IJST EXPLORE International Journal of Science and Technology
CENTRE OF PROFESSIONAL RESEARCH PUBLICATIONS
ISSN 2049-7318
International Journal of Science and Technology Volume 6 No. 5, May, 2017

Longitudinal Study on Change in CD4 Cell Counts On HIV Positive Patients in Nigeria

Yahaya Haruna Umar, Samuel Olanrewaju O.
Department of Statistics, University of Abuja, Abuja- Nigeria

ABSTRACT

The purpose of this study is to examine and compare the utility of the GLMM models for modeling longitudinal data, with particular focus on the change in CD4 cell count in HIV patients started on Antiretroviral therapy(ART) in University of Abuja Teaching Hospital (UATH) Abuja-Nigeria. The determinants of CD4 cell counts as well as the effect of the factors studied on patients CD4+ cell count were shown in the study. The GLMM full model and the GLMM reduced model with the response variable CD4 counts at 36 month with the following predictor variables; time on ART, regimen, BMI, age and baseline CD4 count up to 33 month were fitted and analysis done with correlated data and specified variance components that represent within-subject variance in outcomes and between-subject variation. A comparison between the two models for Goodness of fit using Akaike's Information Criterion (AIC) showed that the full model has the best fit for CD4 count and it hold promise when used with empirical variance estimates. The optimal approach will depend on study design and management goals. 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

Keywords*: HIV, Cluster of Differentiation (**CD4**), GLMM, AIC, Comparison*

1. INTRODUCTION

The CD4 cell count is used in assessing the clinical status of HIV-infected individuals, in making informed decisions regarding the initiation of ART and in monitoring the success of such therapy. Several cohort studies and clinical trials have shown that CD4 count is the strongest predictor of subsequent disease progression and survival. The use of 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 relatively objective and simple marker to follow. Finally, the cost of CD4 counts has become more affordable, including in developing countries.

CD4 cell count and HIV Ribonucleic acid (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. While CD4 cell count is the strongest determinant of mortality in HIV-positive patients who adhere to ART, few studies conducted in rich settings have directly employed the interactive relationship between CD4 count, viral load, and time on treatment. The CD4 count and viral load 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.

2. GLOBAL PERSPECTIVE OF HIV/AIDS AND NIGERIAN SITUATION

Human Immunodeficiency virus (HIV) is a lent virus 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, 34million people are living with HIV in the world. Of these, approximately 23.5 million are in sub-Sahara 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 2.5 million 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.

3. CLUSTER OF DIFFERENTIATION 4

Cluster of differentiation (CD4) is a protein that is sometimes expressed on the surface of a class of immune system cells known as T-cells.

T-cells are specialized cells that recognize when other cells have been infected by viruses or other pathogens. In order to determine if a cell is infected, the CD4 positive T-cell looks to see if any of the pathogen protein fragments that it has learned to recognize are displayed on the cell's surface.

There are two subclasses of T-cells:CD4 cells and CD8 cells. CD8 cells are known as cytoxic (cells toxic) T-cells because they kill infected cells directly. In contrast, CD4 cells activate other cells to fight off an infection. CD4 cells are responsible for initiating the antibody response, and they can also tell cells known as macrophages to destroy any bacteria they are carrying.

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;

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

HIV causes AIDS by destroying CD4 cells **(Almonti et al, 2003)**. CD4 cell count, therefore, measures the degree of immune suppression in HIV-positive patients. There is inverse relationship between CD4 count and degree of immune suppression **(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 Nigeria, 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).**

4. MODELLING OF 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, favouring 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 inter individual 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).

5. GENERALIZED ESTIMATING EQUATION MODELS (GEE)

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)**. 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.

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

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 *yi* does not need to be specified. Instead, it is only the marginal distribution of *yij* 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 *yi1* and *yi2* are two univariate normals, rather than assuming that *yi1* and *yi2* 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, unbiased standard errors, even with incorrect specification of intra-cluster 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 correlation without 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:

a) The dependent variable is linearly related to the predictors (when the dependent variable is non-normally distributed a non-identity link function is to be selected);
b) The number of clusters be relatively high (a rule of thumb is no fewer than ten, possibly more than 30; Norton et al., 1996);
c) 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 miss-specified, 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).

6. GENERALIZED LINEAR MIXED EFFECTS MODEL (GLMMs)

Generalized Linear Mixed-effects regression models are now quite widely used for the analysis of longitudinal data (38 papers 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 ($i = 1, 2 \ldots N$ subjects) on occasion j ($j = 1, 2, \ldots n\ j$ occasions):

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \varepsilon_{ij}$$

That is;

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

Mixed models = fixed and random effects.

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 $\sigma 2$. 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 ($\beta 0$, the intercept or initial level, and $\beta 1$, 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.

7. METHODOLOGY

The researcher 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 viral load and CD4 count could be measured. A longitudinal study carried out from 2nd December, 2011 through 30th August, 2016. A total of 50 subjects were sampled. Gender, age at initiation of ART, Baseline weight, CD4 cell count (cells/mm3) taken at the initiation of ART and thereafter every 3 month CD4 cell count up to six months at the initiation of ART.

7.1 General Notation

To set a stage for the statistical discussion to follow, it is helpful to present a unified notation for the various aspects of the longitudinal design. We index the N subjects in the longitudinal study as

i= 1,…., N subjects.

For a balanced design in which all subjects have complete data, and are measured on the same occasions, we index the measurement occasion as

j= 1,….,n observations.

Or in the unbalanced case of unequal numbers of measurements or different time-points for different subjects

j= 1,…., n_1 observations for subject i

The total number of observations are given by

$$\sum_{1}^{n} ni$$

7.2. 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 by 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.

8. DATA LAYOUT

In fixing ideas for the statistical development to follow, it is also useful to apply this previously described notation to describe a longitudinal dataset as follows.

TABLE 1: 2-LEVEL DESIGN IN MULTI LEVEL (GOLDSTEIN 1995)

Subject	Observation	Response	Covariates
1	1	y_{11}	x_{111} … x_{11p}
1	2	y_{12}	x_{121} … x_{12p}
.	.	.	. … x_1n_{1p}
1	n_1	y_1n_1	x_1n_{11} … x_1n_{1p}
.	.	.	. …
.	.	.	. …
.	.	.	. …
N	Nn	y_Nn_N	x_Nn_{N1}… x_Nn_{NP}

In this univariate design, n_i varies by subject and so the number of data lines per subject can vary. In terms of the covariates, if x_r is time-invariate (i.e., a between-subjects variable) then, for a given subject i, the covariate values are the same across time, namely $x_{i1r}=x_{i2r}=x_{i3r}=\ldots=x_{in1r}$.

The above layout depicts what is called 2- level design in the multilevel **(Goldstein 1995)** and hierarchical linear modeling **(Raudenbush and Bryk, 2002)** literatures, namely, repeated observations at level 1 are nested within subjects at level 2. In some cases, subject themselves are nested within sites, hospitals, clinics, workplaces, etc. in this case, design has three levels with level-2 subjects nested within level-3 sites. This study primarily focuses on 2-level design and models

8.1. Analysis Considerations

There are several different features of longitudinal studies that must be considered when selecting an appropriate longitudinal analysis. First, there is the form of the outcome or response measure. If the outcome of interest is continuous and normally distributed, much simpler analysis is usually possible, (e.g., a mixed effects linear regression model). By contrast, if the outcome is continuous but does not have a normal distribution, (e.g., a count) the alternative nonlinear models (e.g., a mixed-effects poison regression model) can be considered for qualitative outcomes, such as binary (yes or no).

Second, the number of subject N is an important consideration for selecting a longitudinal analysis method. The more advanced models (e.g., generalized mixed-effects regression models) that are appropriate for analysis of unbalanced longitudinal data are based on large sample theory and may be inappropriate for analysis of small N studies (e.g., N<50).

Third, the number of observations per subject n_i is also an important consideration when selecting an analytic method. For n_i = 2 for all subjects, a simple change score can be computed and the data can be analyzed using methods for cross-sectional data, such as ANCOVA. When n_i =n for all subjects, the design is to be balanced, and traditional ANOVA models for repeated measurements (i.e., traditional mixed-effects models or multivariate growth curve models) can be used. In the most general case where n_i varies from subject to subject, more general methods are required (e.g., generalized mixed-effects regression models), which are the primary focus of this research.

Fourth, the number and type of covariates is an important consideration for model selection for $E(y_i)$. in the one sample case, we may only have interest in characterizing the rate of change in the population over time. Here, we can use a random-effect regression model, where the parameters of the growth curve are treated as random effects and allowed to vary from subject to subject. In the multiple sample case (e.g., comparison of one or more treatment conditions to control), the model consists of one or more categorical covariates that contrast the various treatment conditions in the design. In the regression case, we may have a mixture of continuous and/or categorical covariate, such as age, sex and race. When the covariates take on time-specific values, (i.e. time-varying covariates), the statistical model must be capable of handling these as well.

Fifth, selection of a plausible variance-covariance structure for the $V(y_i)$ is of crucial importance. Different model specifications lead to (a) homogeneous or heterogeneous variance and/or (b) homogeneous or heterogeneous co variances of the repeated measurements over time. Furthermore, residual autocorrelation among the responses may also play a role in modelling the variance-covariance structure of the data.

Each of these factors is important for selecting an appropriate analytical model for analysis of a particular set of longitudinal data.

8.2 General Approaches

There are several different general approaches to the analysis of longitudinal data.

The first approach, which we refer to as the "derived variable" approach, involves the reduction of the repeated measurements into a summary variable. In fact once reduced, this approach is strictly not longitudinal, since there is only a single measurement per subject. Perhaps the earliest example of the analysis of longitudinal data was presented by **Student (1908)** in his illustration of the t-test. The objective of the study **(Cushny and Peebles, 1905**) was to determine changes in sleep as a function of treatment with the hypnotic drug scopolamine. Although hours of sleep were carefully measured by the investigators, day-to-day variability presented statistical challenges in detecting the drug effect using large sample methods available at the time.

Second, generalized mixed-effects regression models, which form the primary emphasis of this research, are now quite widely used for analysis of longitudinal data. These models can be applied to normally distribute continuous outcomes as well as categorical outcomes and other non-normally distributed outcomes such as counts that have a Poisson distribution. Mixed-effects regression models are quite robust to missing data and irregularly spaced measurement occasions and can easily handle both time-invariant and time-varying covariates. As such, they are sometimes called "full-likelihood" methods because they make full use of all available data from each subject. The advantage is that missing data are ignorable if the missing responses from a given subject. The disadvantage is that full-likelihood methods are more computationally complex than quasi-likelihood methods, such as generalized estimating equations (GEE).

8.3. ASSUMPTIONS IN GENERALIZED ESTIMATING EQUATIONS AND GENERALIZED LINEAR MODELS

1) Generalized estimating equations and generalized linear models do not assume that the dependent variables are not normally distributed.
2) Generalized estimating equations and generalized linear models neither assumes linearity between the predictors and the dependent variables, nor homogeneity of variance for range of the dependent variable
3) There must be linearity in the link function.
4) Multi-co linearity is absent
5) The data must be centred in order to reduce multi-co linearity
6) The dependent data in GEE and GLM are continuous, ordinal, or binary

9. RESULTS AND DISCUSSION

The results describe information on the subjects under study and the changes in the variable of interest by performing longitudinal analysis GLMM.

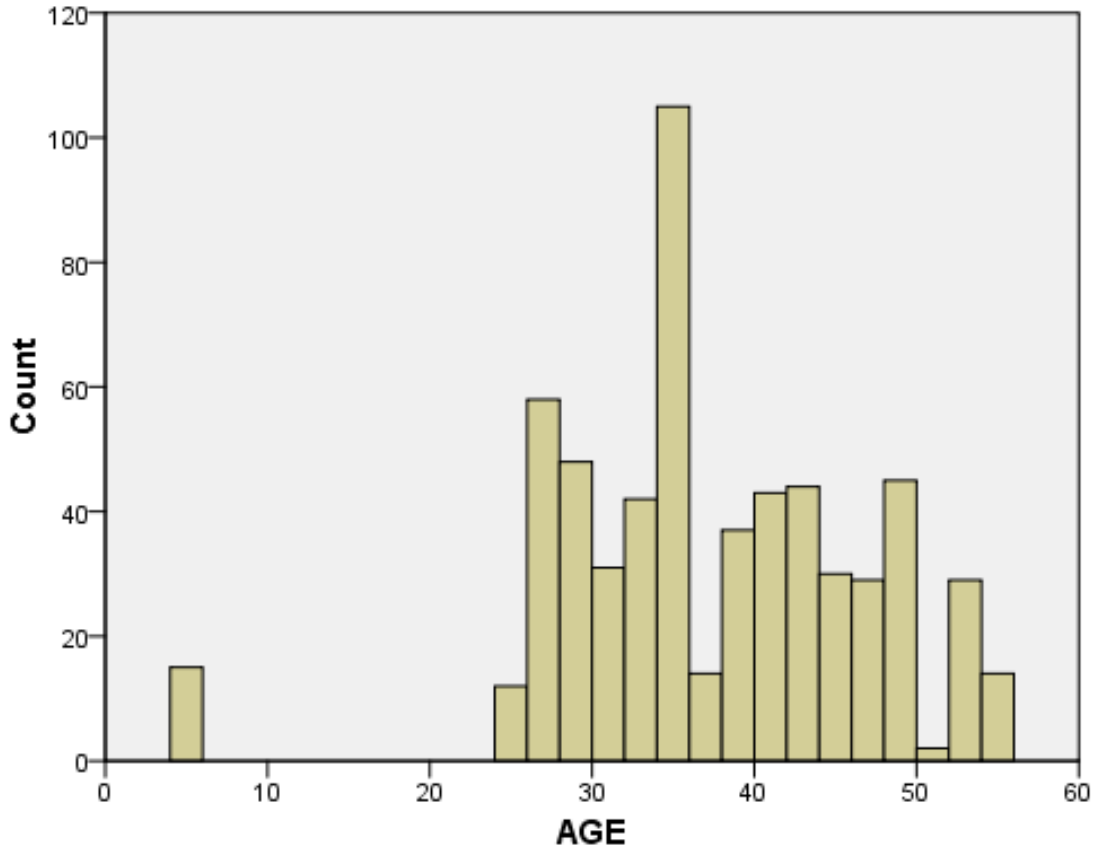

Figure 1: Bar Chart of Age Characteristics

The bar chart indicates the age distribution of the 50 patients initiated on ART, with the highest number of patients at 36 years of age and the lowest number at 52 years of age.

9.1 Data Analysis

Table 2: Descriptive statistics on Age and BMI

	N	Minimum	Maximum	Mean	Std. Deviation
AGE	50	5	54	36.82	9.51309
BMI	50	13.1	51.9	24.98	7.50259
Total	50				

Source: SPSS OUTPUT

Table2 indicates a total of 50 patients observed. The mean age was 36.8211 and SD was 9.51309, also the BMI averaged 24.9796 and SD 7.50259 at initiation of Antiretroviral Therapy.

Table 3: CD4 cell count characteristics

	N	Minimum	Maximum	Mean	Std. Deviation
cd4_initial	50	27	1802	298.16	311.62722
cd4_3	47	27	1802	300.15	313.55958
cd4_6	46	27	1802	319.85	315.55073
cd4_9	45	27	1802	334.42	327.55704
cd4_12	43	27	1435	337.6	296.0958
cd4_15	43	29	1435	369.53	327.0657
cd4_18	39	29	1435	372.97	318.98164
cd4_21	38	29	1435	364.89	276.82015
cd4_24	37	8	1435	365.49	269.523
cd4_27	33	8	1435	379.06	283.43881
cd4_30	30	105	1697	406.63	306.30675
cd4_33	25	105	1697	412.96	347.02575
cd4_36	25	105	1697	422.56	343.27189
Total	25				

Source: SPSS OUTPUT

Table 3 depicts the CD4 cell counts measured from the start of ART up to the 36th month of treatment. The mean CD4 cell counts over each time period is as indicated, as well as the standard deviation (SD) for each mean CD4 cell count.

Table 4: Regimen Characteristics

Regimen type	N	Percent	Cumulative Percent
3TC/D4T/EFV	10	21	21.1
3TC/D4T30/NVP	13	23.4	44.5
3TC/D4T40/NVP	20	40.4	84.9
AZT/3TC/NVP	7	15	100
Total	50	100	

Source: SPSS OUTPUT

Table 4 shows that all patients were on a triple therapy. 63.9% of the subjects were on first line regimen based on 2 nucleoside reverse transcriptase inhibitors (NRTIs) and a non-nucleoside reverse transcriptase inhibitors (NNRTI) i.e. 3TC/D4T(30/40)/NVP, while 10 patients were on 3TC/D4T/EFV a regimen given to patients on Tuberculosis treatment. 7 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).

Table 5: Time on ART (months) characteristics

Class Interval	Frequency	Valid Percent	Cumulative Percent
0 – 12	237	39.7	39.7
13 – 25	163	27.3	67
26 -38	96	21.1	88.1
39 -51	68	11.4	99.5
52 – 64	4	0.7	100
Total	598	100	

Source: SPSS OUTPUT

Table 5 indicates the time distribution of the 50 patients initiated on ART, with the highest frequency of patient's visit (237) at class interval (0 – 12) i.e. from the month of initiation of ART to the 12th month and the lowest frequency of patients' visit (4) at class interval (52 – 64) i.e. from the month 52nd and 64th.

Table 6: Correlation matrix on Repeated CD4+ cell counts

	CD4_at0	CD4_at3	CD4_at6	CD4_at9	CD4_at12	CD4_at15	CD4_at18	CD4_at21	CD4_at24	CD4_at27	CD4_at30	CD4_at33	CD4_at36
CD4_at0	1	.990**	.968**	.952**	.880**	.859**	.799**	.807**	.819**	.810**	.872**	.875**	.849**
CD4_at3		1	.978**	.962**	.883**	.852**	.795**	.803**	.814**	.805**	.868**	.870**	.845**
CD4_at6			1	.983**	.907**	.873**	.820**	.833**	.831**	.835**	.867**	.877**	.852**
CD4_at9				1	.928**	.887**	.817**	.830**	.827**	.831**	.863**	.877**	.850**
CD4_at12					1	.953**	.897**	.826**	.817**	.847**	.821**	.848**	.818**
CD4_at15						1	.949**	.724**	.714**	.719**	.683**	.892**	.865**
CD4_at18							1	.715**	.706**	.703**	.665**	.872**	.843**
CD4_at21								1	.995**	.983**	.934**	.661**	.639**
CD4_at24									1	.977**	.944**	.677**	.655**
CD4_at27										1	.944**	.678**	.655**
CD4_at30											1	.734**	.713**
CD4_at33												1	.983**
CD4_at36													1

**. Correlation is significant at the 0.01 level (2-tailed).

Source: SPSS OUTPUT

The correlation structure was of the form illustrated in table 6. Correlation was significant at the 0.01 level (2-tailed) and 0.05 level (2-tailed). All correlations are estimated separately.

It can be generalized that 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. This implies that the CD4 cell counts are dependent at each level with time.

GENERALIZED LINEAR MIXED MODEL

Table 7: Model 1(Full model)

Scaled residuals:				
Min	1 Q	Median	3 Q	Max
-1.27591	-0.22655	-0.01243	0.21805	1.51593

Source: R-CONSOL output

Linear mixed model fit by REML ['lmerMod']
Formula: CD4.at.36 ~ AGE + BMI + Regimen + Time.on.ART + CD4.at.0 + CD4.at.3 + CD4.at.6 + CD4.at.9 + CD4.at.12 + CD4.at.15 + CD4.at.18 + CD4.at.21 + CD4.at.24 + CD4.at.27 + CD4.at.30 + CD4.at.33 + (1 | Study)
REML criterion at convergence: 254

Table 7.1: Random effects

Random effects:		
Groups Name	Variance	Std.Dev.
Study (Intercept)	3.040e-11	5.513e-06
Residual	7.281e+03	8.533e+01
Number of obs: 25, groups: Study, 4		

Source: R-CONSOL output

Table 7.2: Fixed effects

Fixed effects:			
	Estimate	Std. Error	t value
(Intercept)	-59.5272	117.6655 -0.506	
AGE	-1.2291	5.8501	-0.210
BMI	0.1676	14.8358	0.011
RegimenAZT/3TC/NVP	34.4515	141.3098	0.244
Time.on.ART	2.8610	2.0554	1.392
CD4.at.0	1.7313	1.5621	1.108
CD4.at.3	-1.4493	1.5407	-0.941
CD4.at.6	0.4659	0.8110	0.575
CD4.at.9	-1.0400	0.8178	-1.272
CD4.at.12	1.1162	0.8287	1.347
CD4.at.15	-1.0523	0.7785	-1.352
CD4.at.18	-0.1275	0.2358	-0.541
CD4.at21	6.5994	4.7103	1.401
CD4.at.24	-6.2678	4.7052	-1.332
CD4.at.27	-0.2051	0.4785	-0.429
CD4.at.30	-0.5378	0.4490	-1.198
CD4.at.33	1.8332	0.5085	3.605

Source: R-CONSOL output

Table 8:Model2 (Reduced model)

Scaled residuals:				
Min	1Q	Median	3Q	Max
-1.04703	-0.35225	-0.00959	0.07004	2.22801

Source: R-CONSOL output

Linear mixed model fit by REML ['lmerMod']
Formula: CD4.at.36 ~ CD4.at.0 + CD4.at.3 + CD4.at.6 + CD4.at.9 + CD4.at.12 + CD4.at.15 + CD4.at.18 + CD4.at21 + CD4.at.24 + CD4.at.27 + CD4.at.30 + CD4.at.33 + (1 | Study)
REML criterion at convergence: 280.9

Table 8.1: Random effects

Random effects:		
Groups Name	Variance	Std.Dev.
Study (Intercept)	0	0.00
Residual	7316	85.54
Number of obs: 25, groups: Study, 4		

Source: R-CONSOL output

Table 8.2: Fixed effects

Fixed effects:			
	Estimate	Std. Error	t value
(Intercept)	16.85873	37.38111	0.451
CD4.at.0	0.04795	1.26123	0.038
CD4.at.3	-0.05653	1.34977	-0.042
CD4.at.6	0.30052	0.78225	0.384
CD4.at.9	-0.41369	0.71095	-0.582
CD4.at.12	0.15827	0.64716	0.245
CD4.at.15	-0.14117	0.56354	-0.250
CD4.at.18	-0.12653	0.18670	-0.678
CD4.at21	2.19737	4.07594	0.539
CD4.at.24	-2.09770	4.12327	-0.509
CD4.at.27	-0.04300	0.38959	-0.110
CD4.at.30	-0.12532	0.38548	-0.325
CD4.at.33	1.26998	0.40593	3.129

Source: R-CONSOL output

10. Models Comparison of Goodness of Fit (Using AIC)

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).

10.1. Hypothesis

H_0: full model is not a better fit for change in CD4 cell count

H_1: full model is a better fit for change in CD4 cell count

10.2 Decision Rule

Reject the null hypothesis (H0) if P-Value < α (0.05)

Models:
glmm reduced model: CD4.at.36 ~ CD4.at.0 + CD4.at.3 + CD4.at.6 + CD4.at.9 + CD4.at.12 + CD4.at.15 + CD4.at.18 + CD4.at21 + CD4.at.24 + CD4.at.27 + CD4.at.30 + CD4.at.33 + (1 | Study)

glmm full model: CD4.at.36 ~ AGE + BMI + Regimen + Time.on.ART + CD4.at.0 + CD4.at.3 + CD4.at.6 + CD4.at.9 + CD4.at.12 + CD4.at.15 + CD4.at.18 + CD4.at21 + CD4.at.24 + CD4.at.27 + CD4.at.30 + CD4.at.33 + (1 | Study)

Table 9: Model Comparison

	Df	AIC	BIC	log Lik	deviance	Chisq Chi	Df	Pr(>Chisq)
glmm.model2	15	305.04	323.33	-137.52	275.04			
glmm.model	19	302.79	325.94	-132.39	264.79	10.259	4	0.03628 *

Table 9 shows the P-Value (0.03628) < α (0.05), thus we reject the null hypothesis that the full model is not a good fit for change in CD4 cell count. Also from the AIC statistics the result indicates that the full model has a lower AIC value (302.79). Hence it can be concluded that the full model is a better fit for change in CD4 cell count, thus considering all the variables will make the prediction of CD4 cell count to be more precise.

11. CONCLUSIONS AND RECOMMENDATIONS

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.

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.

A comparison between the two models for Goodness of fit using Akaike's Information Criterion (AIC) showed that the full model was the best fit for CD4 count and it hold promise when used with empirical variance estimates. The optimal approach will depend on study design and management goals.

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.

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.

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.

Bock RD. Multivariate Statistical Methods in Behavioral Research. New York: McGrawHill; 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.

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.

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.

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.
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.

Neuhaus JM, Kalbfleisch JD, Hauck WW. A comparison of cluster-specific and populationaveraged 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 HIVinfected 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.

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

