DYNAMIC MODELS FOR TIME-VARYING OUTCOMES: AN APPLICATION TO THE 2015-2017 PATIENT COHORT ON ANTIRETROVIRAL THERAPY AT LUDERITZ HOSPITAL, NAMIBIA

A MINI-THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN BIOSTATISTICS OF THE UNIVERSITY OF NAMIBIA

BY

LINEEKELA GABRIEL

201203315

MAY 2020

SUPERVISOR: PROF L. KAZEMBE (DEPARTMENT OF STATISTICS AND POPULATION STUDIES, UNAM)

Abstract

Patients' adherence to a prescribed medication regimen is one of the most significant barriers to successful antiretroviral therapy (ART). In addition, adherence to ART is one of the key determinants of Human Immunodeficiency Virus (HIV) disease progression, while non-adherence severely compromises treatment effectiveness and leads to unsuppressed virus. The extent of the impact of poor adherence on resulting health measures is often unknown, and typical analyses ignore the time-varying nature of adherence. The main objective of this study was to model time-varying outcomes of patients, while accounting for data missingness and measurement error using dynamic models, with application to a cohort of 154 adult patients initiated on ART between January 2015 and December 2017 at the Luderitz hospital. The outcome variable of this study was viral load which was measured at scheduled follow-up visits of patients. Baseline CD4 count, baseline weight, age at start of ART and gender were the non-dynamic covariates which were measured at the ART initiation, while adherence to ART and weight at follow up were the dynamic covariates measured at follow up visits. This study used mixed effects model and Generalized Estimating Equations (GEE) to model longitudinally measured viral load as a function of the dynamic as well as non-dynamic covariates. To account for missingness in the outcome variable as well as potential measurement error in covariates, a Simulation Extrapolation Inverse Probability Weighted Generalized Estimating Equations (SWGEE) model that incorporates missing and measurement error was used to model the data. The study found that adherence was good in female patients as compared to male patients. Furthermore, the study found that patients with a good adherence rate achieved viral suppression within 12 months of treatment unlike non-adherent patients. In conclusion, viral load of patient's on ART differ across the patients' baseline demographic and clinical characteristics.

Keywords: antiretroviral therapy, dynamic linear models, time-varying outcomes, viral load, dynamic covariates, non-dynamic covariates

List of Publication(s)/Conference(s) proceedings

- SUSAN-SSACAB Conference 8-11 September 2019, Cape Town -South Africa

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AIC	Akaike Information Criterion
AIDS	Acquired immunodeficiency syndrome
ART	Antiretroviral Therapy
ARV	Antiretroviral
BIC	Bayesian Information Criterion
CDC	Centre for Disease Control
DLMs	Dynamic Linear Models
GEE	Generalized Estimating Equations
GLM	Generalized Linear Models
HAART	Highly Active Antiretroviral Therapy
HIV	Human Immunodeficiency Virus
ICAP	International Center for AIDS Care and Treatment Programs
LME	Linear Mixed Effects
LOQ	Limit of Quantification
MAR	Missing at Random
MCAR	Missing Completely at Random
MCMC	Markov Chain Monte Carlo

MEMS	Medication Event Monitoring System
MLIC	Missing Longitudinal Information Criterion
MNAR	Missing not at Random
MoHSS	Ministry of Health and Social Services
MRM	Mixed Effects Regression Model
NAMPHIA	Namibia Population-Based HIV Impact Assessment
NLME	Non-Linear Mixed Effects
ODE	Ordinary Differential Equations
PLHIV	People Living with HIV
QIC	Quasi Information Criterion
SAEM	Stochastic Approximation Elimination Method
SWGEE	Simulation Extrapolation Inverse Probability Weighted Generalized Estimating Equations
UNAIDS	United Nations Program on HIV and AIDS
VL	Viral Load
VLS	Viral Load Suppression
WHO	World Health Organization

Acknowledgements

I would like to acknowledge my supervisor, Prof. Lawrence Kazembe for his guidance, input and support during the compilation of this thesis. My special thanks go to my best friend Dr. Josua Mwanyekange for his invaluable contribution, motivation, and encouragement during my studies. I am also highly indebted to my supporting classmates Anna Onesmus and Sesilia Kapenda.

This study was supported through the DELTAS Africa Initiative SSACAB (Grant No. 107754/Z/15/Z). The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS) Alliance for Accelerating Excellence in Science in Africa (AESA) and is supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust (Grant No. 107754/Z/15/Z) and the UK government. The views expressed in this publication are those of the author and not necessarily those of the AAS, NEPAD Agency, Wellcome Trust, or the UK government.

Declarations

I, Lineekela Gabriel, hereby declare that this study is my own work and is a true reflection of my research, and that this work, or any part thereof has not been submitted for a degree at any other institution.

……..…… …… ...…..……

Name of Student Signature Date

CHAPTER 1 : INTRODUCTION

1.1 Background of the study

Human Immunodeficiency Virus (HIV) is the virus that can lead to acquired immunodeficiency syndrome (AIDS) if not treated, and it is a lifetime infection since the body is not capable of getting rid of it. HIV/AIDS is a global pandemic (Cohen, Hellmann, Levy, DeCock and Lange, 2008). Approximately 37.9 million people were infected globally in 2018, of which 1.7 million (4.5%) were children less than 15 years, and 57% were men (United Nations Program on HIV and AIDS [UNAIDS] (2019, p.16)). UNAIDS (2019) also indicated that there were 1.7 million new infections in 2018 and about 770,000 deaths from AIDS worldwide.

Moreover, Sub- Sahara Africa is the region most affected, with an estimated 61% of new HIV infections that occurred in the region in 2018 (UNAIDS, 2019). Namibia is one of the countries with a high infection rate in the sub-continent. Based on the UNAIDS annual report of 2015, the first case of HIV in Namibia was reported in 1986 and since then the prevalence has continued to rise and reached a peak of 22% in 2002. In 2016 about 220, 000 (10%) Namibians were living with HIV (UNAIDS, 2019). The number of new cases for HIV in Namibia slowly declined with only 11,000 new infections reported in 2014, compared to the year 2000 when there were 21,000 new cases of HIV annually (MohSS, 2016).

There is currently no effective cure that exists, but with proper medical care, HIV can be controlled with antiretroviral therapy (ART) if taken the right way every day. Antiretroviral medicine can

prolong the lives of people infected with HIV, keep them healthy, and greatly lower their chance of infecting others (Centre for Disease Control [CDC], 2019). The Government of Namibia is committed to controlling the epidemic (MoHSS, 2016, p.2). According to the Directorate for Special Disease of Namibia, in 2016 general ART coverage was estimated at 76%, while paediatric coverage was estimated at 95%.

In 2015, UNAIDS developed the 90-90-90 targets to be achieved by 2020, with the aim of ending the epidemic by 2030 (UNAIDS, 2017). The Ministry of Health and Social Services (MoHSS) in their national guidelines for ART defined the 90-90-90 goals as follows: at least 90% of all people living with HIV should know their HIV status, at least 90% of all people diagnosed with HIV infection will receive sustained antiretroviral therapy and at least 90% of all people receiving antiretroviral therapy will have viral suppression (MoHSS, 2016, p.2). Consequent to these goals, viral load monitoring in People Living with HIV (PLHIV) has become a crucial activity in health facilities offering ART across the country in order to assess the progress of the last 90 of the 90-90-90 goal. The Namibia Population-Based HIV Impact Assessment (NAMPHIA) of 2017 reported that the prevalence of viral load suppression (VLS) among HIV-positive adults aged 15-64 years on ART in Namibia was 77.4%: 81.7% among females and 69.6% among males (NAMPHIA, 2018). This is slightly lower than the target of 90%. The low VLS is mostly attributed to poor treatment adherence (Achappa et al. 2013; Kim et al. 2018).

Achappa et al. (2013) defined adherence as a patient's ability to follow a treatment plan, take medications at prescribed times and frequencies, and follow restrictions regarding food and other

medications. An adherence to ART of 95% is required as an appropriate level to achieve maximal viral suppression and lower the rate of opportunistic infections (Kim et al., 2018). Adherence to ART is measured through pill counting, that is, at follow-up, if a patient has the correct number of (leftover) pills then they are fully adherent otherwise they are not adherent. Adherence is promoted by proper ongoing support and counseling as well as prescribing simplified and well-tolerated regimens involving as few pills as possible, administered no more than two times per day (MoHSS, 2016, p.20).

HIV virulence trends are estimated using set point viral load which is the concentration of HIV RNA copies in blood and reflects the ongoing virus replication in a person's body (Herbeck et al., 2014). Since viral load is a more sensitive and an earlier indicator of treatment failure, Namibia has transitioned to routine viral load monitoring rather than the CD4 count for treatment monitoring. However, if a patient has virologic failure or shows signs of clinical deterioration, a CD4 count should be done (MoHSS, 2016, p.32). Viral load is an excellent means of monitoring treatment response because ART prevents HIV replication by inhibiting viral enzymes causing the viral load to decline (International Center for AIDS Care and Treatment Programs [ICAP], 2016, p.5). One of the challenges in monitoring HIV treatment is the time-varying pattern of health outcomes and data missingness and measurement error in key variables.

Statistical models are being used to understand the transmission and progression of infections and evaluation of the potential impact of control programs in reducing morbidity and mortality (Turner, 2011). Both static and dynamic models have been used for monitoring health outcomes. Static modeling is an approach to modeling of a problem based on the state at a fixed point in time

whereas dynamic models are general models that contain or depend upon an element of time, especially allowing for interactions between variables, and are used to express and model the behavior of the system over time (Daly et al, 2008). Furthermore, static and individual based models excel at projecting clinical events over the lifetime of unique patients who retain their clinical trajectory history whereas dynamic and population-based models are most often used to model transmission and to project population-level changes in incidence and prevalence over long horizons (Jacobsen & Walesnsky, 2016).

Dynamic models are categorized into linear and nonlinear/non-normal models. Dynamic linear models (DLM) are parametric in that the parameter variation and available information are described probabilistically. They can be seen as a generalization of the regression models allowing for changes in parameters values throughout time (Migon et al., 2005). Campos, Glickman & Hunters (2018) developed a modeling framework for longitudinally recorded health measures which was modeled as a function of time-varying adherence to medication or other time-varying covariates and relied on normal Bayesian dynamic linear models (BDLMs), accounts for time-varying covariates and non-dynamic covariates. Their model can be used to forecast health outcomes as a function of specified patterns of adherence and provides a robust framework for understanding the impacts of poor medication adherence.

With the growing use of longitudinal data in medical health and social science researches, methods of analysis of such data need to be more clearly understood. As one would expect, such datasets are characterized by the fact that repeated observations for subject are correlated. For this reason, this study also touched on the issues of model estimation and model selection for longitudinal data

with time-varying covariates. In addition, missingness in longitudinal data sets may give wrong inferences if not properly accounted for. Therefore, correct statistical analysis of such data requires the modelling of correlation and accounting for the missingness and covariate measurement error.

1.2 Statement of the problem

HIV is a lifetime infection that is currently only controlled through ART, and Namibia has one of highest HIV prevalence in the world. A great commitment to ART is required to help suppress the virus in the body, reduce transmission and boost the immune system. The change in viral load over time is a good indicator of treatment effectiveness, while HIV-load decrease and suppression over time for patients on ART are associated with consistent adherence to ART (Chendi et al., 2019). Adherence changes over time but typical analyses of health outcomes ignore the time-varying nature of adherence to medication and instead examine the relationship between adherence and outcomes via correlations with non-dynamic adherence measures (Campos et al., 2018). Therefore, there is need for analyzing time-varying treatment adherence in association with time-varying outcomes. An analysis of HIV viral load dynamics becomes very important as it provides additional information regarding the prognosis of the disease. Furthermore, viral load in HIV infected persons are measured with errors due to disease progressions, illnesses, and lifestyle factors. Failure to appropriately account for the extent of missingness and measurement errors may lead to biased results. Therefore, there is a need to consider models that allows for the analysis of viral load with missingness and measurement error.

1.3 Objectives of the study

1.3.1 Main objective

The main objective of this study was to model time-varying outcomes of patients, while accounting for data missingness and measurement error using dynamic models, with application to ART data obtained from the Luderitz hospital in Namibia.

1.3.2 Specific objectives

a) To explore the average change of HIV viral load in patients on ART over time
b) To model the change in viral load over time using Mixed effects models and Generalized estimating equations.
c) To investigate the effects of clinical factors and demographic characteristics on viral load.
d) To model viral load longitudinal data adjusting for the bias induced by measurement error in covariates as well as missingness in response variable.

1.4 Significance of the study

According to the Namibia Factsheet by Centre for Disease Control (CDC) of 2017, Namibia has one of the world's highest HIV prevalence rates of 13.8% and HIV is still classified and treated as a special disease. The findings of this study could assist health care workers (HCWs) in developing innovative approaches to ensure full adherence throughout ART. It could also encourage patients to remain fully adherent to their treatment and HCWs to maintain adherence counseling for the success of ART which is viral suppression, zero transmission and zero new infections. Moreover, the findings of this study could help program planners, decision makers and project implementers in fighting the virus and in achieving the 90-90-90 goals.

1.5 Organization of the thesis

The rest of this mini-thesis is organized as follows: Chapter 2 provides the review of dynamic models of HIV, while chapter 3 comprises of the study design and exploratory data analysis. In chapter 4 data was modeled using Generalized estimating equations (GEE) and Mixed effects models (MEM) methods. Chapter 5 extends the models to incorporate responses with missingness and covariate with measurement error. Chapter 6 presents the discussion and conclusion.

CHAPTER 2 : LITERATURE REVIEW

2.1. Review of HIV Models

2.1.1. The Dynamic Models of HIV

Interest in HIV studies have increased in recent literature with great attention drawn from AIDS epidemiology and clinical trials. According to Wu (2005), even though the important findings from HIV dynamic studies have been well presented in modern journals, statistical methods for estimating viral load dynamic parameters have not been well discussed by researchers in this area. Clinicians usually attempt to understand the HIV dynamics by making use of AIDS clinical trials. These clinical trials have considerably improved the knowledge of the pathogenesis of HIV infection and guided for the treatment of HIV patients and evaluation of antiretroviral (ARV) therapies (Huang, Liu, & Wu, 2006).

The analyses of HIV dynamics are often made using models derived through a system of differential equations (ODE) in which interaction of CD4 cells and virus is described. However, statistical inferences based on such models present computational difficulties due to problems with ODE numerical solutions and statistical algorithms. Dynamic linear and non-linear mixed effects (LME/NLME) models are popular in modeling HIV dynamics (rate of changes in viral load) as both include confounding interaction of adherence, time varying, invariant covariates as well as random effects (Yangxin et al., 2014).

2.1.2. Dynamic models with time varying covariates

Since patients' characteristics are often recorded repeatedly over a given period of treatment routine, the measurements collected from the same subject for that period may be correlated while measurements collected from different patients can be assumed to be independent. One of the powerful tools used to model HIV dynamics is linear mixed effects (LME) model introduced by (Laird & Ware, 1982). In this model, time varying covariates as well as both within-subject and between-subject dependency are considered.

Time-varying covariance occurs when a given covariate changes over time during the follow-up period, which is a common phenomenon in clinical trials (Zhang et al., 2018). According to Huang et al. (2003), in most models for HIV dynamic, the treatment effect is assumed to be constant. However, the effect of treatment appears to change over time, probably due to pharmacokinetic variation, fluctuating adherence, the emergence of drug resistant mutations and/or other factors. To better model the actual antiviral responses, they extended previous work to include time varying covariates. They also developed a viral dynamic model to evaluate antiviral response as a function of time-varying concentrations of drug in plasma, and changes over time in phenotypic sensitivity of the virus. Some authors such as Ho et al. (1995); Polis et al. (2001) and Wei et al. (1995) have used a linear mixed-effects model to fit viral dynamic data from the first several days.

Yangxin et al. (2014) analyzed AIDS longitudinal data using nonlinear mixed-effects (NLME) models for HIV dynamics. The results suggest that modeling HIV dynamics and virologic responses with consideration of covariate measurement error and time-varying clinical factors may be important for HIV/AIDS studies in providing quantitative guidance to better understand the virologic responses to antiretroviral treatment and to help evaluation of clinical trial design in

existing therapies. More recent comprehensive review on NLME models can be found in Huang, Wu, Holden-Wiltse & Acosta (2011); Gagne & Huang (2012); Bryan & Heagety (2014).

A study by Huang et al. (2011) investigated the effects of several summary determinants of Medication Event Monitoring System (MEMS) adherence rates on virologic response measured repeatedly over time in HIV-infected patients. They established a mechanism-based differential equation model with consideration of adherence to medication, interacted by virus susceptibility to drug and baseline characteristics, to characterize the long-term virologic responses after initiation of therapy. They found that the baseline viral load had a positive effect on drug efficacy, while the baseline CD4 cell count had a negative effect on it.

2.2 Model definitions

2.2.1 Linear Mixed Effects Model

Linear mixed effects models have become very popular tools for analysis of longitudinal data because they are very flexible and applicable (Van et al., 2010). These models are developed with the idea that individuals in the population have their own mean response profile and subject specific effects over time. According to Laird and Ware (1982), Molenberghs and Verbeke (2001), the longitudinal measurements are fit by using a regression model that allows parameters to vary among subjects. To present linear mixed effects model, let the j^{th} observed longitudinal outcome y_{ij} of object i $(i = 1, \dots, n_i)$ measured at time t_{ij} satisfies $\tilde{b}_{i0} + \tilde{b}_{i1} t_{ij} + \epsilon_{ij}$.

One can redefine y_{ij} by taking $\tilde{b}_i = \left(\tilde{b}_{i0}, \tilde{b}_{i1}\right)'$ as a vector of subject specific parameters to be bivariate normal with mean $(\beta_0, \beta_1)'$ and $p \times p$ variance covariance matrix D and also ϵ_{ij} assumed to be normal distributed with mean zero and variance σ_ϵ^2. Hence, the linear model is given by

$$y_{ij} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})t_{ij} + \epsilon_{ij}, \tag{2.1}$$

with $\tilde{b}_{i0} = (\beta_0 + b_{i0}), \tilde{b}_{i1} = (\beta_1 + b_{i1})$ and $b_i = (b_{i0}, b_{i1})'$ taken as random effects with mean zero. Equation (2.1) can be treated as a special case of general mixed effects model which assumes that the outcome vector y_i of all n_i observations for subject i satisfies

$$y_i = x_i'\beta + z_i'b_i + \epsilon_i, \tag{2.2}$$

where x_i and z_i are $(n_i \times p)$ and $(n_i \times q)$ design matrices of known covariates corresponding to the fixed and random effects respectively.

Furthermore, β is a $p \times 1$ column vector of the fixed-effects regression coefficients βs and b_i represents the random effects coefficients assumed to be normally distributed with mean zero and 2×2 covariance matrix D. Moreover, the $n_i \times 1$ column vector of the residual components ϵ_i are assumed to be independent such that $\epsilon_i = N(0, R)$, where $R \sim \sigma_i^2 I_{n_i}$ with I_{n_i} represents the n_i dimensional identity matrix. This implies that observations taken from subject i are said to be independent conditional on the subject specific random effects (Van et al., 2010).

2.2.2 Generalized Linear Models (GLMs)

The term Generalized Linear Model (GLM) refers to a larger class of models popularized by (McCullagh & Nelder, 1989). In these models, three components are specified, that is the random component which refers to the probability distribution of response variable Y which is assumed to

follow an exponential family of distributions, the systematic component which specifies explanatory variables (predictors) in the model and the link function which describes how the mean of the response and linear combination of the predictors are related. The exponential family takes the general form:

$$f(y|\theta,\phi) = exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right], \tag{2.3}$$

where θ called the canonical parameters and denotes the location, while the ϕ called the dispersion parameter and denotes the scales (McCullagh & Nelder, 1989).

According to Faraway (2014), the most used canonical links for GLMs are presented in **Table** 2.1. For example, the normal density has an identity link, with unit variance, whereas the Poisson has a log-link, and binomial has a logit link. The corresponding variances are provided in the Table.

***Table** 2.1: Canonical links for GLMs*

Family	Link	Variance
Normal	$\varphi = \mu$	1
Poisson	$\varphi = \log\mu$	μ
Binomial	$\varphi = \log\left(\frac{\mu}{1-\mu}\right)$	$\mu(1-\mu)$
Gamma	$\varphi = \mu^{-1}$	μ^2
Inverse Gaussian	$\varphi = \mu^{-2}$	μ^3

2.2.3 Generalized Linear Mixed Effects Model (GLMM)

Generalized linear mixed models (GLMMs) are defined as extensions of linear mixed effect models and generalized linear models to accommodate non-continuous responses such as binary responses or counts (Hedeker, D. (2005). GLMMs are also considered for both fixed and random effects which make them appropriate models to apply to longitudinal data where repeated

observations from the same subject are nested within subjects (Bolker et al, 2009 and Gad & El Kholy, 2012.). Nested observations are repeated measurements for individuals which are organized at more than one level. On the other hand, GLMs consider only the fixed effects models (McCulloch, & Neuhaus, 2014).

According to Ojo et al. (2017), fixed effects models assume that all observations are independent of each other which makes them to be inappropriate models for analysis of correlated data structures such as clustered or multilevel data where observations are nested within groups. Let $y_i, \dots, y_{n_i}$ be independent continuous random vectors for n_i subjects, where $y_i = (y_{i1}, \dots, y_{in_i})$ denotes an observed response vector for subject i. For a given subject i at j^{th} observation, the GLMM (Breslow and Clayton, 1993) is given by

$$\varphi\left(E\left(y_{ij}\middle|x_{ij}, z_{ij}, b_i\right)\right) = x'_{ij}\beta + z'_{ij}b_i, \tag{2.4}$$

where φ denotes the link function linking the conditional mean response $(E(y_{ij}|x_{ij}, x_{ij}, b_i))$ with linear mixed model $\mu_i = x'_{ij}\beta + z'_{ij}b_i$, for which the known covariates x_{ij} and z_{ij} represent the design matrices of dimension $n_i \times p$ and $n_i \times q$ corresponding to the vector $\beta = \left(\beta_1, \dots, \beta_p\right)'$ of regression coefficients (the fixed effects) and $b_i = \left(b_{i1}, \dots, b_{iq}\right)'$(random effects) respectively.

2.2.4. Generalized additive models

Generalized additive models (GAMs) are statistical models that can be used to estimate trends as smooth functions of time. GAMs use automatic smoothness selection methods to objectively determine the complexity of the fitted trend (Simpson, 2018). According to Yang et.al (2012),

Generalized Additive Model (GAM) provides a flexible and effective technique for modelling nonlinear time-series in studies of the health effects of environmental factors. However, GAM assumes that errors are mutually independent, while time series can be correlated in adjacent time points (Yang et al., 2012). When the outcome in a GAM is subject to missing, practical analyses often assume that missingness is missing at random (MAR) however, this assumption can be of suspicion when the missingness is by design (Xie, 2010). In parametric regression, the researcher must choose a functional form to impose on the data, for example, that trend over time is linear. However, GAMs reverse this process by letting the data inform the choice of functional form (Sullivan, Shadish & Steiner, 2015). Sullivan et al. 2015, suggested that GAMs may be very useful both as a form of sensitivity analysis for checking the plausibility of assumptions about trend and as a primary data analysis strategy for testing treatment effects.

2.2.5. Semiparametric Models

Semi-parametric models involve a partly specified regression function in some primary covariates and a non-parametric function in some other secondary covariates (Sutradhar, 2018). These models in a longitudinal setup has recently been discussed extensively both for repeated Poisson and negative binomial count data (Sutradhar, 2018). However, the inferences for semi-parametric Poisson and negative binomial models cannot be applied to longitudinal binary responses through a binary dynamic logit model, as these models unlike the count data models produce recursive means and variances containing the dynamic dependence or correlation parameters. In such case, Sutradhar (2018), considered general multinomial dynamic logit model in a semi-parametric setup can be employed to analyze nominal categorical data in a semi-parametric longitudinal setup. The model was then modified to analyze ordinal categorical data. The ordinal responses are fitted by

using a cumulative semi-parametric multinomial dynamic logit model. Fan, Huang, & Li. (2007), proposed a class of semi-parametric models for covariance function of longitudinal data and the estimation procedures for model for model coefficients using a profile weighted least squares approach.

2.3. Estimation approaches

Frequentist and Bayesian approaches are the two major estimation methods used in studies of HIV dynamics. In the frequentist approach, the maximum likelihood estimation (MLE) methods have been widely used for estimating HIV dynamic model parameters. However, due to complexity of such models, MLE method is computationally intensive compared to Bayesian approach (Chen, 2012).

Different extensions based MLE estimation in NLME models, likelihood approximations such as linearization (Pinheiro and Bates, 1995) or Laplace approximation (Wolfinger, 1993) have been proposed, leading to inconsistent estimates (Ding & Wu, 2001). Guedj et al. (2007) proposed algorithms based on Gaussian quadrature but these algorithms are cumbersome and were not applied to problems with more than three random effects. Wu and Zhang (2002) proposed a semi-parametric approach. Other new algorithms are stochastic EM algorithms as Monte Carlo EM (Wu, 2004). Commenges et al. (2011) proposed an asymptotic distribution of the maximum h-likelihood estimators (MHLE).

Another complexity of viral load analysis is left censoring which occurs when viral load is below a limit of quantification (LOQ). The proportion of subjects with viral load below LOQ has

increased with the development of highly active anti-retroviral treatment (HAART). Although it is known that when ignored, this censoring may induce biased parameter estimates (Samson et al., 2006), several authors did not consider this problem (Ding and Wu, 2001; Wu et al., 2005). Conversely, Hughes (1999); Fitzgerald et al. (2002); Thiebaut et al. (2005); Guedj et al. (2007) proposed different approaches to handle accurately the censored viral load data. Samson et al. (2006) extended the SAEM algorithm to perform maximum likelihood estimation for left-censored data.

The Bayesian approach is an efficient way to incorporate prior information, both point estimates and uncertainties (variances), into analysis to identify more unknown parameters in complex models (Yuan & MacKinnon, 2009). Bayesian estimation methods based on Markov Chain Monte Carlo (MCMC) algorithms and informative priors have first been proposed for complex ODE HIV dynamic models and NLME methods (Putter et al., 2002; Wu et al., 2005; Huang et al., 2006). Yangxin et al. (2014) used a Bayesian NLME joint modeling approach to estimate parameters of viral dynamic models with skew-t distribution in the presence of covariate measurement error. In this model, viral load response, time-varying CD4 covariate with measurement error, and time-dependent drug efficacy a function of multiple treatment factors was fully integrated into the data analysis.

2.4. Model Selection Approaches

Model selection is the task of selecting a statistical model from a model class, given a set of data (Ding, Tarokh & Yang, 2018). The primary objective of model comparison is to choose the simplest model that provides the best fit to the data. In case of missing data, the naive use of only

complete cases can lead to serious deficiencies in its applicability to measure the distance between models. Modeling complete data $D_{obs} = \left(Y_{ij}, i = 1, \dots, n; j = 1, \dots, n_i\right)$ is often preferred but there is a case $\boldsymbol{Y}_i = (Y_{obs,ij}, Y_{mis,ij})$ which is totally a different modeling strategy (Duncan & Duncan, 1994).

2.4.1 Akaike's information criterion (AIC)

Akaike's information criterion (AIC) is a measure of goodness of fit of an estimated statistical model. It is not a test on the model in the sense of hypothesis testing; rather it is a tool for model selection. The AIC penalizes the likelihood by the number of covariance parameters in the model. Akaike (1974) proposed the information criterion

$$AIC = -2\log\left(\mathcal{L}(\boldsymbol{\theta})\right) + 2p, \tag{2.5}$$

for model selection, where, $\mathcal{L}\left(\widehat{\boldsymbol{\theta}}\right) = \int \prod_{i=1}^{n}\prod_{j=1}^{n_i} f_y(\boldsymbol{Y_i}; \boldsymbol{\theta}) dY_{mis,ij}$ is the maximized value likelihood function for the estimated model and p is the number of parameters in the model. In other words, the first term measures the goodness of fit, whereas the second term is interpreted as a penalty for model complexity. The AIC values for candidate models are computed, and then the model that minimizes AIC is selected indicating the less information a model loses, the higher the quality of that model. Therefore, information criterion estimates the expected discrepancy between the unknown true distribution of $\boldsymbol{y}$, which is denoted as q_y, and the estimated distribution $f_y(\widehat{\boldsymbol{\theta}}_y)$. This discrepancy is measured by the incomplete-data Kullback-Leibler divergence.

2.4.2 Bayesian information criterion (BIC)

The Bayesian information criterion (BIC) takes the form of a penalized log-likelihood function where the penalty is equal to the logarithm of the sample size times the number of estimated parameters in the model. Claeskens and Hjort (2008), gave a general form of BIC as

$$BIC(\mathrm{M}) = 2\log \mathcal{L}(\mathrm{M}) - (\log n)p, \tag{2.6}$$

where $\mathcal{L}(\mathrm{M})$ is the maximized value of the likelihood function of model M, p is the number of parameters in the model M. The model with the largest BIC value will be chosen as the best model.

2.4.3 Quasi-information criterion (QIC)

Although the AIC can be used in association with mixed models, it cannot be used with GEEs to select either the optimal set of explanatory variables or correlation matrix, because GEE estimation is based on the quasi-likelihood rather than the maximum likelihood (Barnett et al., 2010). The quasi-likelihood counterpart to the AIC is the QIC, or the "quasi-likelihood under the independence model information criterion" (Pan, 2001). Based on the Kullback–Leibler information defined for the quasi-likelihood under the working independence model, Pan (2001) derived the QIC statistic for a GEE model with given working correlation:

$$QIC = 2\sum_{i=1}^{n}\sum_{j=1}^{n_i} Q\left(\hat{\beta}, \hat{\phi}; Y_{obs;ij}, Y_{mis;ij}, X_{ij}\right) + 2Tr\left(\widehat{\Phi}_I^{-1}\widehat{W}\right), \tag{2.7}$$

where $Q\left(\hat{\beta}, \hat{\phi}; Y_{obs;ij}, Y_{mis;ij}, X_{ij}\right)$ is the log-quasi likelihood under independent model with the substitution of $\hat{\beta}\ and\ \hat{\phi}$, $Tr(A)$ denotes the trace of matrix A, and $\widehat{\Phi}_I = (\sum_{i=1}^{n} D_i^t A_i^{-1} D_i)^{-1}$ with the substitution of $\hat{\beta}\ and\ \hat{\phi}$. A better model is the one with a smaller QIC.

2.4.4 Missing Longitudinal Information Criterion (MLIC)

Shen and Chen (2012) proposed the Missing Longitudinal Information Criterion (MLIC) based on the expected quadratic loss. Suppose $\boldsymbol{\mu}_i^0 = E(\boldsymbol{Y}_i|X_i)$ is the true mean of $\boldsymbol{Y}_i$ and $\hat{\mu}_{ij}$ is the estimated mean of Y_{ij} based on a candidate model. The MLIC statistic for a GEE model with given working correlation is written as

$$MLIC = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(Y_{ij} - \hat{\mu}_{ij}\right)^2 + 2Tr\left(\widehat{\Phi}\widehat{\Im}\right), \tag{2.8}$$

where $\widehat{\Im} = \sum_{i=1}^{n} D_i^t V_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i^0)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i^0)^t D_i|_{\beta=\widehat{\beta},\phi=\widehat{\phi}}$. In practice, $\boldsymbol{\mu}_i^0$ is unknown and estimated by the largest candidate model (Shen & Chen, 2012). Similarly, a model with smaller MLIC value represents a better model. The MLIC has also modified to accommodate monotonically missing response data by applying the WGEE estimation (Shen & Chen, 2012).

2.5. Missing Data

Missing data are a common feature in many areas of research especially those involving survey data in biological, health and social sciences (Chinomona & Mwambi, 2015). Even in well-controlled situations, missing data invariably occur in longitudinal studies (Hedeker & Gibbons, 1997). Subjects can be missed at a measurement wave; these subjects provide data at some but not all study timepoints. Alternatively, subjects who are assessed at a given study timepoint might only provide responses to a subset of the study variables, again resulting in incomplete data. Finally, subjects might drop out of the study or be lost to follow-up, thus providing no data beyond a specific point in time (Hedeker & Gibbons, 1997).

2.5.1 Mechanisms of missing data

a) Missing Completely at Random (MCAR)

The most basic assumption about the missing data is to assume that they were missing completely at random. That is, a subject is missing at a timepoint for completely random reasons. This implies that the missing data indicators do not depend on the dependent variable values that were observed or those that were not observed (Xu & Blozis, 2011).

b) Missing at Random (MAR)

The MAR mechanism occurs when the probability of missing data in a variable is related to some other variable(s) in the data set that is, missing at random after controlling for all other related variables (Graham, 2009). An example of MAR is when subjects drop out of the study because their value of the dependent variable falls below (or exceeds) some critical value. For instance, if subjects in a depression study who have Hamilton depression scores below 15 drop out of the study (i.e.., they are measured at a timepoint with a score below 15, but then are not measured at any future time points).

c) Missing Not at Random (MNAR)

Missing not at random (MNAR) is the situation where the missingness is related to the unobserved dependent variable vector, after taking observed variables into account. The notion here is that there is a relationship between what would have been observed and the missingness. MNAR can occur if subjects are not measured at a given timepoint because their value of the dependent variable falls below (or exceeds) some critical value. For instance, to contrast MNAR with MAR, MNAR occurs if subjects who have Hamilton depression scores below 15 are not measured at that timepoint.

2.5.2. Handling missing data

There are three categories of methods for handling missing data are: case deletion, imputation, and augmentation (McKnight, McKnight, Sidani & Figueredo, 2007). Data deletion is an efficient way of dealing with missing data if missing data are MCAR. Researchers delete cases containing missing values and run a model without missing values. However, as the fraction of missing cases grows, problems such as reduction in statistical power and potential bias will arise. Graham (2009) recommended that, if at least 5% of the cases are missing, one should use multiple imputation or data augmentation.

Barnard & Meng (1999) proposed multiple imputation which has solved the problem of biased uncertainty, it has become the most practical and the best-recommended method in most cases. Among imputation techniques that can generate unbiased parameter estimates under the MAR assumption, most relevant and useful methods are expectation maximization (EM) algorithms and Markov chain Monte Carlo (MCMC) procedures. Missing data may seriously compromise inferences from randomised clinical trials, especially if missingness is not at random and if missing data are not handled appropriately (Sterne et al., 2009).

CHAPTER 3 : STUDY DESIGN & EXPLORATORY DATA ANALYSIS

3.1. Introduction

The standard approach for monitoring treatment outcomes in patients on ART depends on the measurement of HIV-load over time (Chendi, 2019). As viral load is a more sensitive and an earlier indicator of treatment failure, Namibia has transitioned to routine viral load monitoring rather than CD4 count for treatment monitoring (MoHSS, 2014). According to the fifth edition of Namibia National ART guidelines of 2016, all patients initiating therapy routinely have a viral load (VL) assay done at 6 and 12 months after beginning therapy and every 12 months thereafter (but every 6 months for children/adolescents ≤19 years). VL assays are also recommended for patients already on treatment who are showing evidence of immunologic and or clinical failure. Virological failure is defined as a viral load >1,000 copies/ml 6 months after starting ART or viral rebound to >1,000 copies/ml on two consecutive measurements after a period of viral suppression (Johnston et al, 2012). Viral suppression means that a person's viral load has reduced to an undetectable level (<40 copies/ml).

ARV medication adherence is vital for the success of ART. Very high levels of adherence and taking at least 95% of prescribed doses are required to achieve sustained suppression of HIV replication over time. Adherence is promoted through proper ongoing support and counseling. Adherence is also promoted by prescribing simplified, well-tolerated regimens involving as few pills as possible, administered no more than two times per day. This chapter aimed to obtain an in-depth understanding of the effect of ART on viral load and examine the average rate of change in viral load in patients over time.

3.2. Study Design and Setting

3.2.1. Study design

This study followed a retrospective cohort study design, with data of records of PLHIV on ART obtained from the Luderitz Hospital in the !Karas region of Namibia. This study determined the average change in viral load and assessed adherence in all HIV adult patients (patients above the age of 19 years) initiated on ART starting from January 2015 to December 2017 and have been on ART for at least 12 months at the hospital.

3.2.2. Study Setting

Luderitz hospital is situated in a small town of Luderitz in the southern region of Namibia and it is the only state facility that offers ART in the town. The catchment population of the town is approximately 14 000. The first patient enrolled on ART care was enrolled in October 2003 and the first ART initiation was in the same month. As of December 2019, 3473 patients have been enrolled on care while 2624 have been initiated on ART at this facility between October 2003 and December 2019. A total of 151 (5.8%) patients initiated on ART have since been recorded dead, while 222 (8.4%) were lost to follow-up, 1101 (42%) transferred out to other facilities and 1150 (43.8%) were active on care.

Systematic reviews (Fox & Rosen, 2010) show that retention rates are estimated to range from as low as 64% to as high as 94% at 12 months after ART initiation. Thus, one can say that the retention rate at this facility is quite good. Patients on ART are screening for TB at each follow up and get tested for TB if screened positive. 485 (13.9%) patients enrolled on care have been on TB

treatment and 2524(72.7%) have received TB preventative therapy (TPT). Currently there are 1,428 (884 females) patients active on ART at Luderitz Hospital, of which 68 (4.8%) are 19 years and younger, and 149(10.4%) are 50 years and older. The proportion of patients who are virally suppressed (VL<100 copies) stands at 91% as of December 2019, which implies that there is good patient management at the facility.

This study used patients who were initiated on ART between January 2015 and December 2017, this was because viral load monitoring for patients has started in 2015 and the study aimed at working with patients who has at least two viral loads measured. Only adult patients were chosen for this study since viral load monitoring differs between adults and pediatric patients.

3.2.3. Study variables

The outcome variable of this study was viral load, which is the number of virus copies measured at 6 months, 12 months and yearly after initiation of ART. The independent variables were**:** Age at start of ART, baseline weight, gender, WHO Clinical stage, time (in months) on ART, baseline CD4 Count, adherence, and weight at follow up. **Table** 3.1 presents details of the study variables.

***Table** 3.1: Variable description of ART data taken from Luderitz Hospital from 2015 –2017*

Variable	Description	Coding
Age	Age of patients at the start of ART in years	None
Baseline Weight	Weight of patients at the start of ART in kg	None
Gender	Gender of patients	Female=1, Male = 2
Baseline_CD4 Count	Number of cells per cubic millimeter measured at the beginning of ART	None
WHO Clinical stage	WHO clinical stage at start of ART Stage I = Asymptomatic Stage II = Moderate unexplained weight loss Stage III = Unexplained severe weight loss (>10% of presumed or measured body weight) Stage IV= HIV wasting syndrome	Stage I = 1, Stage II = 2, Stage III = 3, Stage IV = 4
Time in months	Observation time of viral load (the first 6^{th} month of ART, month 12 and yearly thereafter)	None
Viral Load	Number of viral particles found in each milliliter of blood measured for individuals in the first 6th month, month 12 and yearly thereafter	None
Adherence	Adherence of patients to ART measured with pill count at follow up. Good=when a patient misses at most 3 doses per month Fair =when a patient misses between 4 and 8 doses per month Poor= missing more than eight doses per month	1=Good 2=Fair 3=Poor
Weight at follow up	Weight of individuals at follow up visits	None

3.2.4. Ethical Clearance

This study was approved by the Research Ethical Committee of the University of Namibia from The Centre of Postgraduate Studies. Permission to use the ART data was granted by the Luderitz Hospital management and patients' confidentiality was ensured by de-linking extracted data from identifiable information.

3.3. Descriptive Statistics

A total of 154 HIV positive patients initiated on ART between January 2015 and December 2017 were included in this study. The baseline characteristics of patients are displayed in **Table** 3.2. Among these patients, 110 (71.4%) were females and 44 (28.6%) were males. About 99 (64.3%) patients were in WHO stage I, 25 (16.2%) were in stage II, 27 (17.5%) were in stage III and 3 (1.9%) were in stage IV. The stages descriptions are given in Table 3.1. A total of 46 (29.9%) patients were between the ages of 20-29 years, 53 (34.4%) were between 30-39 years, 45 (29.2%) were 40-49 years of ages and 10 (6.5%) patients were 50 years and above at the time of ART initiation. A total of 53 (34.4%) had a CD4 count below 200cells/mm, 101 (65.6%) had a CD4 count of 200cells/mm^3 and more at the initiation of ART, while 56 (36.4%) had no initial CD4 count taken.

A total of 133 (86.4%) patients achieved viral suppression (<1000copies/mm^3) at the 12 months of ART, out of which 98 (73.7%) patients were females, 48 (36.1%) were aged between 30 and 39 years, 85 (63.9%) were at clinical stage 1 at ART initiation, 61 (45.9%) patients had baseline weight between 50 and 70 kilograms, 89 (66.9%) had CD4 counts ≥ 200 cells/mm^3 at ART initiation.

A Chi-square test was used to test for association between the outcome variable (viral load) and the independent variables. The results show that gender and WHO stage had p-values (0.046<0.05) and (0.041<0.05) thus it was concluded that there was a significant relationship between gender and viral load as well WHO stage and viral load.

Table 3.2: Demographic and clinical characteristics by viral suppression at 12 months

Variables	Viral suppression at 12 months							
			Yes		No			
	Count (n=154)	%	Count (n=154)	%	Count (n=154)	%	χ^2	P-values
GENDER								
Male	44	28.6	35	79.5	99	20.5	3.980	0.046
Female	110	71.4	98	89.1	12	10.9		
AGE								
20-29 years	46	29.9	40	86.9	6	13.1	1.808	0.613
30-39 years	53	34.4	48	90.6	5	9.4		
40-49 years	45	29.2	37	82.2	8	17.8		
≥ 50 years	10	6.5	8	80	2	20		
WHO stage								
Stage-I	99	64.3	85	85.9	14	14.1	8.297	0.041
Stage-II	25	16.2	21	84	4	16		
Stage-III	27	17.5	26	96.3	1	3.7		
Stage-IV	3	1.9	1	33.3	2	66.6		
CD4 at ART initiation								
< 200	53	34.4	44	83	9	17	0.314	0.842
≥ 200	101	65.6	89	88.1	12	11.9		
Baseline Weight								
< 50kg	26	16.9	20	76.9	6	23.1		
50-70kg	76	49.4	61	80.3	15	19.7		

> 70kg	52	33.8	50	96.2	2	3.8	1.267	0.538

3.4. Statistical Analysis

3.4.1. Notation

In this study, the response variable (viral load) was denoted by $\boldsymbol{Y}_{ijt_i}$, and the time dependent explanatory variables, or covariates were denoted by $\boldsymbol{X}_{ikt_i}$. In addition, n was the number of subjects (patients) in the study, n_i was the number of time points measured for each subject i, q was the number of response variables, and p was the number of explanatory variables, measured for each subject and time. The explanatory variables included the indicator variable making it special events affecting each subject. There were also time independent explanatory covariates, which were denoted as $\boldsymbol{Z}_{ij}$; $for\ i = 1, \dots, n, and\ j = 1, \dots, r$.

3.4.2. Exploratory data analysis

It is a critical process of visualizing the pattern of data as to spot anomalies as well as to check assumptions with the help of the summary statistics and graphical representations. Hence, plotting individual profiles to carefully study the viral load should be done prior to performing any formal model fitting. Therefore, in this study exploratory data analysis was done in order to assess the nature of viral load by exploring the average change in viral load over time and the correlation and covariance structure.

a) Exploring the Mean Structure

The main aim of exploring the mean structure is to choose the fixed effects for the model. To explore the overall mean, the response variable was plotted against time. In addition to the overall mean, the possible differences between the gender groups, baseline CD4 count, WHO clinical stages and adherence were studied by plotting the mean of each group as shown in **Table** 3.3 to **Table** 3.7 and **Figure** 3.1 to **Figure** 3.5.

***Table** 3.3: The mean of patients' viral load taken at each follow-up time*

Time in months	Mean	SD	95% CI
6	34365.5	183237.8	(5424.68, 63306.32)
12	31158.3	193622.3	(68.47, 62248.13)
24	3972.62	18573.69	(237.60, 7707.64)
36	1653.57	7652.25	(0[1] , 3675.96)

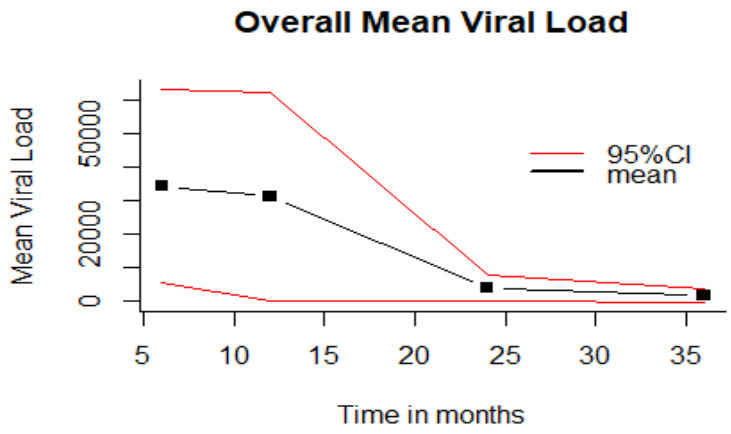

***Figure** 3.1: The overall average profile of the viral load taken at Luderitz hospital, from 2015-2017*

[1] -368.82

In **Table** 3.3, the estimated mean of the viral load showed a decrease over time. This is a good thing as it meant that after the patients were initiated on ART, their viral load has decreased due to the effect of the therapy. **Figure** 3.1 shows a decline in patients' viral load, this implies that the patients' immune systems were boosted, and the progression of the disease declined over time. The mean viral load of patients after 36 months was relatively low 1653.57, this meant that majority of the patients had viral loads less than 1000 copies.

***Table** 3.4: The average profile of the viral load by WHO stage taken at Luderitz hospital*

	WHO STAGE			
Time in months	**1**	**2**	**3**	**4**
6	38447.94	23525.89	33030.79	3348.33
12	35616.45	2323.03	41543.17	5614.00
24	3455.56	1969.28	139.80	44543.36
36	1744.25	1078.69	2602.92	284.01

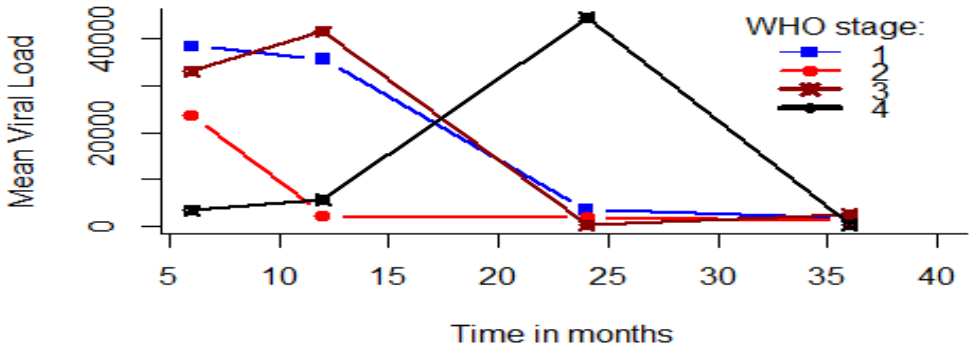

***Figure** 3.2: The average profile of the viral load by WHO stage taken at Luderitz hospital, from 2015-2017*

Table 3.4 and **Figure** 3.2 depict the patients' mean viral load by baseline WHO clinical stage over time. WHO staging is in 4 stages that are categorized according to the CD4 count at baseline or progression of the infection looking at the physical appearance of the patient such as skin

condition, TB status and other opportunistic infections. In **Table** 3.4, viral load was high in patients with initial clinical stage 1 at almost every time point. This could simply be because some patients might have been categorized as stage 1 when they belonged to advanced stages.

***Table** 3.5: The average profile of the viral load by CD4 count group taken at Luderitz hospital*

	CD4 count group	
Time in months	**< 200**	**>= 200**
6	371163.36	12912.86
12	71201.76	475.59
24	14228.37	783.45
36	1858.29	2034.01

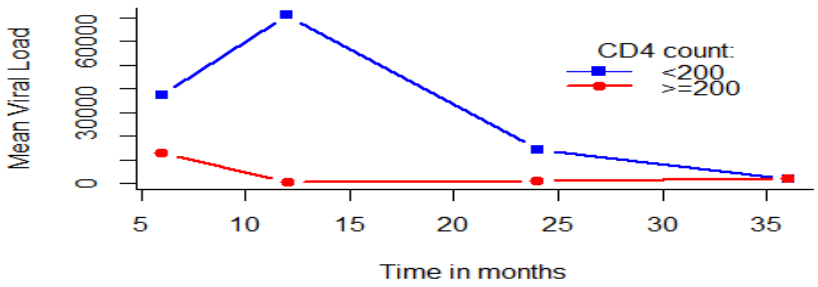

***Figure** 3.3: The average profile of the viral load by CD4 count group taken at Luderitz hospital from 2015-2017*

Table 3.5 and **Figure** 3.3 show the patients' viral load by baseline CD4 count over time. Patients who started the therapy with higher CD4 counts (≥ 200 cells/mm^3) had lower viral load at all the time points as compared to patients who started ART with low CD4 counts. This is because the virus had progressed in the patients with lower CD4 count more than in patients with higher CD4 counts which means that viral load was higher at baseline for patients with fewer CD4 count.

***Table** 3.6: The average profile of the viral load by adherence count group taken at Luderitz hospital*

Adherence			
Time in months	**1**	**2**	**3**
6	34365.50	-	-
12	15922.39	181712.06	336478.30
24	4102.61	29.53	-
36	1705.53	586.42	19.00

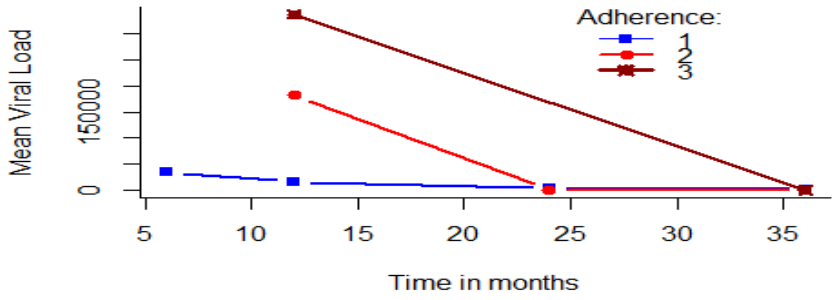

***Figure** 3.4: The average profile of the viral load by adherence group taken at Luderitz hospital from 2015-2017*

Table 3.6 and **Figure** 3.4 depict the patients' viral load by adherence over time. Adherence was grouped in to three categories, 1 for good, 2 for fair and 3 for poor. In this study data, adherence was not consistent among some patients over time. From **Table** 3.6 only at time 12 and 36 where non-adherent patients were observed, with time 12 having the highest mean viral load.

***Table** 3.7: The average profile of the viral load by gender taken at Luderitz hospital*

Gender		
Time in months	**Female**	**Male**
6	16484.77	79783.60
12	4959.93	88982.53
24	7594.37	8317.50
36	1756.65	1063.18

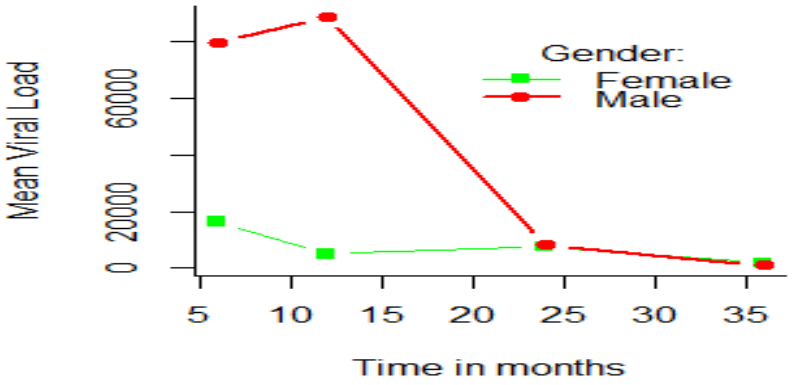

***Figure** 3.5: The average profile of the viral load by gender taken at Luderitz hospital from 2015-2017*

The mean viral load of patients by gender is shown in **Table** 3.7 and **Figure** 3.5. Male patients appeared to have higher viral load than the female patients until the 24 months of the therapy. It also shows that both males and females had decreasing viral load over time.

b) **Exploring variation among individuals**

By considering independent observations, the variability in a response measurement can be summarized using single variance parameter σ^2 (Diggle et al, 2002). Variance computation is given as one half of the expected squared distance between any two randomly selected measurements. However, with longitudinal data the distance between measurements on different subjects is usually expected to be greater than the distance between repeated measurements taken on the same subject (Hedeker & Gibbons, 2006). Taking into account that measurements from the same subject are correlated, the interpretation of variance of repeated viral load is given as $\sigma^2(1-\rho_{jk}) = E\left[\left(\boldsymbol{Y}_{ij} - \boldsymbol{Y}_{ik}\right)^2\right]$ with the assumption that $E(\boldsymbol{Y}_{ij}) = E(\boldsymbol{Y}_{ik})$. Note that $\rho_{jk} > 0$

shows that between-subject variation is greater than the within-subject variation. On the other hand, $\rho_{jk} = 1$ and $\boldsymbol{Y}_{ij} = \boldsymbol{Y}_{ik}$ indicating no variation for repeated viral loads measured on the same subject (Hedeker & Gibbons, 2006).

***Table** 3.8: The variance of the viral load by CD4 count group taken at Luderitz hospital*

CD4 count group		
Time in months	**< 200**	**$>= 200$**
6	17475172665	5024386863
12	54196680907	5061963
24	1516819907	11378275
36	9339872	92962618

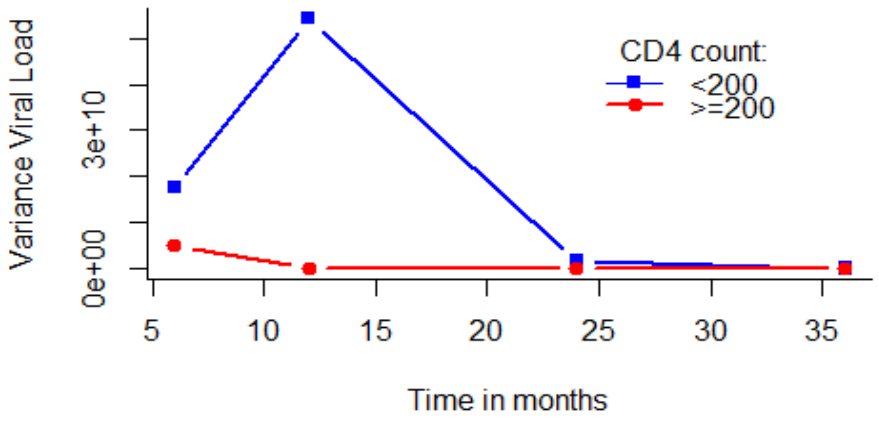

***Figure** 3.6: The variance profile of the viral load by CD4 count taken at Luderitz hospital from 2015-2017*

***Table** 3.9*: The variance of the viral load by WHO stage taken at Luderitz hospital

WHO STAGE				
Time in months	**1**	**2**	**3**	**4**
6	46217514716	8637549740	15600430000	25090561
12	50673358050	85953932	26680890000	62608050
24	241378330	66307638	102210.8	5947255010
36	90767314	6021784	19714300	140461

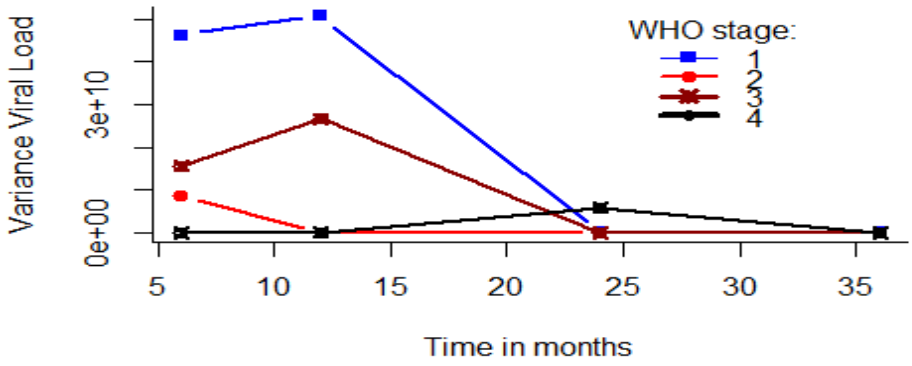

***Figure** 3.7: The variance profile of the viral load by WHO stage taken at Luderitz hospital*

***Table** 3.10: The variance of the viral load by gender taken at Luderitz hospital*

Gender		
Time in months	**Female**	**Male**
6	13208038360	83454177266
12	1093049068	123805257695
24	1948480260	883477608
36	72566411	4666523

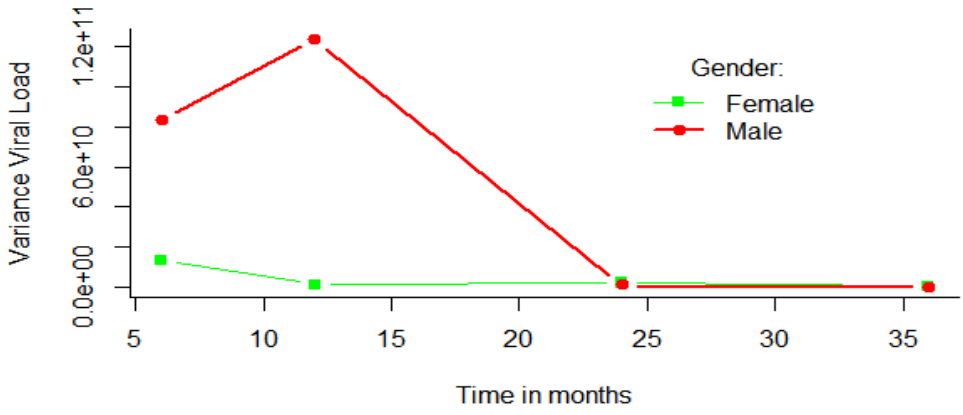

***Figure** 3.8: The variance profile of the viral load by gender taken at Luderitz hospital*

Table 3.8 to **Table** 3.10 and **Figure** 3.6 to **Figure** 3.8 show how viral load varies in patients according to their CD4 count, WHO stage and gender. From figure 3.6, variation was high in patients who started ART with low CD4 count compared to those with high CD4 counts. From figure 3.8, it can be seen that the estimated variance of viral load was higher in males than in female patients over time up to the 12th month of therapy and dropped to the same level after the 24th month of therapy.

c) Correlation Structure

Since outcome variable can be correlated it is useful to understand the strength of correlation across time. In this subsection, correlation structure is explored for understanding components of variation and for identifying a correlation model for regression method.

The correlation matrix used for exploring the correlation structure within the outcome variable is given as

$$Corr(\boldsymbol{Y}_i) = \begin{pmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,n} \\ \rho_{2,1} & 1 & \cdots & \rho_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n,1} & \rho_{n,2} & \cdots & 1 \end{pmatrix}$$

which is useful for comparing the strength of association between pairs of outcomes particularly when the variance σ_j^2 are not constant. Sample estimates of the correlations can be obtained using

$$\rho_{j,k} = \frac{1}{N-1}\sum_i \frac{\left(\boldsymbol{Y}_{ij} - \overline{\boldsymbol{Y}}_{.j}\right)}{\hat{\sigma}_j} \frac{\left(\boldsymbol{Y}_{.j} - \overline{\boldsymbol{Y}}_{.k}\right)}{\hat{\sigma}_k}, \tag{3.1}$$

where $\hat{\sigma}_j$ and $\hat{\sigma}_k$ are the sample standard deviations of $\boldsymbol{Y}_{ij}$ and $\boldsymbol{Y}_{ik}$ respectively. Note that this is done across subjects for times t_i and t_k.

The correlation Matrix

***Table** 3.11: Correlation matrix*

Months	**6**	**12**	**24**	**36**
6	1.000	0.076	0.293	-0.027
12	0.076	1.000	0.026	0.076
24	0.293	0.026	1.000	0.066
36	-0.027	0.076	0.066	1.000

From the correlation matrix in **Table** 3.11, the correlation values are different at each time point, thus an unstructured correlation is appropriate.

3.5. Summary

This chapter aimed at exploring the average change of viral load in patients initiated on ART from January 2015 to December 2017 at Luderitz Hospital in the !Karas region of Namibia. The mean structure has shown that on avaerage, viral load decreased over time in patients until it reached an undetectable level which means that on average the therapy had a positive effect on the patients. Majority (85.9% and 84%) of the patients who started ART at an early stage (stage 1 and 2) of the infection achieved viral suppression within 12 months as compared to those who started therapy at an advanced stage (stage 4), patients who started ART at stage 4 are likely to fail on treatment as seen in table 3.4. Patients whose CD4 count was below 200 at the initiation of ART had a lower suppression rate (83.3%) at 12 months as compared to 87.8% of the ones with CD4 count ≥200 copies without accounting for 36.4% of the records which were missing. It was also found that on average , it takes a year (12 months) for a female patient on ART to achive viral load suppression and 2 years (24 months) for a male patient as seen in table 3.7. This could mean that female patients had a better adherence rate, or they started ART at an earlier stage of the infection as compared to men. Age category of 30–39 years old had the highest viral suppression of 90.6% at 12 months of therapy and the lowest (80%) viral suppression was found in the age category of 50 and above years as seen in **Table** 3.2. After exploring the longitudinal viral load data, it was found to be highly right skewed. Therefore, log transformation was used to make the data less skewed, this is important as it makes the patterns in the data interpretable and helps to satisfy the assumption of normality.

CHAPTER 4 : MODELING VIRAL LOAD USING MIXED EFFECTS MODELS AND GENERALIZED ESTIMATING EQUATIONS

4.1. Background

Studies of HIV dynamics are very important in evaluating the effectiveness of ART. They have significantly improved researchers' and health personnel's understanding of the pathogens of HIV infection and guided the treatment of AIDS patients and evaluation of ART (Huang & Lu, 2008). Therefore, viral load is an essential outcome variable across a wide spectrum of HIV research and surveillance studies (Rose et al., 2015). According to Huang et al. (2015), following ARV treatment, the profile of each subject's viral load tends to follow a dynamic trajectory, indicating multiple phases of decline and increase in viral load. Such multiple-phases can be described by a random change-point model with random subject-specific parameters.

There are several methods for modeling viral load as an outcome variable. Rose et al. (2015) used a log–binomial model and GEE with an exchangeable correlation structure to account for repeated measurements within participants. A semiparametric nonlinear mixed-effects (SNLME) model has been proposed for the complete viral load data which include the third stage viral load data, i.e., the data of those patients who fail the therapy (Ke, C., & Wang, Y., 2001). Haung et al. (2006) compared linear and biphasic nonlinear model performance and found that linear modeling may result in misleading conclusions because one has to truncate the data. Haung et al. (2015) proposed piecewise linear mixed effects models with skew-elliptical distribution to describe the time trend of viral load under Bayesian framework and the findings suggested that it is very important to assume a model with skew distribution in order to achieve reliable results when the data exhibit skewness.

Linear mixed effects models have become popular tools for analysis of longitudinal data because they are considered to be flexible and applicable (Van et al., 2010). They were developed with the idea that each individual in the population has its own mean response profile and subject specific effects over time. This makes mixed effects models more favorable as they take into account the correlation within repeated measurements from the same subject and deals with time between subsequent points unequally (Vangeneugden et al, 2004). As presented in Laird and Ware (1982) and Molenberghs and Verbeke (2001), the longitudinal measurements are fitted using a regression model that allows parameters to vary among subjects.

Rylence et al. (2019), investigated longitudinal lung function trends among HIV infected children in order to describe the evolution of lung disease and access the effect of ART. The analysis was performed using linear mixed effects regression modes with covariate parameters evaluated by likelihood ratio comparison. The models estimated that early ART initiation in life could prevent a deterioration of forced expiratory volume. In the ART-naïve cohort, likelihood ratio comparison suggested an improvement in forced vital capacity during the two years following treatment initiation, but no evidence among participants established on ART.

Abebe (2020) employed linear mixed effects model to evaluate predictors of longitudinal CD4 cell progression of HIV infected children who were under ART. The results revealed that observation time, age, WHO clinical stage, history of TB, and functional status had significantly associated with mean change in the square root of CD4 cell count and they are the predictor of longitudinal CD4 cell change. Furthermore, Koulai et al. (2017) used Bayesian mixed effect model to quantify

the recency of HIV infection at individual at individual level. Characteristics of different biomarkers that affect the ability to estimate recency were explored through simulation. The findings from the analysis suggested that predictive ability was improved by using joint models of two biomarkers, accounting for their correlation, rather than univariate models of single biomarkers.

On the other hand, GEE is an extension of the generalized linear model (GLM) to correlated data (Nelder & Wedderburn, 1972). Repeated measures of the same subject in longitudinal studies are correlated because of the continuity of the measurement over time (Rabe-Hesketh & Skrondal, 2005). To take account of the correlation, specification of a working correlation structure is required, it is important to select an appropriate working correlation structure for the repeated measures per subject in order to enhance efficiency of estimation of the regression parameter (Pardo & Alonso, 2019).

Song, Barnhart, and Lyles (2001) proposed a generalized estimating equations approach to estimate the correlation coefficient between two continuous variables, where one or both may be left-censored. They presented simulation studies to evaluate point and interval estimates of the correlation and compare the GEE results with a maximum likelihood approach. They also conducted a simulation study to explore the robustness of GEE estimates to the normality assumption. The proposed methods were applied to two HIV viral load data sets from clinical studies conducted in Bangkok, Thailand. From their findings, the proposed method can be easily extended to incorporate covariates.

Unlike linear mixed effects models which are based on the maximum likelihood theory for independent observations (Hedeker and Gibbons, 2006), the GEE method is based on the quasi-likelihood theory (Wedderburn, 1974), and no assumption is made about the distribution of response observations. Therefore, some of the statistics derived under the full likelihood theory cannot be applied to GEE directly. This includes, Akaike's information criterion (AIC; Akaike, 1974), a widely used method for model selection in linear mixed effects models, is not applicable to GEE. However, Pan (2001) proposed a model-selection method for GEE and termed it quasi-likelihood under the independence model criterion (QIC). This criterion can also be used to select the best working correlation structure in GEE analyses.

Although, these models have been used in the analysis longitudinal data in several studies, the comparison of models with different specifications is not a common practice. In this chapter, the use of mixed-effects models and generalized estimating equations with time varying viral load response, to model population characteristics and individual variations was proposed.

4.2. Statistical Methods

4.2.1. Notations

Suppose there were n patients under a longitudinal study that collects n_i repeated viral load measurements for i^{th} patient ($i = 1, \dots, n$). Let Y_{ij} denote the viral load measurements for i^{th} patient at time $(j = 1, \dots, n_i)$ with $Y_i = \left(Y_{i1}, \dots, Y_{in_i}\right)^T$. Let $X_i = \left(X_{i1}, \dots, X_{ip}\right)^T$ denote $p-$dimensional vector of the time-invariant and time-varying covariates measured at baseline and subsequent follow-up time for patient i.

4.2.2. Outcome Process Model

In general, the repeated viral load measurements of i^{th} patient measured at follow-up time j can be expressed as follows:

$$Y_i = \begin{bmatrix} Y_{11} & \cdots & Y_{1n_i} \\ \vdots & \ddots & \vdots \\ Y_{n1} & \cdots & Y_{nn_i} \end{bmatrix}.$$

(i) Linear Mixed-Effects model

The trajectory of the viral load the following model will be used Dynamic Linear Mixed-Effects model,

$$y_{ij} = \boldsymbol{\mu}_i^*(t_{ij}) + \varepsilon_{ij} \quad (4.1)$$

$$= \boldsymbol{\beta} X_i^T + \eta_i(t_{ij}) + \varepsilon_{ij},$$

where X_i is a vector of p non-dynamic covariates, β are the corresponding coefficients and η_i a stochastic process that may depend on dynamic covariates,

$$\eta_i(t_{ij}) = \boldsymbol{\Phi} c_i(t_{ij}) + Z_i(t_{ij})\boldsymbol{w}_i, \quad (4.2)$$

where $c_i(t_{ij})$ are time-varying covariates and Z_i is the design vector corresponding to the $q \times 1$ vector $\boldsymbol{w}_\text{i}$ of random effects. The random effects part models correlations due to the repeated measurements within subjects. In this study subject specific random effects model was used to model the correlation structures. Let $\boldsymbol{w}_i = (w_{i0}, w_{i1})$ represent a random intercept and a random slope for subject i, assume that $\boldsymbol{w}_i$ follows a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma_w}$ such that $\boldsymbol{w}_i \sim MVN_q(0, \boldsymbol{\Sigma_w})$, where

$$\boldsymbol{\Sigma}_w = \begin{pmatrix} \sigma_{w_0}^2 & \rho_w \sigma_{w_0} \sigma_{w_1} \\ \rho_w \sigma_{w_0} \sigma_{w_1} & \sigma_{w_1}^2 \end{pmatrix}.$$

That is,

$$\begin{bmatrix} w_{i0} \\ w_{i1} \end{bmatrix} \sim MVN \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{w_0}^2 & \rho_w \sigma_{w_0} \sigma_{w_1} \\ \rho_w \sigma_{w_0} \sigma_{w_1} & \sigma_{w_1}^2 \end{bmatrix} \right\}.$$

(ii) Generalized Estimating Equations

GEE extends generalized linear models to correlated data but differs from mixed effects models in that GEE explicitly fits a marginal model to data (McCullagh and Nelder, 1989). The probability distributions used in Generalized Linear Models are related to the one of GEE because they are all from the exponential family of distributions. The density function of any member of the exponential family takes the following form:

$$f(y; \theta, \phi) = exp\left[\frac{y\theta - b(\theta)}{a(\phi)} - c(y, \phi)\right], \tag{4.3}$$

where a, b and c are the functions that varies from distribution to distribution (Diggle, 2002). The parameter θ is called the canonical parameter. Let $\mu_i = (\mu_{i1}, \dots, \mu_{in_i})$ be a mean vector of response variable, where μ_{ij} is the corresponding j^{th} mean. The responses are assumed to be independent across patients but correlated within each patient. The marginal model specifies that a relationship between μ_{ij} and the covariates X_{ij} is written as follows:

$$g(\mu_{ij}) = X_{ij}^T \beta \tag{4.4}$$

where g is a known link function and β is an unknown $p \times 1$vector of regression coefficients. The conditional variance of Y_{ij} given X_{ij} is specified as $Var(Y_{ij} | X_{ij}) = V(\mu_{ij})\phi$, where V is a known variance function of μ_{ij}. Mostly, V and ϕ depend on the distributions of outcomes.

4.2.3. Likelihood Functions

Likelihood function for Linear Mixed effects model

Consider $Y_i = \left(Y_{i1}, \dots, Y_{in_i}\right)^T$ as the observed viral load measurements for patient i and $f(\cdot \mid \cdot)$ as a conditional density function. Let $\boldsymbol{\theta}$ be the unknown vectors of population parameters of a linear mixed effects model. The likelihood function of the viral load longitudinal data based on the observed data for the model is given as follows:

$$\mathcal{L}(\boldsymbol{\theta}|.) = \int \prod_{i=1}^{n} \left[\left\{ \prod_{j=1}^{n_i} f\left(Y_{ij} \middle| X_{\boldsymbol{i}}, \boldsymbol{w_i}, \boldsymbol{\theta}\right) \right\} \right] f(\boldsymbol{w}_i|\boldsymbol{\theta})\, d\boldsymbol{w}_i,$$

$$= \int \prod_{i=1}^{n} \left[\left\{ \prod_{j=1}^{n_i} f(Y_{ij}|X_{\boldsymbol{i}}, \boldsymbol{w_i}, \boldsymbol{\theta}) \right\} \right] f(\boldsymbol{w}_i|\boldsymbol{\theta})\, d\boldsymbol{w}_i, \qquad (4.5)$$

where, $f(Y_{ij}|X_{\boldsymbol{i}}, \boldsymbol{w_i}, \boldsymbol{\theta}) = (2\pi\sigma_\epsilon^2)\, exp\left({}^{-1}/_{2\sigma_\epsilon^2}\left[Y_i - \left(\boldsymbol{\beta}X_i + \eta_i(t_{ij})\right)\right]^2\right)$ is the probability density function of the viral load outcomes conditional on random effects and $f(\boldsymbol{w}_i|\boldsymbol{\theta}) = (2\pi)^{{}^{q}/_{2}}|\boldsymbol{\Sigma_w}|^{-1/2} exp\left(-{}^{1}/_{2}\, \boldsymbol{w_i'}\boldsymbol{\Sigma_w^{-1}}\, \boldsymbol{w}_i\right)$ is defined as the probability density function of the random effects (Gumedze, & Dunne, 2011).

Quasi-Likelihood for GEE

The estimating equation of equation (4.5) is the derivative of the log likelihood set equal to zero

$$\frac{\partial}{\partial\beta}\log\mathcal{L}(y)=\sum_{i=1}^{n}\frac{y_i-\mu_i}{a(\phi)V(\mu_i)}\frac{\partial\mu_i}{\partial\beta}\ . \tag{4.6}$$

Integrating equation (4.6) by adding an arbitrary constant to an anti-derivative and obtaining another anti-derivative gives,

$$\begin{aligned}\int\frac{\partial}{\partial\beta}\log\mathcal{L}(y)d\beta_x d\beta_z &= \int\sum_{i=1}^{n}\frac{y_i-\mu_i}{a(\phi)V(\mu_i)}\frac{\partial\mu_i}{\partial\beta}d\beta \\ &= \int\sum_{i=1}^{n}\frac{y_i-\mu_i}{a(\phi)V(\mu_i)}d\mu_i.\end{aligned} \tag{4.7}$$

Integrating equation (4.7) in this case will not yield a true log-likelihood, but instead generates something referred to as a quasi-likelihood (although it might be better to call it a quasi-loglikelihood). The formal definition of the quasi-likelihood is that it is the anti-derivative of the generalized estimating equation evaluated at the parameter estimates (Wedderburn, 1974)

$$Q(y;\mu)=\int_{y}^{\hat{\mu}}\frac{y_i-u}{a(\phi)V(\mu)}d\mu. \tag{4.8}$$

4.3. Data Analysis

The study population included all HIV/AIDS patients initiated on antiretroviral therapy (ART) follow-up from January 2015 to December 2017 at the Luderitz Hospital in the !Karas region of Namibia. One response variable was considered in this study, which was the longitudinal viral load of HIV adult patients initiated on ART. A viral load test is used to measure the amount of HIV virus in a sample of blood, the number of copies per milliliter (copies/ml) of blood were measured at 6 months, 12 months and yearly after. Predictor variables of 154 patients considered

for the response variable were gender, follow-up time, weight at baseline, adherence, age at baseline, WHO stage, weight at follow-up time and CD4 count at baseline.

Thus, the model functions with covariates were specified as follow:

Mixed effect model

$$log(ViralLoad_{ij}) = \beta_0 + \beta_1 Time_{ij} + \beta_2 Gender_i + \beta_3 Baseline.Weight_i + \beta_4 Adherence_{2i} + \beta_5 Adherence_{3i} + \phi_1 Adherence_{2i} \times Time_{ij} + \phi_2 Adherence_{3i} \times Time_{ij} + \beta_6 Follow.up.weight_i + \beta_7 Age_i + \beta_8 WHO.Stage_{2i} + \beta_9 WHO.Stage_{3i} + \beta_{10} WHO.Stage_{4i} + \phi_3 Weight.at.followup_i \times Time_{ij} + \beta_{11} Baseline_{CD4_i} + w_{i0} + w_{i1} Time_{ij} + \epsilon_{ij}.$$

Generalized Estimating Equations (GEE)

$$log(ViralLoad_{ij}) = \beta_0 + \beta_1 Time_{ij} + \beta_2 Gender_i + \beta_3 Baseline.Weight_i + \beta_4 Adherence_{2i} + \beta_5 Adherence_{3i} + \phi_1 Adherence_{2i} \times Time_{ij} + \phi_2 Adherence_{3i} \times Time_{ij} + \beta_6 Follow.up.weight_i + \beta_7 Age_i + \beta_8 WHO.Stage_{2i} + \beta_9 WHO.Stage_{3i} + \beta_{10} WHO.Stage_{4i} + \phi_3 Weight.at.followup_i \times Time_{ij} + \beta_{11} Baseline\_CD4_i.$$

The analysis was conducted as follows:

First, fixed effects model was used to analyze viral load longitudinal measures using the lm function in the LMER package in R. Secondly, linear mixed-effects model was fit by REML in R. A comparison of the above stated models was made, and the most appropriate model was chosen.

In addition, the Generalized Estimating Equations (GEE) with different working correlation structures was used, with the analysis implemented using the geepack in R.

The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were used for model selection to evaluate and select the best model for fixed and linear mixed effects models, while for GEE model, the Quasi-Information Criterion (QIC) was used.

4.4. Results

4.4.1 Mixed Effects Model

Table 4.1 shows parameter estimation of a full model for both fixed effects model and linear mixed effects model. From the table, the fixed effect model had two variables that had a significant effect on the model which were gender and WHO-Stage IV with p-values of 0.002 and 0.029, respectively. On the other hand, the mixed effects model had only one significant variable which was baseline weight with a p-value of 0.049. Initially, non-significant variables were being avoided and removed one by one starting with the most non-significant variable then compare the two models. First, adherence with time interaction was removed since it was the most non-significant with p-values of 0.513 and 0.198, the model was fit again but there was no significant difference, then weight at follow-up and time were also removed. Having done this, the final reduced model with fewer variables is given below and the results are presented in **Table** 4.2.

***Table** 4.1: Parameter estimation of LM-Fixed Effects Model and Linear Mixed Effects Model*

	LM-Linear Model (Fixed Effects)	**LME-Linear Mixed Effects Model**

$Coef$	$log(Est)$	t-value	Std Error	p-value	$log(Est)$	t-value	Std Error	p-value
Fixed Effects:								
Intercept (β_0)	7.059	6.614	1.154	0.000	6.696	5.109	1.317	0.000
Time (β_1)	-0.072	-1.215	0.060	0.223	-0.065	-1.266	0.052	0.207
Gender (β_2):								
ref.Female								
Male	0.844	2.995	0.282	0.002	0.671	1.730	0.391	0.074
Baseline Weight (β_3)	-0.036	-1.639	0.022	0.102	-0.043	-1.973	0.022	0.049
Weight at Follow (β_4)	-0.021	-0.783	0.027	0.434	-0.0149	-0.662	0.023	0.508
Adherence:								
ref.Good								
Fair (β_5)	1.235	0.443	2.787	0.658	1.8429	0.861	2.142	0.389
Poor(β_6)	2.779	1.668	1.666	0.096	1.811	1.349	1.344	0.178
Adherence ×Time:								
ref.Good								
Fair(ϕ_1)	0.022	0.186	0.119	0.513	-0.096	-1.030	0.094	0.303
Poor(ϕ_2)	-0.065	-0.654	0.100	0.198	0.0402	-0.480	0.084	0.631
Weight at Follow up×Time (ϕ_3)	0.001	1.289	0.001	0.198	0.0009	1.246	0.001	0.214
Age (β_7):	0.020	1.372	0.014	0.170	0.0202	1.016	0.020	0.311
WHO.Stage:								
ref.Stage I								
Stage II (β_8)	-0.356	-1.140	0.312	0.255	-0.412	-0.897	0.459	0.370
Stage III(β_9)	-0.428	-1.266	0.338	0.206	-0.451	-0.978	0.462	0.328
Stage IV (β_{10})	1.729	2.180	0.793	0.029	-0.040	-0.479	1.171	0.632
Baseline CD4:								
ref. < 200								
≥ 200 (β_{11})	-0.588	-1.566	0.376	0.118	0.578	1.044	0.524	0.297
Random Effects:								
$\sigma^2_{w_0}$					4.507			
$\sigma^2_{w_1}$					0.004			
ρ					-0.579			
σ^2_{ϵ}					2.271			
Model selection:								
AIC	2067.295				2002.516			
BIC	2137.153				2083.977			
LogLik	-1016.648				-981.258			

The reduced model is as follows:

$$\log(ViralLoad_{ij}) = \beta_0 + \beta_1 Gender_i + \beta_2 Baseline.Weight_i + \beta_3 Adherence_{2i}$$

$$+ \beta_4 Adherence_{3i} + \beta_5 Age_i + \beta_6 WHO.Stage_{2i} + \beta_7 WHO.Stage_{3i}$$

$$+\beta_8 WHO.Stage_{4i} + \beta_9 Baseline\_CD4_i + w_{i0} + w_{i1} Time_{ij} + \epsilon_{ij}$$

***Table** 4.2: Parameter estimation of LM-Fixed Effects Model and Linear Mixed Effects Model for reduced model*

		LM-Linear Model (Fixed Effects)				**LME-Linear Mixed Effects Model**		
Coef	$log(Est)$	**t-value**	**S.E**	**p-value**	$log(Est)$	**t-value**	**S.E**	**p-value**
Fixed Effects:								
Intercept (β_0)	6.020	7.942	0.758	0.000	5.802	5.354	1.084	0.000
Gender (β_1):								
ref.Female								
Male	0.837	2.982	0.281	0.002	0.671	1.787	0.389	0.025
Baseline Weight (β_2)	-0.042	-4.801	0.009	0.000	-0.043	-3.495	0.012	0.001
Adherence								
ref.Good								
Fair (β_3)	1.619	1.652	0.980	0.099	-0.293	-0.403	0.727	0.687
Poor (β_4)	1.888	2.241	0.843	0.026	1.269	1.918	0.662	0.056
Age (β_5):	0.021	1.492	0.014	0.137	0.027	1.368	0.020	0.173
WHO.Stage								
ref.Stage I								
Stage II (β_6)	-0.356	-1.150	0.309	0.251	-0.379	-0.834	0.454	0.406
Stage III (β_7):	-0.426	-1.295	0.329	0.196	-0.406	-0.892	0.455	0.373
Stage IV (β_8):	1.758	2.238	0.786	0.026	1.584	1.373	1.154	0.172
Baseline CD4 :								
ref.< 200								
≥ 200 (β_9):	-0.5921	-1.607	0.368	0.1087	-0.493	-0.951	0.519	0.343
Random Effects:								
$\sigma^2_{w_0}$					4.679			
$\sigma^2_{w_1}$					0.004			
ρ					-0.619			
σ^2_{ϵ}					2.259			
Model selection:								
AIC	2059.405				1963.635			
BIC	2108.716				2024.903			
LogLik	-1017.702				-966.818			

In **Table** 4.2, three estimated variance components ($\sigma^2_{w_0}, \sigma^2_{w_1}$ and σ^2_{ϵ}) are shown. These are the random effects variance and the residual variance for the LMEM. The residual variance is $var(\epsilon_{ij}) = \sigma^2_{\epsilon} = 2.259$ and for the random effects, $var(w_0) = \sigma^2_{w_0} = 4.679$ and $var(w_1) = \sigma^2_{w_1} = 0.004$. Assuming normal distrubution of random effects $\sqrt{4.679} = 2.163$, implies that 95% of female patients had a mean viral load value between $e^{5.802 - 2.163 \times 1.96} = 4.771$ and $e^{5.802 + 2.163 \times 1.96} = 29959.338$. The total variability between patients is estimated as $\sigma^2_{w_0} + \sigma^2_{w_1} = 4.679 + 0.004 = 4.683$, while the total valiability within patients is 2.259.

In addition, the total variation in viral load values is estimated to be $2.259 + 4.683 = 6.942$. The proportion of total variability that is attributed to within-patient variation is given by $2.259/6.942 = 0.325(32.5\%)$, while the proportion of total variability attributed to between-patient variation of Viral load values is $4.683/6.942 = 0.675(67.5\%)$. Hence, less than half of the variation is explained by the residuals. The correlation $\rho = -0.619$ indicates a negative correlation between intercept and slope of linear time effect for the random part. This implies that when patient's intercept increase by one unit of standard deviation, their slope would decrease by 0.619 standard deviation.

All fixed effects parameters in both Fixed Effects and Mixed-effects models have specific interpretation. From table 4.2, the intercept $e^{\beta_0} = e^{6.020} = 411.579$ and $e^{\beta_0} = e^{5.802} = 330.961$, are the estimates of the i^{th} female patient mean viral load value given that her adherence was good, she was in clinical stage I and her CD4 count was below 200 cells/mm3 . In the same way, the coefficients for gender were $\beta_1 = 0.837$ and $\beta_1 = 0.671$, hence the mean viral load value for the i^{th} male patient were $e^{0.837} = 2.309$ and $e^{0.671} = 1.956$ times high than female patient and their difference was significant $(p - value = 0.003 \; and \; 0.025)$ at 5% level of

significance. The cofficients for fair adherence were $\beta_3 = 1.619$ and $\beta_3 = -0.293$, for poor adherence the coefficients are $\beta_4 = 1.888$ and $\beta_4 = 1.269$ implies that the viral load for a patient with fair adherence or poor adherence were $e^{1.619} = 5.048$ and $e^{-0.293} = 0.746$ or $e^{1.888} = 6.606$ and $e^{1.269} = 3.557$ times higher than that of a patient with good adherent. The coefficients for WHO stage II were $e^{\beta_6} = e^{-0.356} = 0.700$ and $e^{\beta_6} = e^{-0.379} = 0.685$ which indicates that the mean viral load was lower than that of patients who were in WHO stage I. The coefficients for WHO stage III were $e^{\beta_7} = e^{-0.426} = 0.653$ and $e^{\beta_7} = e^{-0.406} = 0.666$ which indicates that the mean viral load was lower than that of patients who were in WHO stage I . The coefficients for WHO stage IV were $e^{\beta_8} = e^{1.758} = 5.801$ and $e^{\beta_8} = e^{1.584} = 4.874$ which implies that the mean viral load was 5.801 and 4.874 times higher than that of patients who were in WHO stage I . The coefficients for CD4 count were $e^{-0.592} = 0.553$ and $e^{-0.493} = 0.611$ which implies that the mean viral load was lower in patients with less than 200 $cell/mm^3$.

Now a comparison between the models using AIC and log likelihood ratio test was peformed to choose the best model. In **Table** 4.1 and **Table** 4.2, the AIC value of fixed effects model decreased from 2067.295 to 2059.405 which shows that the model with fewer variables (reduced model) was improved as compared to the full model (model with all variables). This result was confirmed by the likelihood ratio test($p < 0.000$). Similarly, the AIC value for Linear Mixed Effects model also decreased from 2002.516 to 1963.635 which implies that the reduced model was better compared to the full model, hence the model with fewer variables was preferred. The mixed effects model had smaller AIC values in both models as compared to the fixed effects model, thus it was preferred to the fixed effects model.

4.4.2 Generalized Estimating Equations

In this subsection the data was analyzed using the Generalized Estimating Equations. Two different working correlation structures were considered (Unstructured and Independence) and compared. In order to build the GEE model, the model with all the variables was first considered. That is

$$log(ViralLoad_{ij}) = \beta_0 + \beta_1 Time_{ij} + \beta_2 Gender_i + \beta_3 Baseline.Weight_i + \beta_4 Adherence_{2i} + \beta_5 Adherence_{3i} + \phi_1 Adherence_{2i} \times Time_{ij} + \phi_2 Adherence_{3i} \times Time_{ij} + \beta_6 Follow.up.weight_i + \beta_7 Age_i + \beta_8 WHO.Stage_{2i} + \beta_9 WHO.Stage_{3i} + \beta_{10} WHO.Stage_{4i} + \phi_3 Weight.at.followup_i \times Time_{ij} + \beta_{11} Baseline\_CD4_i$$

***Table** 4.3: Comparison of unstructured and independence working correlation structures based on the full model*

	GEE with Unstructured correlation				GEE with Independence correlation			
Coef	$log(Est)$	**Naïve S.E**	**Robust S.E**	**p-value**	$log(Est$	**Naïve S.E**	**Robust S.E**	**p-value**
Intercept (β_0)	6.982	1.297	1.438	0.000	7.059	1.154	1.478	0.000
Time (β_1)	-0.076	0.057	0.051	0.188	-0.072	0.060	0.055	0.188
Gender (β_2): ref.Female								
Male	0.788	0.357	0.363	0.019	0.844	0.281	0.361	0.019
Baseline Weight (β_3)	-0.046	0.021	0.018	0.079	-0.036	0.022	0.0207	0.079
Adherence: ref.Good								
Fair (β_4)	0.653	2.405	7.746	0.838	1.235	2.787	6.055	0.838
Poor (β_5)	1.873	1.477	1.100	0.106	2.779	1.666	1.717	0.106
Follow-up Weight (β_6)	-0.011	0.024	0.021	0.136	-0.021	0.027	0.026	0.408
Adherence ×Time: ref.Good								
Fair (ϕ_1)	0.000	0.107	0.288	0.926	0.022	0.118	0.237	0.926
Poor (ϕ_2)	-0.037	0.095	0.036	0.195	-0.065	0.100	0.050	0.195

Weight at Follow up×Time ($\boldsymbol{\phi_3}$)	0.001	0.001	0.001	0.407	0.001	0.001	0.001	0.142
Age ($\boldsymbol{\beta_7}$)	0.021	0.018	0.021	0.338	0.019	0.014	0.021	0.215
WHO. Stage :								
ref.Stage I								
Stage II ($\boldsymbol{\beta_8}$)	-0.295	0.404	0.465	0.428	-0.356	0.312	0.449	0.428
Stage III ($\boldsymbol{\beta_9}$)	-0.435	0.423	0.394	0.273	-0.428	0.338	0.390	0.273
Stage IV ($\boldsymbol{\beta_{10}}$)	1.773	1.034	1.445	0.215	-0.428	0.793	1.395	0.215
Baseline CD4 ($\boldsymbol{\beta_{11}}$):								
ref.< 200								
≥ 200	-0.514	0.478	0.552	0.291	-0.589	0.376	0.558	0.291
Model selection:								
QIC	2468.9				2485.9			
Quasi-Likelihood	-1207.9				-1213.9			
Trace	26.500				29.100			

To compare the two working correlation structures, naïve and robust standard error estimates for both correlation structures were considered first to see how close the results were to each other. As shown in **Table** 4.3, naïve and robust standard error estimates for the unstructured correlation were close to each other as compared to those of the independence correlation. This meant that unstructured correlation was a good working correlation structure for the data. Comparing the two working correlation structures using Quasi-Likelihood Criterion (QIC), the unstructured working correlation had QIC of 2468.9 compared to 2485.9 value for the independence working correlation. Therefore, from both considerations, the model with unstructured working correlation structure was preferred.

Using the unstructured working correlation structure, significant variables were selected using p-values. From **Table** 4.3, weight at follow up, adherence with time interaction and weight at follow up with time interaction were not significant at 5% level of significance. As discussed earlier, non-

significant variables were removed one by one starting from the most non-significant. The most non-significant variable was adherence with time interaction with p-values of 0.999 and 0.304, it was therefore removed. This process was repeated for the second time whereby weight at follow up and weight at follow up with time interaction were removed and the final model was as given below.

$$log(ViralLoad_{ij}) = \beta_0 + \beta_1 Time_{ij} + \beta_2 Gender_i + \beta_3 Baseline.Weight_i + \beta_4 Adherence_{2i} + \beta_5 Adherence_{3i} + \beta_6 Age_i + \beta_7 WHO.Stage_{2i} + \beta_8 WHO.Stage_{3i} + \beta_9 WHO.Stage_{4i} + \beta_{10} Baseline\_CD4_i$$

***Table** 4.4: Comparison of unstructured and independence working correlation structures based on the reduced model*

	GEE with Unstructured correlation				**GEE with Independence correlation**			
Coef	$log(Est)$	**Naïve S.E**	**Robust S.E**	**p-value**	$log(Est)$	**Naïve S.E**	**Robust S.E**	**p-value**
Intercept (β_0)	5.932	0.998	1.148	0.000	5.981	0.784	1.136	0.000
Time (β_1)	-0.003	0.012	1.148	0.794	-0.02	0.011	0.011	0.839
Gender (β_2): ref.Female								
Male	0.788	0.356	0.362	0.032	0.838	0.281	0.358	0.019
Baseline Weight (β_3)	-0.042	0.011	0.556	0.001	-0.042	0.009	0.013	0.001
Adherence: ref.Good								
Fair (β_4)	0.588	0.881	2.177	0.787	1.603	0.985	1.668	0.337
Poor (β_5)	1.349	0.717	0.728	0.064	1.890	0.844	1.102	0.068
Age (β_6)	0.024	0.018	0.020	0.235	0.021	0.014	0.020	0.292
WHO. Stage : ref.Stage I								
Stage II (β_7)	-0.287	0.401	0.465	0.537	-0.359	0.310	0.454	0.428
Stage III (β_8)	-0.392	0.416	0.404	0.331	-0.425	0.329	0.401	0.288
Stage IV (β_{19})	1.829	1.028	1.428	0.200	1.748	0.788	1.369	0.202
Baseline CD4 (β_{10}): ref.< 200								
≥ 200	-0.551	0.472	0.557	0.322	-0.597	0.400	0.553	0.280

Model selection		
QIC	2464.3	2477.4
Quasi-Likelihood	-1213.5	-1218.5
Trace	18.6	20.2

As seen in **Table** 4.4, the naïve and robust standard error estimates for unstructured correlation were still close to each other as compared to those of the independence correlation. This implies that using the reduced model, the unstructured working correlation structure was the best for the study data. Using the unstructured working correlation structure model, the intercept ($e^{\beta_0} = 376.908$) is an estimate of the average viral load at baseline for i^{th} female patient which was significantly different from zero (p<0.00). Also, the coefficient for gender ($\beta_2 = 0.788$), indicates the average viral load in male patients was 2.199 times higher in male patients.

Similarly, comparing the two working correlation structures for the reduced models using QIC, the following values 2464.3 and 2477.4 for unstructured and independence working correlation structures were obtained respectively. Again, the model with unstructured working correlation structure was preferred.

4.5. Comparison of Mixed effects and GEE models

To compare the two models, their respective standard error estimates were used, for the GEE model the robust standard errors for unstructured correlation structure were used. To compare the models full model results for LMEM (**Table** 4.2) and GEE (**Table** 4.3) were used. This was because the most non-significant covariates were removed from the final models, so it was not possible to compare two models having different numbers of covariates. The standard error

estimates of LMEM were smaller than that of the GEE. In other words, the LMEM fitted the data with smaller disturbance than GEE, therefore, LMEM model was better than GEE.

4.6. Summary

This chapter modeled the change of viral load in HIV patients on ART using longitudinally measured viral load data. Since the data was correlated and continuous, two models (GEE and MRMs) were applied. In modeling with MRMs, two models were used which were fixed-effects models and mixed effects models. From the MRMs' final results model (**Table** 4.2), gender and baseline weight were found to be significant factors of viral load at 5% significance level. For GEE models, two correlation structures were chosen for modeling the data, which were unstructured and independence structures. From the results of the GEE (**Table** 4.4), only gender and baseline weight were found to be significant predictors for vrial load at 5% significance level. The findings are discussed in details in chapter 6.

CHAPTER 5 : MODELING VIRAL LOAD WITH RESPONSE MISSINGNESS AND COVARIATE MEASUREMENT ERROR

5.1. Background

In health sciences, and biostatistics fields, longitudinal studies are conducted in which repeated measurements are collected from the same subject over time in order to monitor disease progression and treatment outcomes. However, incomplete data are quite common in such studies, this is because subjects may not be available to be measured or observed at all the time points. Moreover, a subject can be missing at one follow-up time and then measured again at one of the next, resulting in non-monotone missing data patterns. Such data present considerable challenges in statistical inference for statisticians (Carroll et al., 2006; Yi, 2008).

In literature, there have been considerable interest in accommodating either incompleteness or covariate measurement error under random effects models. In addition, there have been extensive research on either covariate measurement error or missing responses, but relatively little work have been done to address both simultaneously, (Yi, Ma & Carroll, 2012). Furthermore, there is a need to fill up this gap as longitudinal data do often have both characteristics (Yi, Liu, & Wu, 2011). This problem has been discussed by several authors (Yi, Liu, & Wu, 2011; Huang & Dagne, 2012; Yi, 2008; Xiong and Yi, 2019).

Yi, Liu and Wu (2011) investigated the effects on inference when both missing responses and error in covariates exist. They proposed a two-stage modeling approach for generalized linear mixed

models in order to modulate the response process in connection with covariates and conducted simulation studies in order to assess its performance. They further demonstrated that substantial finite- sample biases would be induced if missingness and measurement error are not properly accounted for. This ignorance has an impact on estimation of every regression coefficient, including the covariate effect for error-prone covariate and the one for precisely measured covariate.

Huang and Dagne (2012), studied the simultaneous impact of skewness, missingness, and covariate measurement error by jointly modeling the response and covariate processes based on a flexible Bayesian semiparametric nonlinear mixed effects (SNLME). They found that it was important to take the CD4 measurement errors and viral load missing data into account. They further found that the missing viral load data were likely to be non-ignorable and, thus, estimates under a non-ignorable missing data model might be more reliable than those under an ignorable missing data model.

Yi (2008) proposed a simulation based marginal method to adjust for the bias induced by measurement error in covariates as well as by missingness in response. The proposed method focused on modeling the marginal mean and variance structures, and the missingness at random mechanism was assumed and it does not require the full specification of the distribution of the response variable.

Based on ignorable data, standard software for longitudinal data that accommodates unbalanced observations can be used. These include the SAS procedures MIXED, GLIMMIX, NLMIXED,

SPlus, and R functions lme and nlme (Ibrahim & Molenberghs, 2009). Such tools eliminate complete-case bias by incorporating all available information. However, in the non-ignorable case, methods that do not model the missing data mechanism are subject to bias. Xiong and Yi, 2019 developed an R package, called swgee, which implements the method proposed by Yi (2008). Moreover, this package includes additional implementation steps which extend the setting considered by Yi (2008). Xiong and Yi, (2019) mentioned that the swgee method does significantly improve the performance of the GEE analysis. This package employs the simulation extrapolation (SIMEX) algorithm to account for the effect of measurement error in covariates.

Thus, in this chapter the swgee package developed by Xiong and Yi, 2019 to adjust for the bias induced by measurement error in covariates as well as missingness in response variable was used to analyze the longitudinal viral load data. This was also implemented in order to significantly improve the results of the GEE model in chapter 4.

5.2. Statistical Methods

5.2.1. Notations

Suppose there were n patients under a longitudinal study that collects n_i repeated viral load measurements for i^{th} patient ($i = 1, \dots, n$). Now, let Y_{ij} denote viral load measurements for i^{th} patient at time ($j = 1, \dots, n_i$) with $Y_i = \left(Y_{i1}, \dots, Y_{in_i}\right)^T$. Let $X_i = \left(X_{i1}, \dots, X_{in_i}\right)^T$ denote vector of the time-invariant and time-varying covariates measured at baseline and subsequent follow-up time for patient i. Also let, $Z_i = \left(Z_{i1}, \dots, Z_{in_i}\right)^T$ be the vector of covariates which are error-free. Suppose D_{ij} is the missingness indicator of i^{th} patient at time j that takes 1 for presence and 0 for absence, where $D_i = \left(D_{i1}, \dots, D_{in_i}\right)^T$.

In this study there was only one type of missing data mechanism, that is Y_{ij} is missing due to dropout if Y_{ij} and all subsequent measurement $(Y_{ij+1}, \dots, Y_{in_i})$ are missing, meaning the patient missed n_i^{th} measurements and never came back for later measurements. Although it could be possible that some dropouts might have reappeared subsequently if the study had continued beyond n_i measurements, this could not be identified based on the observed data and thus it will not be considered in this study. The missing indicator is defined by

$$D_{ij} = \begin{cases} D_{obs} = 1, & if\ \ Y_{ij}\ \ is\ observed \\ D_{mis} = 0, & if\ Y_{ij}\ is\ missing \end{cases}.$$

5.2.2. **Outcome Process Model**

For $i = 1, \dots, n$ and $j = 1, \dots, n_i$, let $\mu_{ij} = E(Y_{ij}|\boldsymbol{X}_i, \boldsymbol{Z}_i)$ and $v_{ij} = var(Y_{ij}|\boldsymbol{X}_i, \boldsymbol{Z}_i)$ be conditional expectation and variance of Y_{ij}, given the covariates X_i and Z_i, respectively. The influence of the covariates on the marginal response mean modeled by means of regression model:

$$g(\mu_{ij}) = X_{ij}^T \beta_x + Z_{ij}^T \beta_z, \tag{5.1}$$

where $\boldsymbol{\beta} = (\beta_x^T, \beta_z^T)^T$ is the vector of regression parameters and $g(.)$ is a specified monotone function.

To model the variance of Y_{ij}, consider

$$v_{ij} = h(\mu_{ij}; \varphi), \tag{5.2}$$

where $h(\mu_{ij}; \varphi)$ is a given function and φ is the dispersion parameter that is known or to be estimated. The parameter φ is treated as known here with emphasis setting on estimation of the β parameter.

Model for the missing data mechanism

Suppose that $D_{i1} = 1$ for every patient i. In order to reflect the dynamic nature of the observation process over time, a MAR mechanism for the missing process is assumed. This implies that, given the covariates, the missingness probability depends on the observed responses but not unobserved response components (Little & Rubin, 2002). Let $\pi_{ij} = P(D_{ij} = 1|D_{ij-1} = 1, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{Y}_i)$ and $\lambda_{ij} = P(D_{ij} = 1|\boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{Y}_i)$, then

$$\lambda_{ij} = \prod_{i=1}^{n} \prod_{j=2}^{n_i} \pi_{ij}. \tag{5.3}$$

The logistic regression model for dropout process is given by:

$$logit(\pi_{ij}) = \boldsymbol{u}'_{ij}\boldsymbol{\alpha}, \tag{5.4}$$

where $\boldsymbol{u_{ij}}$ is the vector consisting of information of the covariates $\boldsymbol{X}_i, \boldsymbol{Z}_i$ and the observed responses.

Suppose that $Y_i = (Y_{obs;ij}, Y_{mis;ij})$, where Y_i consists of observed and missing viral load measurements and $f(\cdot \mid \cdot)$ as a conditional density function. Let $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ be the unknown vectors of population parameters of GEE model repeated measures and of the missingness mechanism, respectively.

Let $\pi_{ij}(Y_{ij}) = \Pr(D_{ij} = 1|Y_{ij}^T, X_i^T, Z_i^T)$ be the missing probability function at time j. The missingness probability is modelled by $logit(\pi_{ij}) = \alpha_0 + \alpha_a Y_{ij} + \alpha_b X_i^T + \alpha_c Z_i^T$, where $\boldsymbol{\alpha} = \{\alpha_a, \alpha_b, \alpha_c\}$ is a vector that describes how the missingness at follow-up j depends on the measurements $Y_i^T = \{Y_{i1}, .., Y_{in_i}\}$, and in particular, α_c is a $p \times 1$ parameter vector that governs the missingness and covariate associations. Therefore, $f(D_i|\boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\phi})$ is the density function of missingness indictor, with the missing data assumed be generated by the logistic model

$$Prob\left(D_{ij}=1|Y_{ij}^{T},X_{i}^{T},Z_{i}^{T}\right)=\frac{\exp\left(\alpha_{0}+\alpha_{a}Y_{ij}+\alpha_{b}X_{i}^{T}+\alpha_{c}Z_{i}^{T}\right)}{1+\exp\left(\alpha_{0}+\alpha_{a}Y_{ij}+\alpha_{b}X_{i}^{T}+\alpha_{c}Z_{i}^{T}\right)} \quad (5.5)$$

Note that, when Y_i is not observed, $Y_{miss,ij}$ is sampled from its conditional distribution.

Measurement error model

Let $\boldsymbol{W}_{ij}$ be the observed measurements of the covariates $\boldsymbol{X}_{ij}$. The covariates $\boldsymbol{X}_{ij}$ and their observed measurements $\boldsymbol{W}_{ij}$ are assumed to follow a classical additive measurement error model:

$$\boldsymbol{W}_{ij}=\boldsymbol{X}_{ij}+\boldsymbol{\epsilon}_{ij}, \quad (5.6)$$

where the $\boldsymbol{\epsilon}_{ij}$ are independent of $\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}$ and $\boldsymbol{Y}_i$. Note that $\boldsymbol{\epsilon}_{ij}$ follows $N(\mathbf{0}, \boldsymbol{\Sigma}_{\epsilon})$ with covariance matrix $\boldsymbol{\Sigma}_{\epsilon}=\begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{21} & \sigma_1^2\end{pmatrix}$.

5.3. Application

The study population includes all HIV/AIDS patients initiated on antiretroviral therapy (ART) follow-up from January 2015 to December 2017 at the Luderitz Hospital in the !Karas region of Namibia. One response variable was considered in this study, which was the longitudinal viral load of HIV adult patients initiated on ART. A viral load test is used to measure the amount of HIV in a sample of blood, the number of copies per milliliter (copies/ml) of blood were measured at 6 months, 12 months and yearly after. Predictor variables of 154 patients considered for the response variable were gender, follow-up time, weight at baseline, adherence, Age at baseline, WHO stage, weight at follow-up time and CD4 count at baseline.

The response and the covariates were specified by the following regression model:

$$log(ViralLoad_{ij}) = \beta_0 + \beta_1 Time_{ij} + \beta_2 Gender_i + \beta_3 Baseline.Weight_i + \beta_4 Adherence_i$$
$$+ \beta_5 Weight\_at\_followup_i + \beta_6 Adherence_i \times Time_{ij}$$
$$+ \beta_7 Weight.at.followup_i \times Time_{ij} + \beta_8 Age_i + \beta_9 WHO.Stage_{2i}$$
$$+ \beta_{10} WHO.Stage_{3i} + \beta_{11} WHO.Stage_{4i} + \beta_{12} CD4_i,$$

where $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9, \beta_{10}, \beta_{11}$ and β_{12} are regression coefficients.

The error in both risk factors Adherence ($X_{ij,1}$) and weight at follow-up ($X_{ij,2}$) is assumed by $\boldsymbol{W}_{ij} = \boldsymbol{X}_{ij} + \boldsymbol{\epsilon}_{ij}$. Adherence and weight at follow up were measured together with viral load, thus if the patient did not go for viral load these too will be missing. One would want to assess the effect of baseline age Z_{ij} in the missing data process. Hence, in this study, the missing data process was specified by the logistic regression model:

$$logit\ \pi_{ij} = a_0 + \alpha_a Y_{ij-1} + \alpha_b X_{ij-1,1} + \alpha_c X_{ij-1,2} + \alpha_d Z_{ij-1},$$

for $j = 1,2,3,4$.

The concern here is how measurement error in Weight at follow-up and Adherence ART impacts the estimation of parameter $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_9, \beta_{10}, \beta_{11}, \beta_{12})^T$. The following settings were used; $B = 50, \lambda_M = 2$ and $M = 5$. In this application, the covariance matrix of measurement errors in the measurement error model was set as

$$\Sigma_\epsilon = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_1^2 \end{pmatrix} = \begin{pmatrix} 0.25 & 0 \\ 0 & 0.25 \end{pmatrix}.$$

***Table** 5.1: Coefficients associated with the response process*

	SWGEE			
Parameters	$\boldsymbol{log(Est)}$	**S.E**	**t-value**	**p-value**
Intercept (β_0)	4.060	2.225	1.82	0.068
Time (β_1)	-0.012	0.074	-0.170	0.866
Gender (β_2):				

ref.Female				
Male	0.831	0.357	2.33	0.020
Baseline Weight (β_3)	-0.027	0.023	-1.200	0.229
Adherence (β_4)	2.837	1.379	2.060	0.040
Weight at follow up (β_5)	-0.032	0.027	-1.170	0.241
Adherence ×Time (β_6)	-0.085	0.051	-1.680	0.094
Weight at follow up ×Time (β_7)	0.002	0.001	2.20	0.028
Age (β_8)	0.024	0.020	1.220	0.221
WHO. Stage :				
ref.Stage I				
Stage II (β_9)	0.194	0.522	1.220	0.711
Stage III (β_{10})	-0.152	0.421	-0.360	0.718
Stage IV (β_{11})	1.986	1.931	1.030	0.304
Baseline CD4 (β_{12}):				
ref.< 200				
≥ 200	-0.812	0.074	-0.170	0.866

***Table** 5.2: Coefficients associated with the missing process*

SWGEE				
Parameters	**Est**	**S.E**	**t- value**	**p-value**
α_0	2.580	0.874	2.950	0.003
α_a	-0.000	0.000	-0.050	0.956
α_b	-0.710	0.517	-1.370	0.170
α_c	-0.010	0.008	-1.240	0.216
α_d	-0.004	0.013	-0.310	0.757

Table 5.2 shows the Coefficients associated with the missing process. From this table, the estimate of α_a is -0.000 with a p-value of 0.956, which suggests that the probability that a patient will miss their next follow up was not significantly related to their previous viral load. The estimate of α_b is -0.710 with a p-value of 0.170, suggests that the probability that a patient will miss the next follow up was not significantly influenced by their adherence status. This implies that patients might have come for follow up but their decisions were not based on their previous adherence status. The estimate of α_c is -0.010 with a p-value 0.216, indicates that the effect of weight at

follow-up in the missing data process was not statistically significant. The estimate of α_d is -0.004 with a p-value of 0.757, this means that baseline age did not have an impact on the missingness model. In this analysis, some measurements of the viral load were missing while gender, baseline weight and WHO stage were completely observed. The missing values of viral load were MCAR since the probability of observing viral load was independent of gender, baseline weight and WHO stage and the values of viral load that were observed or would have been observed. Therefore, these results suggested that the assumption of MCAR was more appropriate than that of MAR.

5.4. Summary

Problems associated with incomplete gathered data in longitudinal and clinical trials have received considerable attention in recent times (Molenberghs and Verbeke 2000; Fitzmaurice et al. 2004; Molenberghs and Verbeke 2005; Molenberghs and Kenward 2007; Daniels and Hogan 2008; Fitzmaurice et al. 2008). However, analysis of longitudinal data with response missingness and covariate measurement error received little attention. In this chapter, a longitudinal study on viral load of HIV/AIDS patients was used to account for effects of response missingness and covariate measurement error on estimation of response model parameters. The swgee R package by Xiong and Yi, 2019 was used for all the analysis. These findings are discussed in details in chapter 6.

CHAPTER 6 : DISCUSSION, CONCLUSION AND RECOMMENDATIONS

6.1. Discussion

This study aimed to accomplish four research objectives, which were:

- To explore the average change of HIV viral load in patients on ART over time
- To model the change in viral load over time using Mixed effects models and Generalized estimating equations.
- To investigate the effects of clinical factors and demographic characteristics on viral load.
- To model viral load longitudinal data adjusting for the bias induced by measurement error in covariates as well as missingness in response variable.

This was a retrospective cohort study which used data from 154 patients initiated on ART at Luderitz hospital between January 2015 and December 2017. Adherence to ART among these patients was quite good over time and different models (Mixed effects model, GEE and models for missing data and covariates with measurement error) for modeling correlated data were used in the study with the best models selected using AIC and QIC. The results were close to one another, but the model which incorporated missing data and measurement error was most preferred.

After accounting for missingness and measurement error, adherence to ART was found to have a significant effect on viral load. Viral load decreased and suppression over time was associated with consistent adherence to antiretroviral therapy (ART). Patients with good adherence throughout

ART tend to have their viral load suppressed within 12 months of the therapy. Despite adherence being found to be an important factor in influencing treatment outcome, there are no standard guidelines for its measurement (Nachega et.al, 2014).

The analysis of gender difference found that viral load in female patients was significantly low than that of male patients. This was supported by the fact that the progression rates to AIDS and clinical manifestations of diseases associated with HIV infection differs between women and men because of biological and socioeconomic factors (Nicastri & Angeletti, 2005). According to the WHO guidelines, virological failure is observed when patients sustain a viral load > 1000 copies/ml after 6 to 12 months of ART, the average viral load of males in this study was still >1000 at 12 months of ART. In this study, 85% of females started ART at stage 1 and 2, whereas only 65% of the males started therapy at early stages. In general, females were found to start ART at a less advanced disease stage, with higher CD4 and at earlier stage of the infection. Similar findings were reported in other studies (Cornell, Schomaker & Garone, 2012).

Although baseline CD4 count did not have a significant effect on viral load in this study, it is believed that patients who started ART with low CD4 count often have high viral load and usually take long to have their viral load suppressed or might have worse treatment outcomes. CD4 counts are collected for clinical reasons to evaluate the stage of HIV infection. Patients who enter treatment at a significantly more advanced stage of HIV infection get predisposed to increased mortality and worse treatment outcomes (Mosha et.al, 2013). In this study the records of baseline CD4 counts were poor, this could be because baseline CD4 count was not a requirement for one to start ART thus it was not measured for many patients. Furthermore, this study found that WHO staging did not have a significant effect on time-varying viral load. According to Jaffar et al, (2008)

the accuracy of the WHO clinical stage criteria is unknown in a normal health service delivery setting, where training and clinical support of clinical staff can vary.

The findings of this study were limited by a few issues. The data did not reflect what happened to the patients who did not have viral load at all time-points and there was a high percentage (64.5%) of loss to follow-up patients by the end of the 36th month, but this may however be attributed to attrition. Also, some clinical information that may have been useful in the modeling of viral load in response to ART such as ARV regimen and socioeconomic factors were missing in the ART data. Furthermore, the patient cohort used in this study was relatively small (154) and might not have represented all HIV-infected patients on ART at the Luderitz hospital.

6.2. Conclusion

The main objective of this study was to model and study the change in longitudinally measured viral load given time-varying adherence of patients on ART at the Luderitz Hospital in the !Karas region of Namibia. The study found that viral load was higher in male patients at baseline and takes longer (approximately 24 months) for them to achieve viral suppression compared to female patients on ART. In addition, gender, weight at follow-up and adherence were found to be significant predictors of patients' viral load at 5% significance level, although baseline weight was found to be significant in the mixed effects models and the GEE models for both correlation structures. This could be due to missingness and measurement error of covariates which were not accounted for in these models. In conclusion, viral load in patients on ART differ by patients' demographic characteristics (age and gender) and clinical characteristics (baseline CD4 and adherence to ART).

6.3 Recommendations

It is crucial for the special disease program to continue monitoring patients on ART and their viral load for treatment outcome. This study recommended that HCWs should put in an effort in monitoring viral load of patients on ART on the scheduled viral load visit as well as in recording it. In addition, they should strengthen the tracing of patients who have missed their viral load appointment or are lost to follow up. This study did not have access to the patients' ARV regimen and socioeconomic factors, thus it recommended that future researchers may add these as they might have a huge effect on adherence to ART as well as on viral load. Furthermore, the policy makers should come up with a standard approach for measuring adherence since tablet count does not really reflect the true picture- a patient coming back with a correct number of tablets does not necessarily mean that they have taken their medicine. Project managers should put in place a well monitored data cleaning system to ensure their data is of quality.

REFERENCES

Abebe TN. (2020). Evaluation of CD4 Cell Progression among HIV Infected Children Initiating ART: A Case of Adama Referral Hospital and Medical College, Ethiopia.

Achappa, B., Madi, D., Bhaskaran, U., Ramapuram, J. T., Rao, S., & Mahalingam, S. (2013). Adherence to antiretroviral therapy among people living with HIV. *North American journal of medical sciences*, *5*(3), 220.

Akaike, H. (1974). *A new look at statistical model identification*. IEEE Transactions on Automatic Control AC-19: 716–723.

Barnard, J., & Meng, X. L. (1999). Applications of multiple imputation in medical studies: from AIDS to NHANES. *Statistical methods in medical research*, *8*(1), 17-36.

Barnard, J., & Rubin, D. B. (1999). Miscellanea. Small-sample degrees of freedom with multiple imputation. *Biometrika*, *86*(4), 948-955.

Barnett, A. G., Koper, N., Dobson, A. J., Schmiegelow, F., & Manseau, M. (2010). Using information criteria to select the correct variance–covariance structure for longitudinal data in ecology. *Methods in Ecology and Evolution*, *1*(1), 15-24.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J. S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution*, *24*(3), 127-135.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, *88*(421), 9-25.

Bryan, M., & Heagerty, P. J. (2014). Direct regression models for longitudinal rates of change. *Statistics in medicine*, *33*(12), 2115–2136. doi:10.1002/sim.6102

Buonaccorsi, J. P. (2010). *Measurement error: models, methods, and applications*. Chapman and Hall/CRC.

Campos, L. F., Glickman, M. E., & Hunter, K. B. (2018). Measuring Effects of Medication Adherence on Time-Varying Health Outcomes using Bayesian Dynamic Linear Models. *arXiv preprint arXiv:1811.11072*.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. CRC press.

CDC, 2019. HIV [PDF file]. Retrieved from https://www.cdc.gov/globalhealth/countries/namibia/pdf/Namibia_Factsheet.pdf

Chen, R. (2012). Bayesian Inference on Mixed-effects Models with Skewed Distributions for HIV longitudinal data. PhD thesis, University of South Florida, 1 2012.

Chendi, B. H., Assoumou, M. C. O., Jacobs, G. B., Yekwa, E. L., Lyonga, E., Mesembe, M., ... & Ikomey, G. M. (2019). Rate of viral load change and adherence of HIV adult patients treated with Efavirenz or Nevirapine antiretroviral regimens at 24 and 48 weeks in Yaoundé, Cameroon: a longitudinal cohort study. *BMC infectious diseases*, *19*(1), 194.

Chinomona, A., & Mwambi, H. (2015). Multiple imputation for non-response when estimating HIV prevalence using survey data. *BMC public health*, *15*(1), 1059.

Claeskens, G., & Hjort, N. L. (2008). Model selection and model averaging. *Cambridge Books*.

Cohen, M. S., Hellmann, N., Levy, J. A., DeCock, K., & Lange, J. (2008). The spread, treatment, and prevention of HIV-1: evolution of a global pandemic. *The Journal of clinical investigation*, *118*(4), 1244-1254.

Commenges, D., Jolly, D., Drylewicz, J., Putter, H., & Thiébaut, R. (2011). Inference in HIV dynamics models via hierarchical likelihood. *Computational Statistics & Data Analysis*, *55*(1), 446-456.

Cornell, M., Schomaker, M., Garone, D. B., Giddy, J., Hoffmann, C. J., & Lessells, R. (2012). International Epidemiologic Databases to Evaluate AIDS Southern Africa Collaboration Gender differences in survival among adult patients starting antiretroviral therapy in South Africa: a multicentre cohort study. *PLoS Med*, *9*(9), e1001304.

Cui, J., Antoniou, A. C., Dite, G. S., Southey, M. C., Venter, D. J., Easton, D. F., ... & Hopper, J. L. (2001). After BRCA1 and BRCA2—what next? Multifactorial segregation analyses of three-generation, population-based Australian families affected by female breast cancer. *The American Journal of Human Genetics*, *68*(2), 420-431.

Dagne, G., & Huang, Y. (2012). Bayesian inference for a nonlinear mixed-effects Tobit model with multivariate skew-t distributions: application to AIDS studies. *The international journal of biostatistics*, *8*(1), 10.1515/1557-4679.1387 /j/ijb.2012.8.issue-1/1557-4679.1387/1557-4679.1387.xml. doi:10.1515/1557-4679.1387

Daly, H., Ortiz, A., Dwivedi, Y. K., Paul, R. J., Santos, J., & Sarriegi, J. M. (2008). Developing a Dynamic View of Broadband Adoption. In *Handbook of Research on Global Diffusion of Broadband Data Transmission* (pp. 322-336). IGI Global.

Daniels, M. J., & Hogan, J. W. (2008). *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*. CRC Press.

Diggle, P., Diggle, P. J., Heagerty, P., Liang, K. Y., Heagerty, P. J., & Zeger, S. (2002). *Analysis of longitudinal data*. Oxford University Press.

Ding, A. A., & Wu, H. (2001). Assessing antiviral potency of anti-HIV therapies in vivo by comparing viral decay rates in viral dynamic models. *Biostatistics*, *2*(1), 13-29.

Ding, J., Tarokh, V., & Yang, Y. (2018). Model selection techniques: An overview. *IEEE Signal Processing Magazine*, *35*(6), 16-34.

Druyts, E., Dybul, M., Kanters, S., Nachega, J., Birungi, J., Ford, N., ... & Mills, E. J. (2013). Male Gender and the risk of mortality among individuals enrolled in antiretroviral therapy programs in Africa: a systematic review and meta-analysis. *Aids*, *27*(3), 417-425.

Duncan, S. C., & Duncan, T. E. (1994). Modeling incomplete longitudinal substance use data using latent variable growth curve methodology. *Multivariate Behavioral Research*, *29*(4), 313-338.

Fan, J., Huang, T., & Li, R. (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *Journal of the American Statistical Association*, *102*(478), 632-641.

Faraway, J. J. (2014). *Linear models with R*. CRC press.

Fitzgerald, A. P., DeGruttola, V. G., & Vaida, F. (2002). Modelling HIV viral rebound using non-linear mixed effects models. *Statistics in Medicine*, *21*(14), 2093-2108.

Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (Eds.). (2008). *Longitudinal data analysis*. CRC press.

Fox, M. P., & Rosen, S. (2010). Patient retention in antiretroviral therapy programs up to three years on treatment in sub-Saharan Africa, 2007–2009: systematic review. *Tropical medicine & international health*, *15*, 1-15.

Fuller, W. A. (2009). *Measurement error models* (Vol. 305). John Wiley & Sons.

Gad, A. M., & El Kholy, R. B. (2012). Generalized linear mixed models for longitudinal data. *International Journal of Probability and Statistics*, *1*(3), 41-47.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology*, *60*, 549-576.

Guedj, J., Thiébaut, R., & Commenges, D. (2007). Practical identifiability of HIV dynamics models. *Bulletin of mathematical biology*, *69*(8), 2493-2513.

Gumedze, F. N., & Dunne, T. T. (2011). Parameter estimation and inference in the linear mixed model. *Linear Algebra and its Applications*, *435*(8), 1920-1944.

Gustafson, P. (2003). *Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments*. Chapman and Hall/CRC.

Hardin, J. W., and J. M. Hilbe. (2003). *Generalized Estimating Equations*. Boca Raton, FL:

Hedeker, D. (2005). Generalized linear mixed models. *Encyclopedia of statistics in behavioral science*.

Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological methods*, *2*(1), 64.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis* (Vol. 451). John Wiley & Sons.

Herbeck, J. T., Mittler, J. E., Gottlieb, G. S., & Mullins, J. I. (2014). An HIV epidemic model based on viral load dynamics: value in assessing empirical trends in HIV virulence and community viral load. *PLoS Comput Biol*, *10*(6), e1003673.

Ho, D. D., Neumann, A. U., Perelson, A. S., Chen, W., Leonard, J. M., & Markowitz, M. (1995). Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature*, *373*(6510), 123.

Huang, Y., & Dagne, G. (2012). Bayesian semiparametric nonlinear mixed-effects joint models for data with skewness, missing responses, and measurement errors in covariates. *Biometrics*, *68*(3), 943-953.

Huang, Y., & Lu, T. (2008). Modeling long-term longitudinal HIV dynamics with application to an AIDS clinical study. *The Annals of Applied Statistics*, *2*(4), 1384-1408.

Huang, Y., Dagne, G. A., Zhou, S., & Wang, Z. (2015). Piecewise mixed-effects models with skew distributions for evaluating viral load changes: a Bayesian approach. *Statistical methods in medical research*, *24*(6), 730-746.

Huang, Y., Liu, D., & Wu, H. (2006). Hierarchical Bayesian methods for estimation of parameters in a longitudinal HIV dynamic system. *Biometrics*, *62*(2), 413-423.

Huang, Y., Rosenkranz, S. L., & Wu, H. (2003). Modeling HIV dynamics and antiviral response with consideration of time-varying drug exposures, adherence and phenotypic sensitivity. *Mathematical biosciences*, *184*(2), 165-186.

Huang, Y., Wu, H., Holden-Wiltse, J., & Acosta, E. P. (2011). A dynamic Bayesian nonlinear mixed-effects model of HIV response incorporating medication adherence, drug resistance and covariates. *The annals of applied statistics*, *5*(1), 551.

Huang, Y., Yan, C., Wu, H., & Zhang, X. (2014). Simultaneous Inference for HIV Dynamic Models with Skew-t Distribution Incorporating Mismeasured Covariate and Multiple Treatment Factors. *Statistics in Biopharmaceutical Research*, *6*(3), 213-228.

Hughes, J. P. (1999). Mixed effects models with censored data with application to HIV RNA levels. *Biometrics*, *55*(2), 625-629.

Ibrahim, J. G., & Molenberghs, G. (2009). Missing data methods in longitudinal studies: a review. *Test*, *18*(1), 1-43.

ICAP, 2016. STANDARD OPERATING PROCEDURES ON VIRAL LOAD MONITORING FOR ICAP CLINICAL STAFF AND HEALTH CARE WORKERS [PDF file]. Retrieved from http://files.icap.columbia.edu/files/uploads/VL-SOP__Revised_July_2016.pdf

Jacobsen, M. M., & Walensky, R. P. (2016). Modeling and cost-effectiveness in HIV prevention. *Current HIV/AIDS Reports*, *13*(1), 64-75.

Jaffar, S., Birungi, J., Grosskurth, H., Amuron, B., Namara, G., Nabiryo, C., & Coutinho, A. (2008). Use of WHO clinical stage for assessing patient eligibility to antiretroviral therapy in a routine health service setting in Jinja, Uganda. *AIDS research and therapy*, *5*(1), 4.

Johnston, V., Fielding, K. L., Charalambous, S., Churchyard, G., Phillips, A., & Grant, A. D. (2012). Outcomes following virological failure and predictors of switching to second-line antiretroviral therapy in a South African treatment program. *Journal of acquired immune deficiency syndromes (1999)*, *61*(3), 370–380. https://doi.org/10.1097/QAI.0b013e318266ee3f.

Ke, C., & Wang, Y. (2001). Semiparametric nonlinear mixed-effects models and their applications. *Journal of the American Statistical Association*, *96*(456), 1272-1298.

Kim, J., Lee, E., Park, B. J., Bang, J. H., & Lee, J. Y. (2018). Adherence to antiretroviral therapy and factors affecting low medication adherence among incident HIV-infected individuals during 2009–2016: a nationwide study. *Scientific reports*, *8*(1), 3133.

Koulai, L., Presanis, A., Murphy, G., Suligoi, B., & De Angelis, D. (2017). Quantifying the recency of HIV infection using multiple longitudinal biomarkers. *arXiv preprint arXiv:1706.02508*.

Kuhn, E., & Lavielle, M. (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics & Data Analysis*, *49*(4), 1020-1038.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 963-974.

Lavielle, M., & Mentré, F. Estimation of population pharmacokinetic parameters of saquinavir in HIV patients and covariate analysis with MONOLIX.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.

Liu, W., & Wu, L. (2007). Simultaneous inference for semiparametric nonlinear mixed-effects models with covariate measurement errors and missing responses. *Biometrics*, *63*(2), 342-350.

Lyles, R. H., Williams, J. K., & Chuachoowong, R. (2001). Correlating two viral load assays with known detection limits. *Biometrics*, *57*(4), 1238-1244.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Methods*. 2nd edition. London: Chapman and Hall. Pp. 150-300.

McCulloch, C. E., & Neuhaus, J. M. (2014). Generalized linear mixed models. *Wiley StatsRef: Statistics Reference Online*.

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction*. Guilford Press.

Migon, H. S., Gamerman, D., Lopes, H. F., & Ferreira, M. A. (2005). Dynamic models. *Handbook of statistics*, *25*, 553-588.

MoHSS, 2014. National Guidelines for Antiretroviral Therapy Revised Fourth Edition August 2014 [PDF file]. Retrived from http://www.oit.org/wcmsp5/groups/public/---ed_protect/---protrav/---ilo_aids/documents/legaldocument/wcms_140598.

MoHSS, 2016. National Guidelines for Antiretroviral Therapy Fourth Edition August 2016 [PDF file]. Retrieved from http://92.222.142.204/Portals/0/adam/Content/VVys6XEqAkiCUujlnxr3qA/File/na_national_guidelines_art.pdf

Molenberghs, G., & Kenward, M. (2007). *Missing data in clinical studies* (Vol. 61). John Wiley & Sons.

Molenberghs, G., & Verbeke, G. (2000). *Linear mixed models for longitudinal data*. New York, NY: Springer New York.

Molenberghs, G., & Verbeke, G. (2001). A review on linear mixed models for longitudinal data, possibly subject to dropout. *Statistical Modelling*, *1*(4), 235-269.

Mosha, F., Muchunguzi, V., Matee, M., Sangeda, R. Z., Vercauteren, J., Nsubuga, P., ... & Vandamme, A. M. (2013). Gender differences in HIV disease progression and treatment outcomes among HIV patients one year after starting antiretroviral treatment (ART) in Dar es Salaam, Tanzania. *BMC public health*, *13*(1), 38.

Nachega, J. B., Parienti, J. J., Uthman, O. A., Gross, R., Dowdy, D. W., Sax, P. E., ... & Giordano, T. P. (2014). Lower pill burden and once-daily antiretroviral treatment regimens for HIV infection: a meta-analysis of randomized controlled trials. *Clinical infectious*

NAMPHIA, 2018. Namibia Population-Based HIV Impact Assessment Namphia 2017 [PDF file]. Retrieved from https://phia.icap.columbia.edu/wpcontent/uploads/2018/10/33462%E2%80%A2NAMPHIA-SS_A4_B.v41.pdf

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, *135*(3), 370-384.

Nicastri, E., Angeletti, C., Palmisano, L., Sarmati, L., Chiesi, A., Geraci, A., ... & Vella, S. (2005). the Italian Antiretroviral Treatment Group. Gender differences in clinical progression of HIV-1-infected individuals during long-term highly active antiretroviral therapy. *AIDS*, *19*(6), 577-583.

Ojo, O. B., Lougue, S., & Woldegerima, W. A. (2017). Bayesian generalized linear mixed modeling of Tuberculosis using informative priors. *PloS one*, *12*(3), e0172580.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, *57*(1), 120-125.

Pardo, M. C., & Alonso, R. (2019). Working correlation structure selection in GEE analysis. *Statistical Papers*, *60*(5), 1447-1467.

Perelson, A. S., & Nelson, P. W. (1999). Mathematical analysis of HIV-1 dynamics in vivo. *SIAM review*, *41*(1), 3-44.

Pinheiro, J. C., & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of computational and Graphical Statistics*, *4*(1), 12-35.

Polis, M. A., Sidorov, I. A., Yoder, C., Jankelevich, S., Metcalf, J., Mueller, B. U., & Dimitrov, D. S. (2001). Correlation between reduction in plasma HIV-1 RNA concentration 1 week after start of antiretroviral treatment and longer-term efficacy. *The Lancet*, *358*(9295), 1760-1765.

Putter, H., Heisterkamp, S. H., Lange, J. M. A., & De Wolf, F. (2002). A Bayesian approach to parameter estimation in HIV dynamical models. *Statistics in medicine*, *21*(15), 2199-2214.

Rabe-Hesketh, S., & Skrondal, A. (2008). *Multilevel and longitudinal modeling using Stata*. STATA press.

Rose, C. E., Gardner, L., Craw, J., Girde, S., Wawrzyniak, A. J., Drainoni, M. L.,. & Marks, G. (2015). A comparison of methods for analyzing viral load data in studies of HIV patients. *PloS one*, *10*(6).

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592.

Rylance, S., Rylance, J., McHugh, G., Majonga, E., Bandason, T., Mujuru, H., ... & Ferrand, R. A. (2019). Effect of antiretroviral therapy on longitudinal lung function trends in older children and adolescents with HIV-infection. *PloS one*, *14*(3).

Samson, A., Lavielle, M., & Mentré, F. (2006). Extension of the SAEM algorithm to left-censored data in nonlinear mixed-effects model: Application to HIV dynamics model. *Computational Statistics & Data Analysis*, *51*(3), 1562-1574.

Shen, C. W., & Chen, Y. H. (2012). Model selection for generalized estimating equations accommodating dropout missingness. *Biometrics*, *68*(4), 1046-1054.

Simpson, G. L. (2018). Modelling palaeoecological time series using generalised additive models. *Frontiers in Ecology and Evolution*, *6*, 149.

Song, J., Barnhart, H. X., & Lyles, R. H. (2001). A Gee Approach for Estimating the Correlation between Left-Censored Variables. *Mathematics Preprint Archive*, *2001*(11), 163-181.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., ... & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, *338*, b2393.

Sullivan, K. J., Shadish, W. R., & Steiner, P. M. (2015). *An introduction to modeling longitudinal data with generalized additive models: Applications to single-case designs* (Vol. 20, No. 1, p. 26). American Psychological Association.

Sutradhar, B. C. (2018). Semi-parametric Dynamic Models for Longitudinal Ordinal Categorical Data. *Sankhya A*, *80*(1), 80-109.

Thiébaut, R., Jacqmin-Gadda, H., Babiker, A., Commenges, D., & Cascade Collaboration. (2005). Joint modelling of bivariate longitudinal data with informative dropout and left censoring, with application to the evolution of CD4+ cell count and HIV RNA viral load in response to treatment of HIV infection. *Statistics in medicine*, *24*(1), 65-82.

Turner, K. (2011). Introduction to Infectious Disease Modelling. S*exually Transmitted Infections*; **87:**21. http://dx.doi.org/10.1136/sti.2010.046342

UNAIDS, 2017. Ending AIDS: progress towards the 90–90–90 targets [PDF file]. Retrieved from https://www.unaids.org/en/resources/documents/2017/20170720_Global_AIDS_update_2017

UNAIDS, 2019. Global HIV and AIDS statistics [PDF file]. Retrieved from https://www.unaids.org/sites/default/files/media_asset/UNAIDS_FactSheet_en.pdf

Van Montfort, K., Oud, J. H., & Satorra, A. (Eds.). (2010). *Longitudinal research with latent variables*. Springer Science & Business Media.

Vangeneugden, T., Laenen, A., Geys, H., Renard, D., & Molenberghs, G. (2004). Applying linear mixed models to estimate reliability in clinical trial data with repeated measurements. *Controlled clinical trials*, *25*(1), 13-30.

Verbeke, G. (2000). Linear mixed models for longitudinal data. *Springer Series in Statistics*, 30-50.

Verbeke, G., & Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, *91*(433), 217-221.

Wang, C. Y., Huang, Y., Chao, E. C., & Jeffcoat, M. K. (2008). Expected estimating equations for missing data, measurement error, and misclassification, with application to longitudinal nonignorable missing data. *Biometrics*, *64*(1), 85-95.

Wedderburn, R. W. M. (1974). *Quasilikelihood functions, generalized linear models, and the Gauss–Newton method.* Biometrika 61: 439–447.

Wei, X., Ghosh, S. K., Taylor, M. E., Johnson, V. A., Emini, E. A., Deutsch, P., ... & Saag, M. S. (1995). Viral dynamics in human immunodeficiency virus type 1 infection. *Nature*, *373*(6510), 117.

Weiss, R. E. (2005). *Modeling Longitudinal Data*. New York: Springer.

Wolfinger, R. (1993). Laplace's approximation for nonlinear mixed models. *Biometrika*, *80*(4), 791-795.

Wu, H., & Ding, A. A. (1999). Population HIV-1 dynamics in vivo: applicable models and inferential tools for virological data from AIDS clinical trials. *Biometrics, 55*(2), 410-418.

Wu, H., & Zhang, J. T. (2002). The study of long-term HIV dynamics using semi-parametric non-linear mixed-effects models. *Statistics in Medicine, 21*(23), 3655-3675.

Wu, H., Ding, A. A., & De Gruttola, V. (1998). Estimation of HIV dynamic parameters. *Statistics in medicine*, *17*(21), 2463-2485.

Wu, H., Huang, Y., Acosta, E. P., Rosenkranz, S. L., Kuritzkes, D. R., Eron, J. J., ... & Gerber, J. G. (2005). Modeling long-term HIV dynamics and antiretroviral response: effects of drug potency, pharmacokinetics, adherence, and drug resistance. *Journal of Acquired Immune Deficiency Syndromes*, *39*(3), 272-283.

Wu, H., Zhao, C., & Liang, H. (2004). Comparison of linear, nonlinear, and semiparametric mixed-effects models for estimating HIV dynamic parameters. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *46*(2), 233-245.

Wu, L. (2004). Exact and approximate inferences for nonlinear mixed-effects models with missing covariates. *Journal of the American Statistical Association*, *99*(467), 700-709.

Xie, H. (2010). Adjusting for nonignorable missingness when estimating generalized additive models. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *52*(2), 186-200.

Xiong, J., & Yi, G. Y. (2019). swgee: An R Package for Analyzing Longitudinal Data with Response Missingness and Covariate Measurement Error. *R J.*, *11*(1), 416.

Xu, S., & Blozis, S. A. (2011). Sensitivity analysis of mixed models for incomplete longitudinal data. *Journal of Educational and Behavioral Statistics*, *36*(2), 237-256.

Yang, L., Qin, G., Zhao, N., Wang, C., & Song, G. (2012). Using a generalized additive model with autoregressive terms to study the effects of daily temperature on mortality. *BMC medical research methodology*, *12*(1), 165.

Yi, G. Y. (2008). A simulation-based marginal method for longitudinal data with dropout and mismeasured covariates. *Biostatistics*, *9*(3), 501-512.

Yi, G. Y., Liu, W., & Wu, L. (2011). Simultaneous inference and bias analysis for longitudinal data with covariate measurement error and missing responses. *Biometrics*, *67*(1), 67-75.

Yi, G. Y., Ma, Y., & Carroll, R. J. (2012). A functional generalized method of moments approach for longitudinal studies with missing responses and covariate measurement error. *Biometrika*, *99*(1), 151-165.

Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological methods*, *14*(4), 301–322. https://doi.org/10.1037/a0016972.

Zhang, J. T., & Wu, H. (2010). Modeling HIV dynamics using unified mixed-effects models. *American Journal of Mathematical and Management Sciences*, *30*(1-2), 83-109.

Zhang, Z., Reinikainen, J., Adeleke, K. A., Pieterse, M. E., & Groothuis-Oudshoorn, C. (2018). Time-varying covariates and coefficients in Cox regression models. *Annals of translational medicine*, *6*(7), 121. doi:10.21037/atm.2018.02.12

Appendix

Appendix A: R codes using LME4 package

```
library("nlme")

library(lme4)

SampleData<-read.csv(file.choose(), head=T) # Load CSV dataset

#==========CREATE DATASET============================

ViralLoad<-SampleData$Viral_Load

Time<-SampleData$Time_in_months

Gender<-as.factor(SampleData$Gender)

Weight_at_Followup<-SampleData$Weight_at_Followup

Weight<-SampleData$Baseline_Weight

Adherence<-as.factor(SampleData$Adherence_ART)

Age<-SampleData$Age

WHO.Stage<-as.factor(SampleData$WHO_STAGE)

CD4<-as.factor(SampleData$CD4_GROUP)

Id<-SampleData$Patient.ID

O<-SampleData$O

#======================DATASET=================================
=

Dat<-data.frame(ViralLoad,Id, Time,Gender, Weight, Adherence,CD4, Age,
Weight_at_Followup, WHO.Stage, O)

#Dat <- within (Dat, Adherence <- relevel (Adherence, ref = 99))

#======================FIXED EFFECTS MODEL  ======================

 Model11<-lm(log(ViralLoad)~Time +Gender+ CD4 + Adherence*Time + Age  + Weight +
Weight_at_Followup*Time +WHO.Stage, data=Dat, na.action=na.omit)

  summary(Model11)

  logLik(Model11)
```

```
AIC(Model11)

#=======================MODEL ONE (LME)=========================

Model55<-lme(log(ViralLoad)~Time + CD4 +Gender+ Adherence*Time + Age + Weight_at_Followup*Time + WHO.Stage,random=~ 1 + Time | Id, data=Dat, na.action=na.omit)

summary(Model55)

anova(Model22 )
```

Appendix B: R codes using geepack package

```
set.seed(1234)
#==========Required packages=========================================
library("geepack")
library(MASS)
library("swgee")
library("gee")
library(lme4)
SampleData<-read.csv(file.choose(), head=T) #Load CSV dataset
#======================CREATEDATASET=========================
ViralLoad<-SampleData$Viral_Load
Time<-SampleData$Time_in_months
Gender<-as.factor(SampleData$Sex)
Weight_at_Followup<-SampleData$Weight_at_Followup
Weight<-SampleData$Baseline_Weight
Adherence<-as.factor(SampleData$Adherence_ART)
Age<-SampleData$Age
WHO.Stage<-as.factor(SampleData$WHO_STAGE)
CD4<-as.factor(SampleData$CD4_GROUP)
Id<-SampleData$Patient.ID
O<-SampleData$O
#======================DATASET=================================
Dat<-data.frame(ViralLoad, Id, Time, Gender, Weight, Adherence, CD4, Age, Weight_at_Followup, WHO.Stage, O)
#=======GEE with Unstructured correlation=============================
Model1<-gee(log(ViralLoad)~Time +Gender+ Adherence*Time + Age + CD4 + Weight + Weight_at_Followup*Time + WHO.Stage, id=Id,data=Dat, family=gaussian(link="identity"), corstr="unstructured")
```

```
summary(Model1)

PValue1<-2*pnorm(abs(coef(summary(Model1))[,5]), lower.tail = FALSE)

 PValue1

#============GEE with Independence correlation=====================

Model2<-gee(log(ViralLoad)~Time + Gender + Adherence*Time + Age + CD4 + Weight + Weight_at_Followup*Time + WHO.Stage, id=Id,data=Dat, family=gaussian(link="identity"), corstr="independence")

summary(Model2)

PValue2<-2*pnorm(abs(coef(summary(Model2))[,5]),lower.tail = FALSE)

PValue2

#====================Quasi Information Criterion=====================

QIC.long.gee <- function(model.R,model.independence)

 {

  #calculates Gaussian Quasi-Likelihood

  AIinverse <- solve(model.independence$naive.variance)

  V.msR <- model.R$robust.variance

  trace.term <- sum(diag(AIinverse%*%V.msR))

  #estimated mean and observed values

  mu.R <- model.R$fitted.values

  y <- model.R$y

  #quasilikelihood for Gaussian model

  Quasi.R = sum(((y - mu.R)^2)/-2)

  QIC <- (-2)*Quasi.R + 2*trace.term

  output <- c(QIC,Quasi.R, trace.term)

  names(output) <- c('QIC','Quasi Lik','Trace')

  output

 }

 sapply(list(Model1, Model2), function(x) QIC.long.gee(x,Model1))
```

Appendix C: R codes using swgee package

```
library("swgee")

SampleData<-read.csv(file.choose(), head=T) #Load csv dataset

set.seed(1000)

sigma <- diag(rep(0.25, 2))

set.seed(1000)

sigma <- diag(rep(0.25, 2))

output3 <- swgee(log(Viral_Load)~Time_in_months + as.factor(CD4_GROUP) + as.factor(Gender)+Baseline_Weight+Adherence_ART*Time_in_months + Age + as.factor(WHO_STAGE)+Weight_at_Followup*Time_in_months,data=SampleData,id=Patient.ID,family=gaussian(link="identity"),corstr="unstractured",missingmodel=O~log(Viral_Load)+Adherence_ART+Weight_at_Followup,SIMEXvariable=c("Weight_at_Followup","Adherence_ART"), SIMEX.err=sigma, repeated=FALSE, B=50, lambda=seq(0, 2, 0.5))

summary(output3)
```