LONGITUDINAL STUDY OF CHANGE IN CD4+ CELL COUNTS ON HIV-POSITIVE PATIENTS INITIATED ON ANTIRETROVIRAL THERAPY AT THECOMPREHENSIVE CARE CENTRE IN KENYATTA NATIONAL HOSPITAL

GENERALIZED ESTIMATING EQUATIONS MODEL

AND

GENERALIZED LINEAR MIXED-EFFECTS MODEL

DR. VALERIA N.B MAKORY

W62/68997/2011

A PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT OF THE DEGREE OF MASTERS OF SCIENCE IN MEDICAL STATISTICS AT THE UNIVERSITY OF NAIROBI, INSTITUTE OF TROPICAL AND INFECTIOUS DISEASES (UNITID)

2014

DECLARATION

This research project is my original work and has not been presented to any other university/institution.

Sign--------------------------------------- Date: --------------------------------

DR Valeria N. B Makory

DECLARATION BY SUPERVISOR

This research project has been submitted for examination with your approval as University Supervisors.

Sign ------------------------------------ Date------------------------

DR Anne Wang'ombe
Senior Lecturer and coordinator, MSC [Medical Statistics] Programme
University of Nairobi

DEDICATION

I would like to dedicate this work to God and my beloved parents- Mr.& Mrs. Charles and Margaret Makory; my very dear loving husband Mr. Erick AbugaGecheo and my precious children LassanaGesare and Liam Nyakawa for the understanding, support and sacrifice in the course of preparation of this work.

ACKNOWLEDGEMENT

I thank God for the far He has brought me. He has proven that His promises are always true and His ways are beyond human understanding.

I am deeply indebted to my supervisor Dr. Anne Wang'ombe and Mr. Lawrence Muthami(KEMRI) for the academic and professional guidance, prompt reviews of my drafts and the final report. I am equally grateful to Prof. Walter Mwanda for his wise leadership, scholastic mentorship and support and all the staffs at UNITID for providing logistic and technical support in the course of this work. I will forever remain thankful to the University of Nairobi.

Gratitude is also owed to KNH and NASCOP for providing me data for my project.

Special thanks to my colleagues Augustine Njuguna, Chepkutto Yegon, Dr. Edwin Kagereki and Dr. Elizabeth Wangia for the support and encouragement towards completion of this project.

Lastly, I extend my appreciation to my parents Mr. & Mrs. Makory for their love and support, my husband Erick Abuga for financial assistance and moral support and my brothers and sisters- Arthur, Diana, Thaddeus, Jayde and Ignatius for their support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIDS	Acquired Immunodeficiency Syndrome
AIC	Akaike Infornation Criterion
APA	American Psychological Association
BMI	Body Mass Index
CCC	Comprehensive Care Centre
CDC	Centers for Disease Control and Prevention
CD4	Cluster of differentiation 4
GBV	Gender Based Violence
HIV	Human Immunodeficiency Virus
HAART	Highly Active Anti-Retroviral Therapy
HRQL	Health-Related Quality of life
KDHS	Kenya Demographic & Health Survey
KNH	Kenyatta National Hospital
MAR	Missing At Random
MCAR	Missing Completely At Random
MRM	Mixed Regression Models
NASCOP	National AIDS and STI Control Programme
SPSS	Statistical Package for Social Sciences
UNHCR	United Nations High Commission for Refugees
USA	United States of America.
UNFPA	United Nations Population Fund.
UN	United Nations
USDHSS	United States Delaware Department of Health and Social Services
PQL	Penalized Quasi-likelihood
WHO	World Health Organization

LIST OF ACRONYMS FOR ANTIRETROVIRAL DRUGS

3TC	Lamivudine
AZT	Zidovudine
D4T	Stavudine
ddi	Didanosine
EFV	Efavirenz
IDV	Indinavir
NVP	Nevirapine
TDF	Tenofovir

ABSTRACT

Background

The measurement of CD4+ T-cell (CD4) counts is a strong predictor of progression to AIDS and a means of monitoring antiviral therapy (ART). Understanding the wayCD4+ cell counts change over time would provide insight into the way patients respond to treatment and how effective treatment is with time. Secondary data from the Comprehensive Care Centre at KNH was utilized in which patients were enrolled and their CD4+ cell counts were regularly monitored thus generating repeated measures of their CD4+ cell counts. The aim of the present study was to assess the parsimony of the Generalized Estimating Equations and Generalized Linear Mixed effects models in assessing the change in the CD4+ count in HIV-positive patients initiated on ART over a period of at least five years.

Methodology

A Retrospective Longitudinal study with data obtained from the CCC at KNH, Nairobi of HIV-Positive patients enrolled and their CD4+ cell counts initially taken on enrolment into the ART programs and thereafter counted every 12 weeks. The study subjects were enrolled in the CCC between the period of 2008 and 2012 and all were over 18 years at the time of enrolment into the ART program. A total of 248 patients formed the sample and were used in the study. Data was explored using basic descriptive statistics and a profile of the mean CD4+ cell count over the period of the study done. GLMM and GEE models were used to model the change in CD4+ cell count over time. GLMM took into account both within and between sources of variation was flexible enough to account for the natural heterogeneity in the population and handle the degree of missing data while GEE allowed for the correlation between observations.

Results

The analysis included 248 individuals. The mean CD4+ cell count was seen to increase with duration of treatment on ART. The patient’s age, ART regimen and WHO clinical staging did not affect the current CD4+ cell count status. However, BMI and CD4+ cell counts after 12 weeks on treatment were significant predictors on the current CD4+ cell count. GEE model was the most parsimonious model compared to GLMM as proven by the lower Standard Error and also the AIC, Likelihood ration and BIC smaller is best fit criterion.

Conclusion

Longitudinal beta regression models are a natural candidate to analyze longitudinal data over time since they account for the bounded range and the skewed distribution of the response variable. However, depending on whether a population-averaged or a subject-specific approach is preferred, researchers should distinguish between a mixed (beta GLMM) and a marginal (beta GEE) model.

CHAPTER ONE

INTRODUCTION

1.1 Background information

CD4+ T lymphocyte (CD4) cell counts are the primary laboratory markers used to track the progression of HIV to AIDS. However, clinicians still debate the appropriate CD4 threshold at which to initiate HIV therapy. The 2013 WHO and the USDHHS guidelines recommend HIV therapy at CD4 cell counts less than 500 cells/μL, a recent departure from the prior guidelines that gave 350 cells/μL the strongest recommendation. While the therapy should be initiated based on individual patient characteristics, societal factors such as resource availability of health staff and a continuous supply of drugs must also be considered before initiating therapy. The debate of when to initiate therapy is also fueled by the lack of evidence from HIV "treatment initiation" randomized clinical trials (due to ethical implications). The scientific literature does, however, include observational treatment initiation studies of varying quality that can bring treatment guidelines closer to the best treatment strategy.

HIV infection leads to severe depletion of CD4 T cells in the gut-associated lymphoid tissue with subsequent reduced levels of circulating CD4 lymphocytes in the peripheral blood. CD4 cells are reduced precipitously in acute HIV infection, but usually rebound over several weeks as HIV-specific CD8 T cells help to lower plasma viremia. In the untreated patient, CD4 T cells subsequently decline over several years. Population-based studies of the natural history of HIV infection among men who have sex with men (MSM) show that the mean CD4 count prior to seroconversion is about 1000 cells/mm^3; CD4 T cell counts decline to a mean of 780cells/mm^3 at six months post-seroconversion and to 670cells/mm^3 at one year of follow-up. Subsequently, the CD4 cell count declines at an average yearly rate of approximately 50cells/mm^3, but there is substantial variation among patient. Significant depletion of CD4 T cells can lead to opportunistic infections and mortality in the untreated patient.

In longitudinal studies, measurements on a variable are collected at multiple points for each subject. A key feature of such longitudinal data is that within-Individual repeated measurements of a variable are correlated, although the measurements across individuals are usually assumed to be independent. Furthermore, the observed usually contains missing data, dropouts, censoring, outliers, and measurement errors, and are often unbalanced, since the numbers and times of measurements may vary across individuals.

There are two commonly used methods for analyzing longitudinal data;

- Generalized linear mixed effects models
- Generalized Estimating Equations models

Mixed effects models are extensions of standard regression models from cross-sectional data. They incorporate the two sources of variations in longitudinal data, between-individual variation and the within – individual correlation, by introducing random effects in the models. A major advantage of mixed effects models is that they allow for individual-specific inference in addition to standard population-average inference. Statistical inference for a mixed effects model is typically based on likelihood methods under the distributional assumptions for random effects and random errors.

Generalized Estimating Equations (GEE) models require specifications of the mean and covariance structure of the response, without any distributional assumptions for the data. Such models are usually more appealing for non-normal data. The resulting parameter estimates are consistent and asymptotically normal as long as the mean structure is correctly specified. However, the GEE estimates are usually not efficient and individual-specific inference is not available for GEE models.

In Longitudinal analysis, parametric models such as a linear mixed model are popularly used to model CD4 cell count change over time. The mixed effects models are parsimonious and efficient when the models are correctly specified.

The purpose of this study is to examine and compare the utility of the GEE and GLMM models for modeling longitudinal data, with particular focus on the change in CD4+ cell count in HIV patients started on Antiretroviral therapy(ART).

1.2 Problem Statement

Having a reliable marker to evaluate disease progression and predict treatment outcomes would be useful for the practitioner and patient alike.Laboratory markers used in monitoring management in HIV-positive patients are HIV-RNA assay (Viral load) and CD4 count. The former is the gold standard; its use is, however, limited because of its cost and technology. Furthermore, there is a mismatch between an undetectable viral load (<50 copies/mL) and the absence of immune reconstitution, which can be confusing to both the treatment provider and patient.

Since the introduction of HAART, much has been studied regarding which factors best predict a patient's success on HAART. The CD4+ cell count is then used in assessing the clinical status of HIV-infected individuals, in making informed decisions regarding the initiation of antiretroviral therapy and in monitoring the success of such therapy.

Several cohort studies and clinical trials have shown that the CD4 count is the strongest predictor of subsequent disease progression and survival. The use of the CD4 count as an independent and reliable marker for treatment outcome is attractive from various aspects. First, CD4 counts are already the most important factor in deciding whether to initiate antiretroviral therapy and opportunistic prophylaxis – all HIV-positive patients in high-income countries, and an increasing number of patients in low-income countries have a baseline CD4 count at entry into care. Secondly, the CD4 count is a relatively objective and simple marker to follow. Finally, the cost of CD4 counts has become more affordable, including in developing countries.

1.3 Justification

CD4 cell count and HIV RNA viral load in response to antiretroviral therapy (ART) are important measures of the efficacy of ART in individual patients and of the effectiveness of ART in populations of patients enrolled in HIV care and treatment programs. However, few data exist on long-term CD4 response to ART among patients receiving care in resource-limited settings, where HIV RNA testing is not generally available or conducted.In addition, CD4 count at the time of ART initiation is an important determinant of the degree of immunologic and virologic response as well as subsequent risk of morbidity and mortality.Among those patients who are able to remain on ART, robust immunologic responses can be maintained for long periods, and the risk of serious morbidity and mortality may eventually diminish to levels observed in the general population.

1.4 Study Limitations

We had to exclude a substantial number of patients due to lack of follow-up CD4 counts. These patients differed systematically from those who were included in the analysis: they were more likely to die or be lost before a follow-up CD4 count could be measured.

1.5 Objectives

1.5.1 Main objective

The aim of this study was to assess the parsimony of the Generalized Estimating Equations and Generalized Linear Mixed effects models in assessing the change in the CD4+ count in HIV-positive patients initiated on ART for a period of at least five years.

1.5.2 Specific Objectives

1. To establish the parsimony of the Generalized Estimating Equations and Generalized Linear Mixed effect models in the analysis of the change in CD4+ cell count.
2. To estimate the most optimal model in assessing the change in CD4+ cell count in HIV patients on ART.

1.6 Research Question

Which is the most parsimonious model to fit longitudinal data on change in CD4+ cell count in HIV-positive patients receiving ART?

CHAPTER TWO
LITERATURE REVIEW

2.0 Introduction to the Chapter

The literature selected and discussed are related and relevant to this study. Studies related to modeling of longitudinal HIV/AIDS data is scarce, however use of GEEs and GLMMs are becoming popular. The literature review given below has several parts; the overview of HIV/AIDs pandemic, modeling of longitudinal data with emphasis on GEE models and GLMM and a comparison of the two models.

2.1 Global Perspective of HIV/AIDS and Kenyan situation

Human Immunodeficiency virus (HIV) is a Lentivirus that causes Acquired Immunodeficiency Syndrome (AIDS) by reducing a person's ability to fight infection. HIV attacks CD4+ cell which is responsible for the body's immune response to infectious agents. An uninfected individual has around 1100 CD4+ cells per milliliter of blood. This CD4+ cells decrease in number with time, so that an infected person's CD4+ cell count can be used to monitor the progression of the disease (Diggle et al 1994).

According to the Joint United Nations Programme on HIV/AIDS (UNAIDS), 34million people are living with HIV in the world. Of these, approximately 23.5 million are in sub-Saharan Africa. Globally, 14.8 million people are eligible for HIV treatment, while 8 million are on ART (UNAIDS, 2012). As at 2011, HIV/AIDS resulted in about 1.7 million deaths and 2.5million new infections (UNAIDS, 2012)

Antiretroviral therapy (ART) services have been available to HIV-positive patients and the guidelines from the National AIDS Control Programme (NASCOP) recommend that patients should initiate treatment when their CD4+ cell count is less than 350cell/ml or when they become symptomatic with HIV infection as in WHO stage I to IV. Once a patient enrolls on the ART treatment, the CD4+ cell count of the patient is examined from time to time to check whether there is an increase in its count to a relatively normal level(>500cells per microliter) or otherwise(NASCOP, 2001). Hence CD4+ cell count is pivotal in determining when to initiate ART and in staging of HIV/AIDS. It gives information on possible treatment failure.

2.2 Cluster of Differentiation 4

CD4 (Cluster of Differentiation 4) refers to the glycoprotein found on the surface of immune cells such as T helper cells, monocytes, macrophages and dendritic cells. In humans, the CD4 protein is encoded by the CD4 gene (Isobe et al., 1986; Ansari-Lari et al., 1996). T lymphocytes are divided into;

- Helper T cells which help in the functions of the immune system.
- Cytotoxic T cells also called Killer cells which kill infected cells (Kumar, 2012)
- Suppressor T cells which are capable of suppressing the function of both cytotoxic and helper T cells.

HIV causes AIDS by destroying CD4+ T cells (Almonti et al.,2003). CD4+ cell count, therefore, measures the degree of immunosuppression in HIV- positive patients. There is inverse relationship between CD$ count and degree of immunosuppression (Akinbami et al., 2012). Few laboratories in resource- restrained countries can afford to perform CD4+ cell count and HIV viral load (Crowe et al., 2003).

In Kenya, CD4+ cell count is routinely done. It plays an important role in deciding when to commence therapy, staging the disease, monitoring disease progression and determining treatment failure. Generally, CD4+ cell count takes priority over viral load if both tests cannot be carried out together because of financial constraints (Crowe et al., 2003). The cost of CD4+ cell count is lower than viral load and it’s increasingly becoming more affordable to patients in developing countries (Mellors et al., 1997; Lutwana et al., 2008)

2.3 Modeling longitudinal data

Statistical methods for the analysis of longitudinal data have advanced dramatically. A straightforward application of Generalized Linear Models to longitudinal data is not appropriate, due to lack of independence among repeated measures obtained on the same individual. There has been extensive statistical literature on extending generalized linear models to the longitudinal-data setting. The standard Generalized Linear Models (GLM) assumes that observations are uncorrelated. The standard approach to analysis of longitudinal data principally involved using the longitudinal data to impute end-points (e.g. last observations carried forward; LOCF) and then to simply discard the valuable intermediate time-point data, favoring the simplicity of analyses of change scores from baseline to study completion.

Laird and Ware(1982) showed that GLMMs and GEE models could be used to perform a more complete analysis of all of the available longitudinal data under much more general assumptions regarding the missing data. The net result was a more powerful set of statistical tools for analysis of longitudinal data that led to more powerful statistical hypothesis tests, more precise estimates of rates of change (and differential rates of change between experimental and control groups).Although longitudinal studies provide far more information than the cross-sectional studies and are therefore now in widespread use, they are not without limitations.

One; Individual differences are the norm rather than the exception.Hence these personal characteristics may be unobserved, leading to unexplained heterogeneity in the population. Modeling this unobserved heterogeneity in terms of variance components that describe subject-level effects is one way to accommodate the correlation of the repeated responses over time and to better describe individual differences in the statistical characterization of the observed data. These variance components are often termed “random effects,” leading to terms like random-effects or mixed-effects regression models.

Two, there is also short-term correlated errors of measurements that are produced by the psychological state that a subject is in during measurement occasions that are close in time. This type of short-term residual correlation tends to decrease exponentially with the temporal distance between measurement occasions. The addition of auto correlated residuals (Chi &Reinsel., 1989, Hedeker 1989) to mixed-effects regression models allows for a more parsimonious analysis of the more subtle features of the longitudinal response process and results in more accurate estimates of uncertainty in parameter estimates, improved tests of hypotheses, and more accurate interval estimates.

In an attempt to provide a more general treatment of longitudinal data, with more realistic assumptions regarding the longitudinal response process and associated missing data mechanisms, statistical researchers have developed a wide variety of more rigorous approaches to the analysis of longitudinal data. Among these, the most widely used include mixed-effects regression models (Laird & Ware 1982) and generalized estimating equation (GEE) models (Zeger& Liang 1986). Variations of these models have been developed for both discrete and continuous outcomes and for a variety of missing data mechanisms.

The primary distinction between the two general approaches is that mixed-effects models are full-likelihood methods and GEE models are partial-likelihood methods. The advantage of statistical models based on partial likelihood is that they are computationally easier than full-likelihood methods, and they generalize quite easily to a wide variety of outcome measures with quite different distributional forms. The price of this flexibility, however, is that partial likelihood methods are more restrictive in their assumptions regarding missing data than are their full-likelihood counterparts. In addition, full-likelihood methods provide estimates of person-specific effects (e.g., person-specific trend lines) that are quite useful in understanding interindividual variability in the longitudinal response process and in predicting future responses for a given subject or set of subjects from a particular subgroup (e.g., a county, a hospital, or a community).

2.3.1 Generalized Estimating Equation (GEE) Models

During the 1980s, alongside the development of mixed-effects regression models for incomplete longitudinal data, the generalized estimating equation (GEE) models were developed (Liang &Zeger 1986 and Zeger& Liang 1986). Essentially, GEE models extend generalized linear models (GLMs) to the case of correlated data. Thus, this class of models has become very popular, especially for analysis of categorical and count outcomes, although they can be used for continuous outcomes as well i.e. GEEs provide a general framework for the analyses of continuous, ordinal, polychotomous, dichotomous and count-independent data.GEE models are termed marginal models, and they model the regression of *y* on *x* and the within-subject dependency (i.e., the association parameters) separately.

$$\mathbf{g(E(Y)) = X\ \boldsymbol{\beta}}$$

The term “marginal” in this context indicates that the model for the mean response depends only on the covariates of interest and not on any random effects or previous responses. In terms of missing data, GEE assumes that the missing data are missing completely at random (MCAR) as opposed to MAR, which is assumed by the models employing full-likelihood estimation.

Conceptually, GEE reproduces the marginal means of the observed data, even if some of those means have limited information because of subject dropout. Standard errors are adjusted (i.e., inflated) to accommodate the reduced amount of independent information produced by the correlation of the repeated observations over time (or within clusters). The most salient feature of marginal models is a regression model, with appropriately specified link function, relating the mean response of each occasion to the covariates. By contrast, mixed-effects models use the available data from all subjects to model temporal response patterns that would have been observed had the subjects all been measured to the end of the study. Because of this, estimated mean responses at the end of the study can be quite different for GEE versus MRM if the future observations are related to the measurements that were made during the course of the study. This leads to a preference for full-likelihood approaches over quasi- or partial-likelihood approaches and MRM over GEE, at least for longitudinal data. There is certainly less of an argument for a preference for data that are only clustered (e.g., children nested within classrooms), in which case advantages of MAR over MCAR are more difficult to justify.

A basic feature of GEE models is that the joint distribution of a subject's response vector y_i does not need to be specified. Instead, it is only the marginal distribution of y_{ij} at each time point that needs to be specified. To clarify this further, suppose that there are two time points and suppose that we are dealing with a continuous normal outcome. GEE would only require us to assume that the distribution of y_{i1} and y_{i2} are two univariatenormals, rather than assuming that y_{i1} and y_{i2} form a (joint) bivariate normal distribution.Thus, GEE avoids the need for multivariate distributions by only assuming a functional form for the marginal distribution at each time point. GEEs provide consistent, asymptotically normal, un biased standard errors, even with incorrect specification of intracluster dependence structure, assuming the mean model is correctly specified and with complete or missing completely at random data (following classification of Rubin, 1976). GEEs also offer two variance estimator algorithms. One algorithm is model-based and it is the only one available in the more popular multi-level models. The second estimator is commonly referred to as robust (or empirical, Huber/White sandwich, model-free,agnostic) , meaning that it is robust to misspecification of the working correlation matrix. Moreover, Cheong et al., 2001 showed, via simulation studies, that even when data are naturally organized within clusters, and the analyses do not account for such clusters, in large sample sizes the robust estimation yields correct standard errors (Raudenbush&Bryk, 2001)

Liang and Zeger, (1986) allow for the correlation between observations without the use of explicit probability model for the origin of the correlation, so there is no explicit likelihood. They are suitable when the random effects and their variances are not of inherent interest as they allow for the correlationwithout explaining its origin. The focus is on estimating the average response over the population ("population-averaged" effects) rather than the regression parameters that would enable prediction of the effect of changing one or more components of X on a given individual. GEEs are usually used in conjunction with Huber-White standard errors.

The assumptions maintained by the GEE method are that (1) the dependent variable is linearly related to the predictors (when the dependent variable is non-normally distributed a nonidentity link function is to be selected); (2) the number of clusters be relatively high (a rule of thumb is no fewer than ten, possibly more than 30; Norton et al., 1996); (3) the observation in different clusters be independent.

To augment the efficiency of GEEs, Prentise (1988, Zhao &Prentise,1990) introduced a variation called GEE2, which requires the correct specification of both mean model and the correlation structure. The gain in efficiency, however, seems to be minor (Liang, Zeger&Qaqish, 1992). Moreover, when the correlation structure is misspecified, the GEE2 estimated parameters are non-consistent.

However, over the past 20 years, the GEE approach has proven to be a useful method for the analysis of longitudinal data, especially when the response variable is discrete (binary, ordinal or count outcomes)

2.3.2 Generalized Linear Mixed-effects Models (GLMMs)

Generalized Linear Mixed-effects regression models are now quite widely used for the analysis of longitudinal data (38papers in 2005, 62 in 2006, 83 in 2007 and 17 in 2008 to date). These models can be applied for normally distributed continuous outcomes as well as categorical outcomes and other non-normally distributed outcomes such as counts that have a Poisson distribution. Literature review found that many analyses (58%, n=537) used GLMMs inappropriately. The most frequent and severe problem was the use of Penalized Quasi-likelihood (PQL) in situations where it may be biased (Breslow 2005) and the second most common misuse of GLMMs involved the analysis of random effects with too few level; (16%, n=462) of analysis estimated random effects for factors with fewer than four levels, which is not wrong but leads to imprecise estimates of the standard deviation. About 11% of papers used GLMMs only to analyze normally distributed data.

GLMMs are extensions of Generalized Linear Models to longitudinal data by allowing a subset of the regression coefficients to vary randomly from one individual. They enable for accounting for the within subject association. GLMMs have their foundation in simple random-effects models for binary and count data. From an historical perspective, the papers by Ashford and Sowden (1970), Pierce and Sands (1975), and Korn and Whittemore (1979) laid the conceptual foundations for GLMMs. In GLMMs the marginal likelihood is used as the basis for inferences for the fixed-effects parameters, complemented with empirical Bayes estimation for the random effects.

In GLMMs, the model for the mean response is conditional upon both measure covariates and unobserved random effects; the inclusion of the latter induces correlation among the repeated responses marginally, when averaged over the distribution of random effects.

The Generalized linear mixed-effects regression model for the measurement *y* of individual *i* (*i* = 1, 2 …*N* subjects) on occasion *j* (*j* = 1, 2,.…n_j occasions):

Y ij =β 0 +β 1 t ij +ε ij

That is;

Mixed models = fixed and random effects.

$$Y_{it} = \beta_{0i(random)} + \beta_{time(fixed)} + Error$$

Ignoring subscripts, this model represents the regression of the outcome variable *y* on the independent variable time (denoted *t*). The subscripts keep track of the particulars of the data, namely whose observation it is (subscript *i*) and when this observation was made (the subscript *j*). The independent variable *t* gives a value to the level of time and may represent time in weeks, months, etc. Since *y* and *t* carry both *i* and *j* subscripts, both the outcome variable and the time variable are allowed to vary by individuals and occasions.

In linear regression models, the errors $\varepsilon_{i\ j}$ are assumed to be normally and independently distributed in the population with zero mean and common variance σ^2. This independence assumption makes the typical general linear regression model unreasonable for longitudinal data. This is because the outcomes *y*are observed repeatedly from the same individuals, and so it is much more likely to assume that errors within an individual are correlated to some degree. Furthermore, the above model posits that the change across time is the same for all individuals since the model parameters (β_0, the intercept or initial level, and β_1, the linear change across time) do not vary by individuals. For both of these reasons, it is useful to add individual-specific effects into the model that will account for the data dependency and describe differential time trends for different individuals.

The addition of auto correlated residuals (Chi & Reinsel 1989, Hedeker 1989) to mixed-effects regression models allows for a more parsimonious analysis of the more subtle features of the longitudinal response process and results in more accurate estimates of uncertainty in parameter estimates, improved tests of hypotheses, and more accurate interval estimates. Thus correct specification of the correlation structure augments efficiency (Y. -G. Wang & Carey, 2003) and several specifications are commonly adopted; Independent, Exchangeable, Autoregressive, Stationery M, M-independent or non-stationery, Unstructured and specified or fixed. The choice among the several specifications should be based on substantive reasons, and sensitivity analyses of the different specifications of the correlation are recommended (Y. –G. Wang & Carey; Zorn, 2001).

Regarding parameter estimates, for continuous and normally distributed outcomes, Hedeker et al. (1994) noted that the fixed effects estimates are not greatly affected by the choice of model. However, the estimates of the standard errors, which determine the significance of these parameter estimates, are influenced by the choice of model. In general, when a source of variability is present but ignored by the statistical model, the standard errors will be underestimated. Underestimation of standard errors results since the statistical model assumes that, conditional on the terms in the model, the observations are independent. However, when systematic variance is present but ignored by the model, the observations are not independent, and the amount of independent information available in parameter estimation is erroneously inflated.

2.4 Model comparison

Model comparison and model checking in the GLMM and GEE framework is not straightforward and suitable methods are sparse. In general, if GLMMs are estimated using a full likelihood approach, models can be compared using information criteria such as Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC). AIC and BIC are measures of the likelihood, penalized for the complexity of the model.

The choice between the two depends mainly on the specific scientific question of interest. GLMMs are most useful for making inferences about individuals and tracking individual trajectories, while the marginal model is more useful for inferences about population or sub-population averages. No model is a priori more suitable for the analysis of HRQL data than the other. It has been argued that mixed models may be more appropriate in epidemiological research as they allow a better understanding of the underlying mechanisms. Also, they have a close relationship to matched-pair design methods often used in epidemiologic and public health research. Due to the individual-specific interpretation of regression coefficients, the GLMM is also most meaningful for time-varying covariates. In contrast, the interpretation of time-invariant or between-subject covariates in the GLMM is less intuitive or even misleading since they also only allow a within-subject interpretation which is difficult to imagine. For example, if a beta GLMM is used to estimate treatment effects on HRQL in clinical trials, the respective treatment arm coefficient is interpreted as the difference in outcomes between two individuals with the same covariate values and the same random effects bi, differing only in their treatment arm. It does not describe the average treatment effect which is usually of major interest in intervention studies, especially if preference-based HRQL measures are used in economic evaluation studies. Therefore, the marginal model may be more suitable in many applications in public health research. Also, it has been argued that many epidemiologic methods such as stratified methods are essentially population-averaged methods. Differences between beta GLMM and beta GEE also exist with respect to the handling of missing data: In practice, the beta GLMM may be more convenient since it remains valid under the MAR assumption which is usually more plausible in quality of life studies than the MCAR assumption made by the beta GEE.

A common approach to compare regression models and assess goodness of fit is to consider likelihood-based statistics which evaluate the probability of the observed data under the model.

CHAPTER THREE

METHODOLOGY

3.0 Introduction to the chapter

This chapter describes the method that was used to meet the study objectives. Generalized Estimating Equations Model and Generalized Linear Mixed-effects Models were used on longitudinal data and evaluated using Akaike Information Criterion. A description of the research setting, research tools used, research procedure and the ethical issues relating to the study are also given.

3.1 Study design

It is a retrospective longitudinal study carried out from 1st of January 2008 through 31st December 2012. A total of 248 subjects were sampled. Gender , age at initiation of ART, Baseline weight, CD4 cell count(cells/mm^3) taken at the initiation of ART and thereafter every six-weekly CD4 cell count up to six months and world health clinical staging at the initiation of ART. GEE model and GLMM were performed with the Akaike Information Criterion (AIC) being used to compare the efficiency of the models.

3.2 Study Area and Population of interest

The study was carried out in Kenyatta National Hospital located in Nairobi, the capital city of Kenya. Kenyatta National Hospital is a teaching and referral hospital and receives patients from all over Eastern and Central Africa and has optimum service provision and best outcomes in interventions administered to patients presenting to it. It has a Comprehensive Care Centre whichoffers a wrap-around approach to HIV disease management. A multidisciplinary team provides specialized HIV care to address the diagnosis and treatment of opportunistic infections and other HIV/AIDS-related co-morbidities, antiretroviral therapy management, laboratory monitoring of patients, antiretroviral adherence counseling, nutrition counseling, social work services, and supportive counseling. Currently, 6642 patients are actively being followed up in the CCC with approximately 5500 patients receiving ARV therapy and another 950 patients who do not qualify to be initiated on ARVs on palliative care. Patients qualify for ARV therapy when CD4 counts fall below 350 or when they are in WHO clinical stage III or IV. Most of the patients (95%) on ARV are on 1st line treatment regimen which consists of an NRTI backbone and an NNRTI.

The study population consisted of HIV-Positive patients above 18 years who were initiated on ART between 1st of January 2008 through 31st December 2012.

3.2.1 Sample size

Sample size determination was addressed in the original study.

The following formula was used for sample size computation:

$n = (z_{1-\alpha/1}+z_{1-\beta})^2 2\sigma^2/ \mu_1-\mu_2$

Where α = significant level (0.05)

$1-\beta$ = the power of the study (90%)

$Z_{1-\alpha/2}$ =Z-value attributed to $\alpha/2$ (1.96)

$Z_{1-\beta}$ = Z-value attributed to $1-\beta$ (1.28)

$\mu_1-\mu_2$ = the expected difference between the subjects on TDF + 3TC +EFV or NVP and AZT + 3TC + NVP or EFV

This gave a total of 288 but only those whose CD4 cell count was complete were analyzed giving a total of 248 subjects.

3.2.2 Sampling procedure

The sampling frame for the study was all cases of HIV-Positive patients above 18 years started on ART between January 1st 2008 and 31st December 2012. Sampling was carried out using Simple Random Sampling (computer generated random numbers).

3.3 Variables

The count variable of interest was the CD4 cell count (cells/mm^3) with repeated measurements every six weeks up to six months, the WHO clinical staging at initiation of ART and thereafter upto six months as well as change in BMI from initiation of ART up to six months and age at start of study.Categorical variable was ART various regimens.

3.3.1 The Response Variable

The response or outcome variable in this study is the CD4+ Cell count at 60 weeks from the date of ART initiation.

3.3.2 The Predictor Variables

The predictor variables in longitudinal analysis are called covariates. These are explanatory variables which are assumed to influence the outcome of CD4+ cell count after initiation of ART and are as below:

1. Age in years (at ART initiation)……………………………………
2. Baseline BMI………………………………………………………….
3. Regimen type………………………………………………………….
4. WHO staging (at ART initiation)…………………………………….
5. Baseline CD4+ cell count……………………………………………..
6. CD4+ cell count at 12 weeks………………………………………….
7. CD4+ cell count at 24 weeks…………………………………………..
8. CD4+ cell count at 36 weeks…………………………………………….
9. CD4+ cell count at 48 weeks……………………………………………
10. CD4+ cell count at 60 weeks…………………………………………

3.4 Statistical Analysis

Descriptive statistics of continuous variables was done to show the distribution of patient factors, as well as the minimum and maximum CD4 cell counts. Correlation between the repeated measurements was done and significant relationships chosen. GLMM and GEE models were fitted to the CD4 cell count data to explore predictive relationships with age, BMI, ART regimen and WHO clinical staging. The likelihood ration and the BIC were used to determine the most optimal model.

Statistical significance level for analyses in this study was taken as $p<0.05$.

The basic Generalized Estimating Equations model is presented below as;

Y_{ij}(CD4 count at 60weeks)= $\beta_0 + \beta_1(\text{age}) + \beta_2(\text{BMI}) + \beta_3(\text{ART regimen}) + \beta_4(\text{cd4t0}) + \beta_5(\text{cd4t12}) + \beta_6(\text{cd4t24}) + \beta_7(\text{cd4t36}) + \beta_8(\text{cd4t48})$ + CORR + Error

The basic Generalized Linear Mixed- effects Model is;

$Y_{ij} = \beta_{0j} + \beta_{1j(\text{cd4t0})} + \beta_{2j(\text{cd4t12})+}\ \beta_{3j(\text{cd4t24})}\ldots\ldots\ldots + \beta_{5j(\text{cd4t48})} + b_{i,0} + b_{i,1(\text{cd4t0})} + b_{i,2(\text{cd4t12})} + b_{i,3(\text{cd4t24})}\ldots\ldots + b_{i,4(\text{cd4 48})} + \text{Error}_{ij}$

Data was analyzed using SPSS Statistical software version 20

3.5 Ethical Consideration

The primary study was approved by the KNH/UON Ethical Review Committee before implementation. All facets of the relevant ethics was adequately addressed, that is details and importance of the study was explained to the recruited patients with particular emphasis on the fact that the study would help health workers in understanding and making informed decisions regarding the initiation of antiretroviral therapy and in monitoring the success of such therapy and willing participants were asked to sign an informed consent form, hence was not replicated here except for a formal application and subsequent acquisition of the original datasets.

3.6 Validity and Reliability

Content validity is based on the adequacy with which the items in an instrument measure the attributes of the study (Nunnally.,1978). Content validity of the method was ensured through constructive criticism from colleagues in class and the supervisors who have extensive experience in research. Reliability is the extent to which any measuring method yields the same results on repeated trials (Carmines & Zeller, 1979). The reliability of the method will be ensured through fitting the model to hypothetical datasets. Furthermore, the reliability and validity of the results will be obtained through member checks to help indicate whether the findings appeared to match with perceived authenticity. This will be done in order to limit the distorting effects of random errors on the findings.

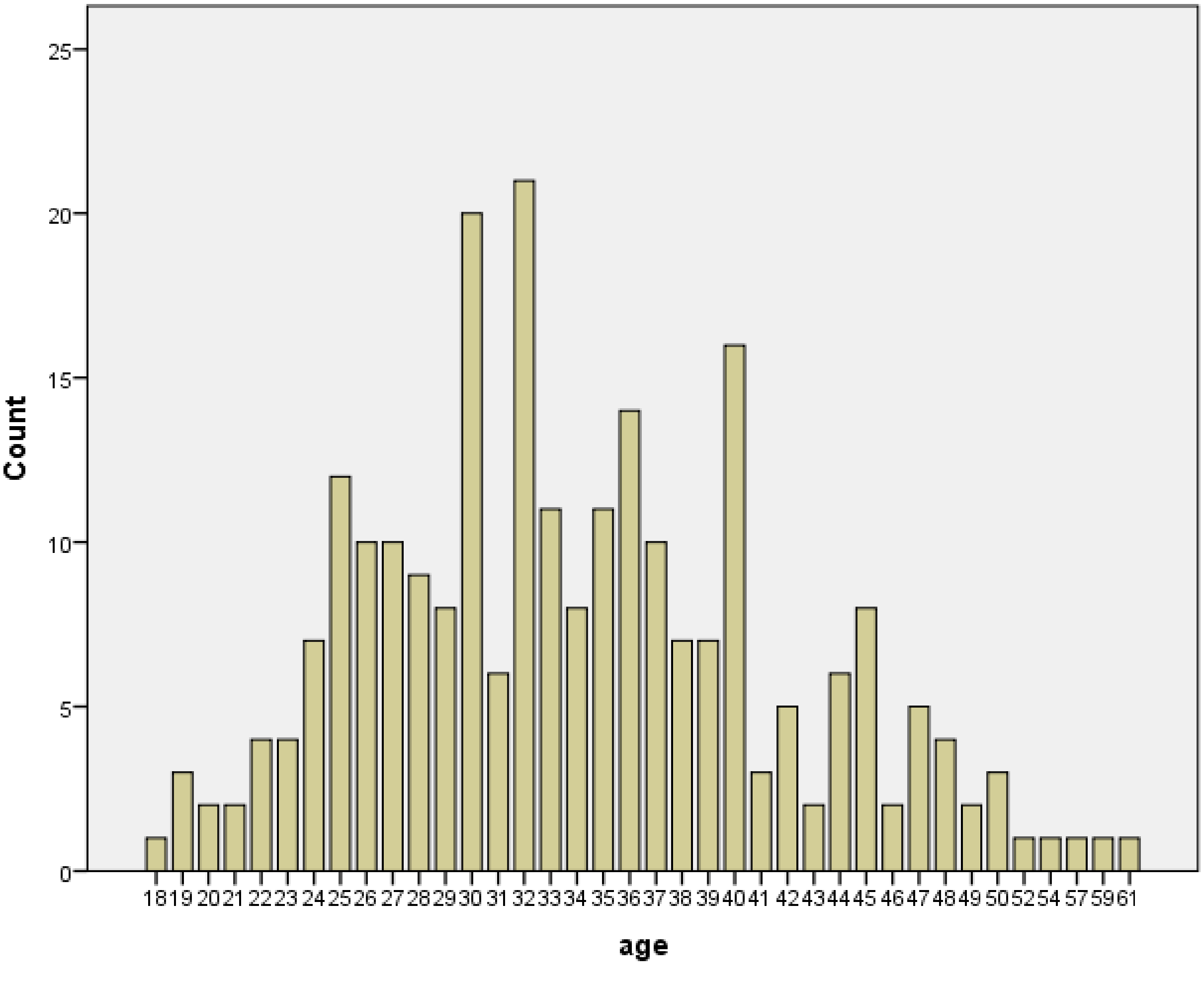
CHAPTER FOUR

RESULTS

4.0 Introduction

In this chapter, the results of the study are described and the analysis of the data presented. Data analysis was done using SPSS version 20. The results describe information on the subjects under study and the changes in the variable of interest by performing longitudinal analysis using GEE model and GLMM.

Figure 1 Bar graph of Age characteristic



The table indicates the age distribution of the 248 patients initiated on ART, with the highest number of patients at 32 years of age and the lowest number at 18, 52,54,57,59, 61 years of age.

Table 1 Descriptive statistics on Age and BMI

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Age	248	18	61	34.04	7.915
BMI at baseline	248	12.45	36.63	21.0063	3.74829
Valid N (listwise)	248				

There were a total of 248 patients observed. The mean age was 34.04(SD 7.915) and the baseline BMI averaged 21.0063(SD 3.74829) at initiation of Antiretroviral Therapy.

Table 2 CD4+ cell count characteristics

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
cd4 count at time0	248	1	773	209.03	122.787
cd4 count at 12weeks	248	50	1180	351.83	170.059
cd4 count at 24weeks	248	78	1121	420.74	189.749
cd4 count at 36weeks	248	59	1389	501.60	213.934
cd4 count at 48weeks	248	111	1359	540.64	217.087
cd4 count at 60weeks	248	88	1467	587.59	244.286
Valid N (listwise)	248				

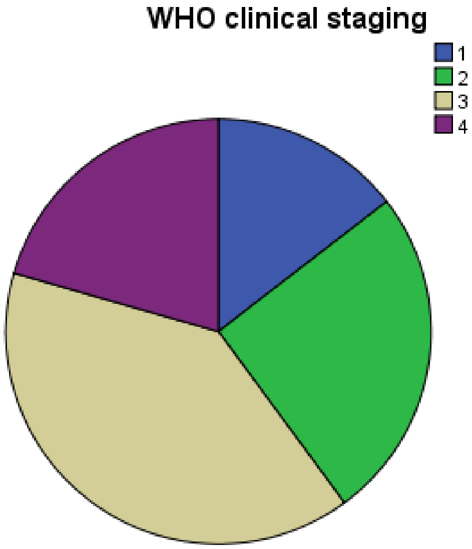
A total of 248 patients were observed and there CD4+ cell counts measured from the start of ART up to 60 weeks of treatment. The mean CD4+ cell counts over each time period is as indicated above as well as the standard deviation (SD) for each mean CD4+ cell count.

Table 3 Regimen characteristics

Regimen	N	%
3TC/D4T30/NVP	**70**	**28.2**
3TC/D4T40/NVP	**151**	**60.9**
3TC/D4T/EFV	**22**	**8.9**
AZT/3TC/NVP	**5**	**2.0**
3TC/TDF/NVP	**0**	**0**
AZT/DDI/NVP	**0**	**0**
TOTAL	**248**	**100**

The table shows that all patients were on a triple therapy. 89.1% of the subjects were on first line regimen based on 2 nucleoside reverse transcriptase inhibitors (NRTIs) and a non-nucleoside reverse transcriptase inhibitor(NNRTI)i.e. 3TC/D4T(30/40)/NVP, while 22 subjects were on 3TC/D4T/EFV a regimen given to patients on Tuberculosis treatment. 5 patients were on AZT/3TC/NVP regimen which was previously a first-line regimen that is being phased out due to adverse drug reactions caused primarily by zidovudine (AZT).

Figure 2 World Health Organization Clinical Staging



WHO clinical staging

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	36	14.5	14.5	14.5
	2	63	25.4	25.4	39.9
	3	98	39.5	39.5	79.4
	4	51	20.6	20.6	100.0
	Total	248	100.0	100.0	

A total of 36 subjects (14.5%) were in WHO clinical stage 1, 63 subjects (25.4%) were in WHO clinical stage 2, 98 subjects (39.5%) were in WHO clinical stage 3 and a total of 51 subjects (20.6%) were in WHO clinical stage 4.

Table 4 Regression analysis

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	T	Sig.
		B	Std. Error	Beta		
1	(Constant)	-83.921	69.989		-1.199	.232
	Age	-.101	.953	-.003	-.106	.915
	BMI at baseline	4.885	2.123	.075	2.301	.022
	cd4 count at time0	-.095	.078	-.048	-1.225	.222
	cd4 count at 12weeks	.077	.082	.054	.938	.349
	cd4 count at 24weeks	.168	.080	.130	2.099	.037
	cd4 count at 36weeks	.205	.068	.179	3.005	.003
	cd4 count at 48weeks	.672	.063	.597	10.710	.000
	ART regimen	2.851	11.668	.008	.244	.807
	WHO clinical staging	8.739	8.258	.034	1.058	.291

a. Dependent Variable: cd4 count at 60weeks

The BMI at baseline, CD4+ cell count at 24 weeks, CD4+ cell count at 36weeks and CD4+ cell count were the predictor variables that were statistically significant as shown by the multivariate analysis results.

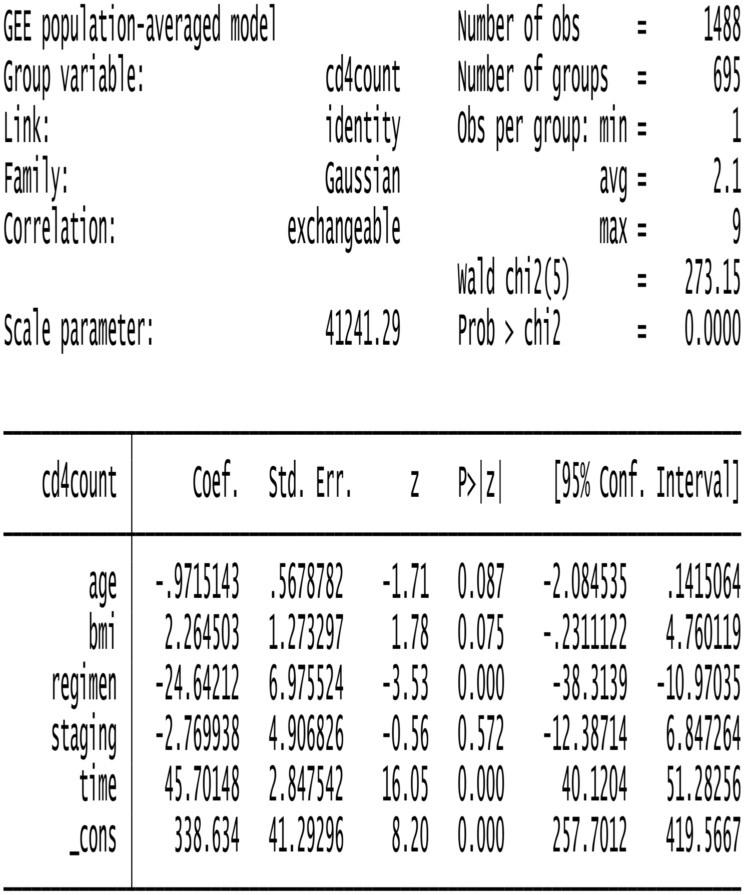
The other predictor variables i.e. Age, CD4+ cell count at baseline, CD4+ cell count at 24 weeks on ART treatment, ART regimen and WHO clinical staging were not statistically significant.

Figure 3Correlation matrix on Repeated CD4+ cell counts

	CD4t0	CD4t12	CD4t24	CD4t36	CD4t48	CD4t60
CD4t0	**1**	**0.597**	**0.525**	**0.439**	**0.425**	**0.379**
CD4t12		**1**	**0.81**	**0.702**	**0.68**	**0.66**
CD4t24			**1**	**0.784**	**0.743**	**0.736**
CD4t36				**1**	**0.815**	**0.79**
CD4t48					**1**	**0.861**
CD4t60						**1**

The correlation structure was of the form illustrated above. Correlation was significant at the 0.01 level (2-tailed) and 0.05 level (2-tailed). The structure is unstructured with decreasing correlation for further time periods. All correlations are estimated separately. The CD4+ cell counts have a within-person correlation that is high for observations close together in time, but the correlation tends to decrease with increasing time separation between the measurements.

Table 5 Generalized Estimating Equations Model



```
GEE population-averaged model              Number of obs      =      1488
Group variable:                 cd4count   Number of groups   =       695
Link:                           identity   Obs per group: min =         1
Family:                         Gaussian                  avg =       2.1
Correlation:                exchangeable                  max =         9
                                           Wald chi2(5)       =    273.15
Scale parameter:                41241.29   Prob > chi2        =    0.0000

------------------------------------------------------------------------------
    cd4count |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.9715143   .5678782    -1.71   0.087    -2.084535    .1415064
         bmi |   2.264503   1.273297     1.78   0.075    -.2311122    4.760119
     regimen |  -24.64212   6.975524    -3.53   0.000     -38.3139   -10.97035
     staging |  -2.769938   4.906826    -0.56   0.572    -12.38714    6.847264
        time |   45.70148   2.847542    16.05   0.000      40.1204    51.28256
       _cons |    338.634   41.29296     8.20   0.000     257.7012    419.5667
------------------------------------------------------------------------------
```

This model describes how the population *as a whole* behaves, *not how an individual within that population* will behave

It is generally understood that likelihood ratio tests have better statistical properties than Wald tests. The generalized estimating equations are not a form of maximum likelihood estimation, thus likelihood ratio tests are not available. So one can go ahead with the Wald test that is reported.

The GEE model provided a good fit to the data given the significant goodness of fit test (Prob> chi2= 0.00000). The ART regimen and CD4+ cell count(repeated measures) were found to be significant predictors of the current CD4+ cell count at alpha<0.05.

Table 6 Generalized Linear Mixed-effects Model

```
Residuals:
    Min      1Q  Median      3Q     Max
-506.42 -122.79  -27.24   91.99  912.88

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 242.3421    45.7037   5.302 1.32e-07 ***
age          -2.0606     0.6448  -3.196  0.00142 **
bmi           3.8972     1.4329   2.720  0.00661 **
regimen     -38.5415     7.7851  -4.951 8.24e-07 ***
staging      -0.6204     5.5674  -0.111  0.91128
time         72.5740     2.9777  24.372  < 2e-16 ***
---
Signif.codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 196.2 on 1482 degrees of freedom
Multiple R-squared:  0.3004,    Adjusted R-squared:  0.298
F-statistic: 127.3 on 5 and 1482 DF, p-value: < 2.2e-16
```

The Generalized Linear Mixed – effects model provided a good fit to the data given the significant goodness of fit test (p-value = 0.000). The age, BMI, regimen and CD4 + cell counts(repeated measures) were found to be significant predictors of the current CD4+ cell count at alpha<0 05.

Table 7 Comparing the GLMM and GEE models

Goodness of fit

GEE	**GLMM**
QIC = 14739938.05	**AIC = 20,455,081**
QICC^b = 14739938.05	**BIC = 20, 460,383**

QIC (Quasi likelihood under independence model criterion)

QICC^b(Corrected quasi likelihood under independence model criterion)

AIC (Akaike Information Criterion)

BIC (Bayesian information Criterion)

The relative fit of GEEs vs GLMMs could not be directly compared. This is because GEEs use quasi-likelihood, while GLMMs use maximum- likelihood framework for model estimation.

Comparative measures such as Akaike's Information Criterion (Burham& Anderson 1978) are used for evaluating relative fit of models for GLMM (Boker et al, 2009), whereas the quasi-likelihood – under- the independence- model information criterion, or QIC (Pan 2001) are used for evaluating relative fit of models for GEE, but there is no criterion that can be used for both.

Also, because parameter estimates from GLMM were conditional, while parameter estimates from GEE were marginal, parameter estimates and significance differed, and their comparison is therefore not appropriate.

CHAPTER FIVE

DISCUSSION

The importance of early treatment was evident from this study. The Initial CD4+ cell count was shown to significantly determine a patient's current CD4+ cell count following initiation on ART. A higher initial CD4+ cell count would result in a better rate of recovery of patients on ART. This agrees with findings of Viviane et al (2009) and Kulkarni et al (2011).

This study did not show any age differentials. However, the BMI was shown to significantly determine a patient's current CD4+ cell count hence a higher baseline BMI predicts greater gains in CD4+ cell counts. This finding is in contrast to the results in the Crum- Cianflone et al study, which showed that obese patients have smaller CD4+ cell count gains. However possible explanations for the relationship between higher BMI and higher CD4+ cell count gains includes the effects of adipokines such as Leptin, differences in thymic size, differences in lymphocyte population dynamics in the gastrointestinal tract and other mucosal sites and differences in T- lymphocyte apoptosis. Therefore, persons with higher BMI may naturally have higher CD4+ cell counts, and the greater CD4+ cell count recovery on ART in HIV- Infected patients with higher BMI could be explained simply by a "return to health" phenomenon.

Since the visits are equally spaced and each subject is scheduled to have a total of six measurements, the unstructured correlation matrix was used. The unstructured covariance structure model often offers the best fit and is most commonly used in longitudinal data as it is the most parsimonious, which requires no assumption in the error structure. A comparison of Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) for the known covariance structures was done and the model best fit for the covariance structure was the autoregressive moving average model. This means that there is correlation between CD4+ cell counts and that the correlation weakens with distance between counts. Thus, even though a patient's CD4+ cell count depended on his/her past CD4+ cell count, the strength of the relationship was stronger with his/her immediate past CD4+ cell count, and weakened with increasing time difference between counts.

In GLMM, the within-subject variation was seen as the deviation between individual observations. Each subject had an individual-specific intercept and slope. Within-subject variations were seen in the magnitude of variation in the deviation between the observations and the individual trajectory. The between- subject variation was represented by the variation among the intercepts, variation (β_i, 0) and the variation among subjects in the slopes i.e. variation ($\beta_{i,}$ 1). The resulting estimated $b_1$the fixed-effect parameter for each predictor in this model, represents the average change in CD4+ cell count for a unit increase in that predictor.

In the GEE model the estimated regression coefficient, β_0 were broad valid estimates that approached the correct value with increasing sample size regardless of the choice of correlation model. The correlation model choice was used simply to weight observation and a good correlation model choice led to more precise estimation of regression coefficients than a poor choice. Based on optimal estimation theory (Gauss Markov theory) the best correlation model choice for efficiency of estimation was the true correlation structure.

CHAPTER SIX

CONCLUSION AND RECOMMENDATIONS

6.1 Conclusion

Generally all patients that were considered in the study between January 2008 and December 2012 had their CD4+ cell count increased at different levels after being put on HAART at a certain initial CD4+ cell count. The determinants of CD4+ cell counts as well as the effect of the factors studied on patients CD4+ cell count were shown in this study.

The Generalized Linear Mixed-effects Model (GLMM) permitted regression analysis with correlated data and specified variance components that represent within-subject variance in outcomes and between-subject variation.

On the other hand, the Generalizing Estimating Equations (GEE) Model was fairly efficient once variance function was correctly specified and between-subject comparisons were nearly efficient once an independence covariance structure was used.

Both GEEs and GLMM hold promise when used with empirical variance estimates. The optimal approach will depend on study design and management goals.

6.2 Recommendation

This study is useful to guide education to the public, particularly patients, and also guide policy and management of treatment. Further studies are recommended to expand understanding and knowledge on the analysis of longitudinal data as well as include more covariates.

REFERENCES

A. Akinbami, O. Oshinaike, T. Adeyemo, A. Adediran, O. Dosunmu, et al., “Hematologic abnormalties in treatment-naïve HIV patients,” Infectious Diseases: Research and Treatment, vol. 3, pp. 45–49, 2010.

Aiuti F, Mezzaroma I: Failure to reconstitute CD4+ T cells despite suppression of HIV replication under HAART. AIDS Rev. 8(2), 88–97 (2006).

Bisson GP, Gross R, Strom JB et al.: Diagnostic accuracy of CD4 cell count increase for virologic response after initiating highly active antiretroviral therapy. AIDS 20(12), 1613–1619 (2006).

Badri M, Lawn SD, Wood R: Utility of CD4 cell counts for early prediction of virological failure during antiretroviral therapy in a resource-limited setting. BMC Infectious Disease.8, 89 (2008).

Begg MB, Parides MK. Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. Stat Med. 2003;22:2591–602. [PubMed]

Bhaumik DK, Roy A, Aryal S, Hur K, Duan N, et al. Sample size determination for studies with repeated continuous outcomes. Psychiatry Ann. 2009;38:765–71. [PMC free article] [PubMed]

Bock RD. Multivariate Statistical Methods in Behavioral Research. New York: McGraw-Hill; 1975.

Brian, W., Viviane, L. and Eleanor, G. (2011). Modeling the impact of Antiretroviral Therapy on the Epidemic of HIV: Curr HIV Research, 9 (6): 367-382

Bryk AS, Raudenbush SW. Hierarchical Linear Models: Applications and Data Analysis Methods. Newbury Park, CA: Sage; 1992.

Callens SF, Kitetele F, Lusiama J et al.: Computed CD4 percentage as a low-cost method for determining pediatric antiretroviral treatment eligibility. BMC Infect. Dis. 8, 31 (2008).

Chi EM, Reinsel GC. Models for longitudinal data with random effects and AR(1) errors. J Am Stat Soc. 1989;84:452–59.

Crum-Cianflone NF, Roediger M, Eberly LE, et al. Infectious Disease Clinical Research Program HIV Working Group. Obesity among HIV-infected persons: impact of weight on CD4 cell count. AIDS.2010;24:1069–1072. [PMC free article][PubMed]

Conaway MR. Analysis of repeated categorical measurements with conditional likelihood methods. J Am Stat Assoc. 1989;84:53–61.

Dang Q, Mazumbar S, Houck PR. Sample size and power calculations based on generalized linear mixed models with correlated binary outcomes. Compute Methods Programs Biomed. 2008;91: 122–27. [PMC free article] [PubMed]

Daka D, Loha E: Relationship between total lymphocyte count (TLC) and CD4 count among peoples living with HIV, southern Ethiopia: a retrospective evaluation. AIDS ResearchTher. 5, 26 (2008).

deLeeuw J, Kreft I. Random coefficient models for multilevel analysis. J Educ Stat. 1986;11:57–85.

Diggle PJ, Heagerty P, Liang K-Y, Zeger SL. Analysis of Longitudinal Data. 2 New York: Oxford Univ. Press; 2002.

F. Lutwama, R. Serwadda, H. Mayanja-Kizza et al., "Evaluation of dynabeads and cytospheres compared with flow cytometry to enumerate CD4+ T cells in HIV-infected ugandans on antiretroviral therapy," Journal of Acquired Immune Deficiency Syndromes, vol. 48, no. 3, pp. 297–303, 2008.

Fitzmaurice GM, Laird NM, Ware JH. Applied Longitudinal Analysis. New York: Wiley; 2004.

Gibbons RD. PhD thesis.Univ. Chicago; 1981.Trend in correlated proportions.

Gibbons RD, Bock RD. Trend in correlated proportions. Psychometrika. 1987;52:113–24.

Gitura B, Joshi MD, Lule GN, Anzala O: Total lymphocyte count as a surrogate marker for CD4+ T cell count in initiating antiretroviral therapy at Kenyatta National Hospital, Nairobi. East Afr. Med. J. 84(10), 466–472 (2007).

Goldstein H. Multilevel Statistical Models. 2 New York: Halstead Press; 1995.

Hedeker D, Gibbons RD. Longitudinal Data Analysis. New York: Wiley; 2006.

Hedeker D, Gibbons RD, Du Toit SHC, Patterson D. SuperMix—a program for mixed-effects regression models. Chicago: Sci. Software Int; 2008.

Hu FB, Goldberg J, Hedeker D, Flay BR, Pentz MA: Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes.

Jones CY, Hogan JW, Snyder B, et al. Overweight and human immunodeficiency virus (HIV) progression in women: associations HIV disease progression and changes in body mass index in women in the HIV Epidemiology Research Study cohort. Clin Infect Dis. 2003;37:S69–S80. [PubMed]

Keiser O, MacPhail P, BoulleA et al.: Accuracy of WHO CD4 cell count criteria for virological failure of antiretroviral therapy. Trop. Med. Int. Health 14(10), 1220–1225 (2009).

Laird, N. and Ware, J. (1982), ‘Random-effects models for longitudinal data’, Biometrics 38, 963-974.

Lederman M, McKinnis R, Kelleher D, et al. Cellular restoration in HIV infected persons treated with abacavir and a PI: age inversely predicts naive CD4 cell count increase. AIDS.2000;14:2635–2642. [PubMed]

Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika.1986; 73:13–22.

Little RJA.Modeling the drop-out mechanism in repeated-measures studies. J Am Stat Assoc. 1995; 90:1112–21.

Liu LC, Hedeker D. A mixed-effects regression model for longitudinal multivariate ordinal data.Biometrics.2006; 62:261–68. [PubMed]

Longford NT. A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects.Biometrika.1987; 74:817–27.

Marziali M, De Santis W, Carello R et al.: T-cell homeostasis alteration in HIV-1 infected subjects with low CD4 T-cell count despite undetectable virus load during HAART. AIDS 20(16), 2033–2041 (2006).

Mellors JW, Rinaldo CR Jr, Gupta P. White RM, Todd JA, Kingsley LA. Prognosis in HIV-1 infection predicted by the quantity of virus in plasma, Science 1996; 272; 1167-70.

Molenberghs GM, Thijs H, Jansen I, Beunckens C, Kenward MG, et al. Analyzing incomplete longitudinal clinical trial data. Biostatistics.2004; 5:445–64. [PubMed]

Neuhaus JM, Kalbfleisch JD, Hauck WW. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. Int Stat Rev. 1991; 59:25–35.

Nies-Kraske E, Schacker TW, Condoluci D et al.: Evaluation of the pathogenesis of decreasing CD4+ T cell counts in human immunodeficiency virus type 1-infected patients receiving successfully suppressive antiretroviral therapy. J. Infect. Dis. 199(11), 1648–1656 (2009).

Phillips AN, Pillay D, Miners AH, Bennett DE, Gilks CF, Lundgren JD: Outcomes from monitoring of patients on antiretroviral therapy in resource-limited settings with viral load, CD4 cell count, or clinical observation alone: a computer simulation model. Lancet 371(9622), 1443–1451 (2008).

Singer JD, Willett JB. Applied Longitudinal Data Analysis. New York: Oxford Univ. Press; 2003.

Verbeke G, Molenberghs G. Linear Mixed Models for Longitudinal Data. New York: Springer; 2000.

Viviane, D., Fink, L.V., Benita, Y., Robert, S.H., Harrigan, R. and Julio, S.G.M, (2009). Association between HIV-1 RNA Level and CD4+ Cell Count among untreated HIV-infected individuals, American Journal of Public Health Vol(99)51

WHO and UNAIDS: AIDS epidemic update. WHO UNAIDS (2009). http://data.unaids.org/pub/ EPISlides/2007/2007_epiupdate_en.pdf

Wolfinger RD. Covariance structure selection in general mixed models. Commun Stat Simulation Comput.1993; 22:1079–106.

Yari A, Passo FS, Yari V et al.: SMARThivCD4mos: a complexity-free and cost effective model technology for monitoring HIV patients CD4 number in resource-poor settings. Bioinformation 2(6), 257–259 (2008).

Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes.Biometrics.1986; 42:121–30. [PubMed]

Zeger SL, Liang KY, Albert PS. Models for longitudinal data: a generalized estimating equation approach. Biometrics.1988; 44:1049–60. [PubMed]