 2. Randomized, double-blinded trial to evaluate the efficacy and safety of zidovudine plus lamivudine versus zidovudine plus lamivudine plus indinavir in HIV-infected antiretroviral-naïve patients. AIDS 2000; 14:367-374.
5. Phair J et al. The risk of Pneumocystis carinii pneumonia among men infected with HIV1. New England Journal of Medicine 1990; 155: 961-974
6. Patrice Severe, Paul Leger et al. Antiretroviral Therapy in a Thousand Patients with AIDS in Haiti. N Engl J Med 2005; 353: 2325-2334.
7. UNAIDS and WHO. AIDS epidemic update, December 2005. Available at: http://www.who.int/hiv/epi-update2005_en.pdf.
8. WHO and UNAIDS. Progress on global access to HIV antiretroviral therapy: a report on "3 by 5" and beyond: March 2006. Available at: http://www.who.int/hiv/progreport2006_en.pdf.
9. Lawn S, Myer L, Harling G, et al. Determinants of mortality and non death losses from an antiretroviral treatment service in South Africa: implications for program evaluation. Clinical Infectious Diseases 2006; 43:770–6.
10. Barbra A. Richardson et al. Hormonal contraception and HIV-1 disease progression among postpartum Kenyan women. Lippincott Williams & Wilkins 2007

11. A. Wanchu et. al. Prevention of opportunistic infections in HIV infection by pentoxiphylline. Indian J Med Res 124, December 2006, pp 705-708

## APPENDICES

### Appendix 1: LME results in R

```
> ck1$gender <- factor(ck1$gender,levels=c('1','2'))
> contrasts(ck1$gender)
       Female
Male        0
Female      1
```

Having established the contrast-coding for gender, the linear mixed model in equation 1 is fit as follows:

```
> y <-lme (cd4change ~ timtest + ageyr + gender + ageyr*gender, random = ~ timtest | hopeid, data=ck1)
> summary(y)
Linear mixed-effects model fit by REML
Data: ck1
     AIC      BIC    logLik
4226.505 4260.724 -2104.252

Random effects:
 Formula: ~timtest | hopeid
 Structure: General positive-definite, Log-Cholesky parametrization
             StdDev    Corr
(Intercept) 108.13143 (Intr)
timtest      10.15265 -0.563
Residual    101.02324

Fixed effects: cd4change ~ timtest + ageyr + gender + ageyr * gender
               Value Std.Error  DF   t-value p-value
(Intercept) -58.45677 74.43075 246 -0.785385  0.4330
timtest      18.55559  4.95564  85  3.744337  0.0003
ageyr         2.40284  1.81147 246  1.326457  0.1859
gender2     286.36908 80.66163 246  3.550251  0.0005
ageyr:gender2 -6.50355  2.11374 246 -3.076799  0.0023

Correlation:
              (Intr) timtst ageyr  gendr2
timtest       -0.340
ageyr         -0.928 0.021
gender2       -0.822 0.018  0.850
ageyr:gender2  0.795 -0.017 -0.857 -0.977

Standardized Within-Group Residuals:
       Min         Q1          Med         Q3        Max
-2.59239394 -0.48494535 -0.09501877 0.37317815 3.93629687
```