University of Nairobi

School of Mathematics

ISSN: 2410-1397

Master Project in Statistics

Data Driven Longitudinal Model with Application to HIV Differentiated Care Data

Research Report in Statistics, Number 13, 2019

Weunda O. Stephen

June 2019

Submitted to the School of Mathematics in partial fulfilment for a degree in Master of Science in Statistics

Data Driven Longitudinal Model with Application to HIV Differentiated Care Data

Research Report in Statistics, Number 13, 2019

Weunda O. Stephen

School of Mathematics
College of Biological and Physical sciences
Chiromo, off Riverside Drive
30197-00100 Nairobi, Kenya

Master Thesis

Submitted to the School of Mathematics in partial fulfilment for a degree in Master of Science in Statistics

Submitted to: The Graduate School, University of Nairobi, Kenya

Abstract

Background: Differentiated care is a new innovative approach for managing HIV/AIDS where ART treatment services are customized by staggering patients' visits for stable status while reducing unnecessary burdens on the health system. Through provision of differentiated care, the health system can reallocate resources to patients most in need who are failing treatment.
Objective: The main objective of this study is to develop a data-driven longitudinal model which is applicable to HIV differentiated care.
Method: We used routine data of HIV positive patients initiated to ART at the point of care from 4 medical facilities in Nairobi in the year 2018. Since both the GLMM and GEE are extensions of the GLM, we start with a brief overview of GEE then relooked at extensions of GLMM. We specify $f(\mu)$ and $g(\mu)$ to be dependent on the type of response Y_i. For a binary Y_i , we consider $f(\mu)$ as Bernoulli distribution and $g(\mu)$ as the logit function, $g(\mu)$ = $\log[\mu/1-\mu]$ resulting to GLM is the logistic regression.
Results and conclusion: Results show the binary response which was differentiated care category fits well with GLMM. We also found TB-HIV co-infection to be the only significant predictor of differentiated care under both GEE and GLMM.

Master Thesis in Statistics at the University of Nairobi, Kenya.
ISSN 2410-1397: Research Report in Statistics

DISTRIBUTOR: School of Mathematics, University of Nairobi, Kenya

Declaration and Approval

I the undersigned declare that this dissertation is my original work and to the best of my knowledge, it has not been submitted in support of an award of a degree in any other university or institution of learning.

Signature Date

Weunda O. Stephen
Reg No. I56/8041/2017

In my capacity as a supervisor of the candidate's dissertation, I certify that this dissertation has my approval for submission.

Signature Date

Dr. Idah Orowe
School of Mathematics,
University of Nairobi,
Box 30197, 00100 Nairobi, Kenya.
E-mail: orowe@uonbi.ac.ke

Dedication

This research is dedicated to my family; my wonderful parents Francisca and Nelson, and my siblings. Thank you for walking through the arduous journey with me. And to the Almighty God, for granting me the ability to be able to undertake the research. I am grateful.

Contents

List of Figures

List of Tables

Acknowledgments

First and foremost I wish to thank God as all is and has been possible through Him. I would like to sincerely thank Dr. Idah Orowe, whose guidance, patience, words of encouragement, insight and wisdom helped me to complete the project on time. I would also like to sincerely thank Dr. Collins Odhiambo who also helped frame the ideas for the project and facilitated all the inputs for data collection. I would also like to appreciate Mr. Evans Otieno Omondi who provided guidance and insight on how to work with the statistical softwares and formatting of the document. I would also like to acknowledge my parents Francisca and Nelson, and my brothers and sisters for the prayers, direct support and words of encouragement throughout the project. I would like to acknowledge the MSc. Mathematical Statistics class of 2019 for making light of the hard situations and making them manageable.

Weunda O. Stephen

Nairobi, 2019.

1 Introduction

1.1 Background to the Study

There is considerable literature on finding suitable statistical models for HIV/AIDS data so as to be able to administer the right antiretroviral therapy(ART) to the patients. The most widely used data are the CD4 counts and the viral loads. More recently, attempts to model differentiated care have become important.

Differentiated care is a new approach of managing HIV/AIDS whereby ART services are customized to capture the preferences and expectations of HIV positive patients while reducing unnecessary burdens on the health system. Through provision of differentiated care, the health system can reallocate resources to patients most in need. Differentiated care is meant for the HIV positive patients who are stable and therefore they do not need frequent clinical consultations or frequent medical attention. Differentiated care is characterized by: taking the services to the doorsteps of the clients, less frequent visits to clinics, getting drugs out of facilities, communication with health-care providers through some sort of technology without visiting the facility, creating support networks etc (Grimsrud et al., 2017; Organization et al., 2016).

Some markers of HIV/AIDS such as CD4+ counts and viral loads among others are longitudinally measured. In longitudinal studies, there are variables which are repeatedly measured over time, and these variables may be used either as responses or predictors, depending on the objectives of the study . A common challenge of longitudinal studies is that data on some of these variables may be missing at times of interest, or may be measured with errors. Other challenges include correlation of the data which violates the assumption of independence of observations.

Therefore longitudinal data should be modeled using appropriate statistical models for correct statistical inference. The dominant longitudinal models include the generalized linear mixed models (GLMM) and generalized estimating equations (GEE) which are capable of analyzing correlated, non-normal or missing data (Gamerman and Lopes, 1997; Lee and Nelder, 2001; Liang and Zeger, 1986; Magder and Zeger, 1996).

The response variable in this study is differentiated care. This research attempts to develop a statistical model that will be used to predict whether or not a patient should be placed under differentiated care. The relationship between differentiated care and other HIV

variables such as CD4+ counts, viral loads, TB-HIV co-infection, age among others will also be brought to light.

Chapter two of this research discusses some of the existing studies on suitable models for HIV data as well as the progression of HIV/AIDS. Chapter three discusses the application of the longitudinal models to HIV/AIDS data. Chapter four is on results and discussions after analysis of longitudinal data. Finally, conclusions on this study will be discussed in chapter five.

1.2 Problem Statement

Currently in Kenya, all HIV patients are advised to visit the hospital regularly without failing for ART services. This is a one way fits all approach. This approach is so far not effective in combating HIV/AIDS due to three major reasons: First, most patients, as dictated by their lifestyles are not committed to faithfully keeping these appointments and thus with time end up not responding to the prescribed treatment. Such lifestyles include busy work schedules, drug abuse etc.

The second major factor is the limited resources available to conduct these activities within the hospitals. There are few understaffed hospitals. Finally, stigmatization. Some HIV positive patients may be uncomfortable to have their status publicized for fear of discrimination. Ultimately these challenges lead to unresponsiveness to the ART and thus puts the patients at a graver risk (Institut National de Sante' Publique du Quebec, 2014).

In line with the above problems, there are HIV/AIDS prediction systems that are currently existing to manage HIV/AIDS. These prediction systems are so far not effective because they only determine the levels and kinds of ARVs to be administered and they support the one way fits all approach. This leaves the health facilities burdened and the unstable patients may not receive adequate treatment. The expectations and preferences of the stable HIV patients may also be unrealized. In addition, Some of the existing prediction models used regression techniques that cannot support non-normal, missing or correlated data therefore they produced biased statistical inferences (Culshaw, 2006; Degruttola et al., 1991). They did not use the most flexible and powerful longitudinal models to handle non-normal data like GLMM and GEE.

All the studies that have been done so far did not model the relationship between differentiated care and other longitudinally measured HIV/AIDS data. The challenges above call for the implementation of differentiated service delivery. This study attempts to develop a model that will be used to predict whether or not a patient should be placed under differentiated care by using the most flexible and powerful longitudinal models to handle non-normal, missing or correlated data; the GEE and the GLMM.

1.3 Objectives

1.3.1 General Objectives

The general aim of this study is to develop a data-driven longitudinal model with application to HIV differentiated care data.

1.3.2 Specific Objectives

1. To model the association between longitudinally measured HIV variables and differentiated care using GLMM and GEE and to compare the two models.
2. To determine which statistical model is better using goodness of fit.
3. To identify factors which predict differentiated care.

1.4 Justification of the Study

The significance of this study is that; we will be able to accurately derive a longitudinal model to predict whether or not a patient should be initiated in differentiated care. With differentiated ART delivery, patients are able to return to their normal daily routines and not waste time queuing at the clinics. The research will also enable the government to accurately allocate medical resources. With differentiated ART delivery, doctors and nurses will be able to provide quality care without feeling rushed especially to the unstable patients. With differentiated ART delivery HIV/AIDS patients will not be exposed to stigmatization. All these benefits will allow the various people living with HIV to obtain proper care that reflects their preferences and expectations (Organization et al., 2016; Phillips et al., 2015).

2 Literature Review

2.1 The Human Immunodeficiency Virus (HIV)

The Human Immunodeficiency Virus (HIV) leads to Acquired Immuno-Deficiency Syndrome (AIDS), a condition in which the immune system begins to decline, exposing infected individuals to life-threatening opportunistic infections (Callaway et al., 1999). The HIV virus is a retrovirus that attacks the human Cluster of Differential 4 (CD4+) cells, leading to a decline in their natural defenses against pathogenic microorganisms (Rosa et al., 2014). A cure or vaccine for HIV/AIDS does not currently exist. However, great strides have been made in treatment termed as Highly Active Anti-Retroviral Therapy (HAART). HAART is made up of cocktails of at least two to three different classes of antiretroviral therapies and effectively lowers the concentration of the virus in the body by increasing the immune system which is called CD4+ T cells and suppressing the viral loads. In most developed countries, where these drugs are available, a large reduction in HIV-associated morbidity and mortality has been registered.

Archer (2008) states that there are two major phenotypes of HIV virus, namely HIV-1 and HIV-2. HIV -1, which this study will focus on has three strains; labelled as M (Major), O (Outlier) and N (New i.e. not M or O). He also states that the strain that is almost entirely to blame for the global pandemic is the Group M, which has a lot of diversity. HIV-2 is relatively uncommon and is concentrated majorly in the West of Africa. This phenotype is less infectious and progresses slower as compared to HIV -1. HIV in this study, will refer to the more common HIV -1 phenotype.

The majority of the infected people within Kenya are aged between 15-39 years. The prevalence rate of HIV amongst this group is 5.9%. The main contributor to the high incidence of HIV in Kenya has been attributed to the high level of poverty. Organization et al. (2016). Consequentially, a lot of research has gone into trying to come up with a solution to HIV/AIDS with the recent temporal solution being the invention of the Antiretroviral Therapy (ART) drugs, composed of a compound of medicines aimed at slowing down the rate at which the HIV virus replicates itself. However, a bigger quartile of the population of the third world countries is still suffering from logistical challenges such as lack of adequate medical equipment and medical supplies in the hospitals, and the high prices of undertaking the activities and tests. Worst case scenarios have included the introduction of ART in the late stages of a HIV patients (Lopez, 2011).

In the recent decades the use of data relating to HIV protein levels in the plasma, also known as clinical markers have been used to predict progression in HIV-1 infection. The most common clinical markers are the CD4+ counts and the viral loads. The CD4+ count is a measure of the number of white blood cells per milliliter of blood that contain the CD4 glycoproteins. The CD4+ cells are usually developed in response to infections Tunduny (2017). Viral load on the other end, is a measure of the actual number of viral particles per milliliter of blood. This count is more accurate than the CD4+ count since CD4+ cells are usually detected after the drug resistance has been developed, and can also be affected by other factors other than HIV infection, such as other infection (Tunduny, 2017).

2.2 Longitudinal Models for HIV/AIDS Data

2.2.1 Generalized Linear Mixed Models (GLMM)

GLMM are extensions of the generalized linear models(GLM). Breslow and Clayton (1993), McCullagh (81). They are obtained by adding random effects into the GLM. GLMM are useful for modeling the correlation among response variables inherent in longitudinal or repeated measures studies, for accommodating over dispersion among binomial or Poisson responses, and for producing shrinkage estimators in multi-parameter problems. One way to account for the within subject correlation is through the introduction of random effects in generalized linear models which leads to a class of models known as GLMM. GLMM are parametric. They are estimated using maximum likelihood estimation method. GLMM have a wide range of applications which has made them receive significant attention. They are now available in the major software packages such as R and STATA.

GLMM have a computational burden due to the high dimensional numerical integration which has limited past studies of GLMM to the case of simplified models (e.g., random intercept models); to tractable random effects distributions; or to conditional inference for the regression coefficients, conditioning on the random effects (Zeger and Karim, 1991). A variety of ways have been suggested to overcome the computational difficulties so as to improve inference and estimation procedures for the fixed effects in GLMM. They include Gibbs sampling Zeger and Karim (1991), penalized quasi-likelihood and marginal quasi-likelihood Breslow and Clayton (1993), pseudo-likelihood based on approximate marginal models Wolfinger and O'connell (1993), and maximum likelihood with Monte Carlo versions of EM, Newton-Raphson and simulated maximum likelihood algorithms (McCulloch, 1997). These methods require normal distribution assumptions for the random effects. Methods for non-normal random effects are less common and limited to specialized cases (Gamerman and Lopes, 1997; Lee and Nelder, 2001; Magder and Zeger, 1996).

2.2.2 Generalized Estimating Equations (GEE)

GEE advanced by Liang and Zeger (1986) is a class of estimating equations which take into account the correlation arising due to a repeated study design so as to increase the efficiency of standard error estimates. GEE are based on quasi- likelihood theory and can be used for continuous as well as for discrete outcome. Nelder and Wedderburn (1972). The GEE is also an extension of the GLM. One of the advantages of using the GEE is that the solutions are consistent, i.e. the estimate of parameters are nearly efficient and asymptotically Gaussian, even when the time dependence is mis-specified. In a GEE, the parameter estimates are estimated parametrically and the variances are estimated non-parametrically. Hence the GEE is semi-parametric. GEE deals with the within subject correlation caused by the collection of several samples from each subject by adjusting the standard error to compensate for the absence of independence among samples. GEE focus on estimating the average response over the population ("population-averaged effects) rather than the regression parameters that would enable prediction of the effect of changing one or more components of X on a given individual.

2.3 HIV Differentiated Care

Grimsrud et al. (2017); Organization et al. (2016) introduced HIV Differentiated Care. Differentiated care is a client-centred approach that simplifies and adapts HIV services across the cascade to reflect the preferences and expectations of various groups of people living with HIV (PLHIV) while reducing unnecessary burdens on the health system. By providing differentiated care, also called differentiated service delivery, the health system can reallocate resources to those most in need. Differentiated care can be organized based on the specific needs of groups of patients, such as clinical characteristics of patients (e.g. patients with advanced disease), sub-populations (e.g. pregnant and breastfeeding women, adolescents, children, key populations), or context (e.g. low-prevalence vs. high-prevalence settings).

Differentiated care applies across the HIV continuum to all three of the 90-90-90 targets (90% of people living with HIV should know their status; 90% who know their status should be on ART; 90% of those on ART should be virologically suppressed). The aim of differentiated care is to enhance the quality of the client experience. The main driver to adapting service provision is the patient's needs.

Who is a Stable Client?

Organization et al. (2016) stable clients are those PLHIV on ART who are adherent and do not require frequent clinical consultation. They have received ART for at least one year and have no adverse drug reactions that require regular monitoring, no current illnesses such as TB or pregnancy, are not currently breastfeeding, have good understanding of

lifelong adherence and evidence of treatment success (i.e. two consecutive viral load measurements below 1000 copies/mL). In the absence of viral load monitoring, rising CD4 cell counts i.e. CD4 counts > 200 cells/mm3, an objective adherence measure, can be used to indicate treatment success." It is dependent on access to resources, such as routine viral load monitoring.

2.4 Review of Previous Studies

A number of researches have been carried out on finding suitable models for clinical markers of HIV as well as the progression of HIV. The most widely used markers are the Cluster of Differential 4 (CD4+) cells and measures of HIV viral loads. A review of these previous studies will be discussed in this chapter.

Wu and Zhang (2018) studied mixed effects models with censored covariates, with applications in HIV/AIDS Studies. This study brought to light the fact that these models are capable of handling censoring and missing data in covariates, since the "predicted values" based on these models are more reliable than the commonly used empirical covariate models. However, they reported the following limitations: (i) in many applications such models may not be available and (ii) computational difficulties. This is because these models are often non-linear, hence computation is a main challenge in likelihood inference.

In 2018, Yu and Wu (2018) did a research on modeling the relationship between CD4 counts and Viral Loads for Complex HIV/AIDS data. The study established that Viral load and CD4 play a central role in HIV/AIDS studies. The study reported that much of the research in the literature failed to address measurement errors, outliers and missing data, which are typical features of AIDS datasets. To their knowledge, their research was the first attempt to address all these data complications simultaneously. The results in the paper confirmed that CD4 and viral load are negatively associated over time, whether CD4 is viewed as continuous, binary, count or CD4 is viewed as response or covariate. However, the level of strength of the association can be severely mis-estimated if measurement errors and outliers are not addressed in data analysis. Simulation results confirmed that the proposed robust two-step methods for joint LME/GLMM and joint NLME/LME models performed well. The proposed methods can be applied to missing data by incorporating a missing data model (e.g. a binary mixed effects model for the missing data indicators), which also leads to a joint model.

In 2014, Lu (2014) suggested that statistical analyses and modeling have contributed greatly to the understanding of the progression of HIV-1 infection; they also provide guidance for the treatment of AIDS patients and evaluation of ART. Various statistical models, nonlinear mixed-effects models in particular, were used to model the CD4 counts and the viral loads. A common assumption in these methods is all HIV patients come from a homogeneous population following one mean trajectories. This assumption obscures

important characteristic difference between subgroups of patients whose response to treatment and whose disease trajectories are biologically different. This may lead to biased inference. In their research, they developed mixture dynamic model and related Bayesian inferences via Markov chain Monte Carlo (MCMC). Finite mixture models, also known as latent class models, are commonly used to model non-predetermined heterogeneity in a population; they provide an empirical representation of heterogeneity by grouping the population into a finite number of latent classes and modeling the population through a mixture distribution. This important feature may help physicians to better understand a particular patient disease progression and refine the ART strategy in advance.

In 2012, Taye (2012) evaluated the association between the progressions of HIV infection using longitudinally measured CD4 count and its possible predictors via longitudinal analysis methodologies. Statistically two modeling approaches (GEE and GLMM) were compared for the analysis of ART data and found that GLMM exhibited the best fit for these data with small disturbance than GEE. The study also found that on average CD4 count increases in a quadratic pattern over time after patients initiated to ART program (i.e. the immune system increases whereas the progression of the disease goes down due to the therapy). Furthermore, though the choice between GEE and GLMM for longitudinal data can only be made on subject matter grounds, using generalized linear mixed model is much emphasized than generalized estimating equations for correlated data as GEE can only handle the within subject variations through the assigned working correlation structure where as GLMM in addition to within measurement variation, between individual variations can be accounted by incorporating the random effects. Due to that, GLMM fits a given data with a small disturbance than GEE.

Boscardin et al. (1998) did a research on longitudinal models for AIDS marker data. The study reviewed the existing literature including the preferred models which involved mixed effects, stochastic terms and independent measurement error. Adding stochastic terms to standard mixed effects models gives an interpretable and parsimonious method for generalizing the covariance structure of the measurement error and short-term variability.

Hughes et al. (1994), did a research on within-subject variation in CD4 counts in HIV Infection: Implications for patient monitoring. The study found out that Changes in CD4 counts are widely used in monitoring HIV-infected patients for disease progression. They noted that, random fluctuations may obscure clinically significant changes. They assessed for up to 2 years CD4 cell counts from 1020 untreated subjects with HIV infection who were monitored by standardized methods. The within-subject coefficient of variation averaged 25% but was higher in subjects with lower CD4 counts. They reported that using multiple counts and other markers may provide more precise assessment of immune status.

Degruttola et al. (1991) did a research on modeling the progression of HIV Infection.The study found out that statistical modeling of the progression of markers of HIV infection is complicated by three problems: (1) censored information; (2) missing data; and (3) correlation within and between subjects. The studied the CD4 counts. As a result of the three problems, it is difficult to distinguish between different models for the decline in CD4 count over time for HIV-infected individuals. They concluded that models that assume a steady linear decline of CD4 counts on the square root scale and accommodate the three sources of variation mentioned previously provide adequate fits to the study data. They also noted that the linear decline does not apply near the time of seroconversion; this event seems to be accompanied by a sharp drop in CD4 counts.

The previous studies did not model differentiated care using GLMM and GEE. This research attempts to model differentiated care by applying the robust longitudinal models; the GLMM and the GEE which are capable of handling challenges in longitudinal data sets.

3 Methodology

3.1 Research Design

3.1.1 Data Source and its Description

The study used routine data of HIV positive patients initiated to ART at the point of care from 4 medical facilities in Nairobi in the year 2018. The medical record is a software called IQ Care. The data was pulled from IQ Care then stored in excel. The study population is made up of HIV positive patients between 10 and 58 years of age whose viral loads and CD4+ T cells were measured at least once. The data consists of 193 individuals with a minimum of one and a maximum of three measurements per individual. The data was recorded in medical charts by assigning an identification number per individual which helps to find the patients profile easily during his/her next visit time. The data in excel was cleaned and coded in R software version 3.6.0. All analysis was done in R software save for the descriptive statistics in table 4.1 which was done in SPSS version 25. After obtaining Ethical Clearance at Strathmore University Institutional Ethics Review Committee (SU-IERC), the data was collected via extracting the required variables from medical charts in a check list format.

3.1.2 Study Variables

Dependent variable/response variable: Differentiated care
Independent variables/predictors: Age when last visited clinic, Sex, TB-HIV Co-infection, Viral loads and CD4+ T- cells (CD4 count) for each individual measured in every visitation.

3.1.3 Exploratory Data Analysis

It is a technique to visualize the patterns of data relative to research interests. Since exploratory data analysis can serve to discover as much of the information regarding raw data as possible, plotting individual curves to carefully examine the data should be performed first before any formal model fitting is carried out.

3.2 The Statistical Models for Longitudinal Data

Since both the GLMM and GEE are extensions of the GLM, we start with a brief overview of the latter (Lin et al., 2016)

Table 3.1. Variable description and coding for ART data

Variable	Coding
Age	Age when last visited clinic
Gender	Male = 1, Female = 2
TB-HIV Co-infection	Yes = 1 No = 2
Viral loads suppression category	< 1000 copies/ml = 1 > 1000 copies/ml = 2
CD4 counts category	0-200 = 1 200-400 = 2 400-600 = 3 600+ = 4
Differentiated care	Yes = 1 No = 0

3.2.1 Generalized Linear Models (GLM)

McCullagh and Nelder (1989) introduced the concept of Generalized Linear Models (GLM). Consider a sample of n subjects and let $Y_i(x_i)$ denote a continuous response. The classic linear model is given by:

$$Y_i = X_i^T\beta + \varepsilon_i, \varepsilon_i \sim N(0,\sigma^2), 1 \leq i \leq n, \tag{1}$$

where N$(0,\sigma^2)$ denotes a normal distribution with mean μ and σ^2 variance. One major limitation is that it only applies to continuous response Y_i . The generalized linear models (GLM) extend the classic linear model to non-continuous response such as binary. To express the GLM, we first rewrite the linear regression in (3.1) as:

$$Y_i|X_i \sim N(\mu_i,\sigma^2), \mu_i = E(Y_i|X_i) = X_i^T\beta, 1 \leq i \leq n, \tag{2}$$

where $Y_i|X_i$ denotes the conditional distribution of Y_i given X_i and $E(Y_i|X_i)$ denotes the conditional mean of Y_i given X_i. By replacing the normal in (3.2) with other distributions appropriate for the type of response, we obtain the class of GLM:

$$Y_i|X_i \sim f(\mu_i), g(\mu_i) = X_i^T\beta, 1 \leq i \leq n, \tag{3}$$

where $f(\mu)$ denotes some distribution with mean μ and $g(\mu)$ is a function of μ. Since $g(\mu)$ links the mean to the explanatory variables, $g(\mu)$ is called the link function. μ , the pdf of Y, is the random component of the GLM. $X_i^T\beta$ is the systematic component.

The specification of $f(\mu)$ and $g(\mu)$ depends on the type of response Y_i. For a binary Y_i, $f(\mu)$ is the Bernoulli distribution and $g(\mu)$ is often set as the logit function, $g(\mu)$ = $\log[\frac{\mu}{1-\mu}]$. The resulting GLM is the logistic regression.

Inference for GLM can be based on maximum likelihood (ML) or estimating equations (EE). The classic ML provides most efficient estimates, if the response Y_i follows the specified distribution such as the normal in the linear regression in (3.1). In many studies, it may be difficult to specify the right distribution, in which case the ML will yield biased estimates if the specified distribution does not match the data distribution. The modern alternative EE uses an approach for inference that does not require specification of a mathematical distribution for Y_i, thereby providing valid inference for a wider class of data distribution. Since no distribution is required under EE, we may also express the GLM in this case as:

$$\mu_i = E(Y_i|X_i), g(\mu_i) = X_i^T\beta, 1 \leq i \leq n, \tag{4}$$

or simply

$$g(E(Y_i|X_i)) = X_i^T\beta, 1 \leq i \leq n, \tag{5}$$

or equivalently

$$E(Y_i|X_i) = h(X_i^T\beta), 1 \leq i \leq n, \tag{6}$$

where $h = g^{-1}$ is the inverse of $g(\mu)$. When specified without the distribution component, (3.4), (3.5) or (3.6) are also called the semi-parametric GLM. In comparison, (3.3) is called the parametric GLM.

3.2.2 Extension of GLM to Longitudinal Data

Generalized Estimating Equations (GEE)

Let T = time points in a longitudinal study and let Y_{it} and X_{it} denote the same response and predictors, but with t indicating their dependence on the time of assessment $(1 \leq i \leq n, 1 \leq t \leq T)$. For each time t , we can apply the GLM in (3.3) to model the regression relationship between Y_{it} and X_{it} at each point:

$$Y_{it}|X_{it} \sim f(\mu_{it}), \mu_{it} = E(Y_{it}|X_{it}), g(\mu_{it}) = X_{it}^T\beta_t, 1 \leq t \leq T, 1 \leq i \leq n$$

We can then get estimates of β_t for each time point t. However, it is difficult to interpret different β_t across the different time points. Moreover, it is technically challenging to combine estimates of β_t to test hypotheses concerning temporal trends because of inter-dependence between such estimates. The GEE addresses the aforementioned difficulties by using a single estimate β to model changes over time based on multiple assessment times. Since the GEE estimates β using a set of equations that do not rely on assumed distribution $f(\mu_{it})$ in (3.7), the resulting model becomes:

$$\mu_{it} = E(Y_{it}|X_{it}), g(\mu_{it}) = X_{it}^T\beta, 1 \leq t \leq T, 1 \leq i \leq n \quad \textbf{(7)}$$

By comparing the GEE above with the model in (3.4), it is seen that the GEE is an extension of the semi-parametric GLM to longitudinal data. The key difference between (3.4) and (3.7) is that (3.7) is not simply an application of GLM to each of the time points, but rather an extension of the model in (3.4) to provide a single parameter vector β for easy interpretation and estimate this parameter vector by using data from all time points and accounting for correlations between the repeated assessments. Like the semi-parametric GLM, the GEE provides robust statistical inference for a wider class of data distributions.

Suppose that Y_{it} is binary and is modeled by a GEE for binary response as follows:

$$logit(E(Y_{it}|X_{ti})) = X_{it}^T\beta, 1 \leq i \leq n, 1 \leq t \leq T \quad \textbf{(8)}$$

or

$$E(Y_{it}|X_{ti}) = \frac{\exp(X_{it}^T\beta)}{1+\exp(X_{it}^T\beta)}, 1 \leq i \leq n, 1 \leq t \leq T \quad \textbf{(9)}$$

Method of Estimation and Statistical Inference for GEE:

A Quasi-likelihood method of estimation is used since likelihood based methods are not available for testing fit, comparing models and conducting inference about parameters. Inference can only use Wald statistics constructed with asymptotic normality of the estimators together with their estimated covariance matrix. Moreover, even though GEE estimates are consistent with misspecification of the covariance structure, it is important to choose the covariance structure that closely approximates the true underlying one for greater efficiency.

Working Correlation Structures: Because the repeated observations within one subject are not independent from each other, a correction must be made for these within-subject correlations. With GEE, this correction is done by assuming in advance a certain 'working'

correlation structure for the repeated measurements of the outcome variable Y. Before carrying out a GEE analysis, the within-subject correlation structure was chosen based on the results of exploring correlation structure of the observed data. Accordingly, two proposed working correlations were compared.

1. Independence structure: This is the correlation that GEE model assumes by default. It assumes that the correlations between subsequent measurements are assumed to be zero or measurements are independent to each other within individuals.
2. Exchangeable correlation structure (compound symmetry): it assumes the correlations between subsequent measurements are assumed to be the same, irrespective of the length of the time interval.

Generally, assuming no missing data, the $J \times J$ covariance matrix for Y is modeled as:

$$V_i = \phi A_i^{1/2} R_i A_i^{1/2} \tag{10}$$

Where ϕ is a glm dispersion parameter which is assumed 1 for count data, A_i is a diagonal matrix of variance functions, and R_i is the working correlation matrix of Y. GEE can be used to model correlated data with the variance covariance matrix V by iteratively solving the quasi- score equations. The score function of a GEE for β has the form:

$$\sum_{x=1}^{N} (\frac{\partial \mu_i}{\partial \beta_t}) V_i^{-1} (Y_i - \mu_i) = 0 \tag{11}$$

Where μ_i is the fitted mean, which is given by $g(\mu_{it})$ = $X_{it}\beta$ for covariates $X = X_{i1}, X_{i2}, \cdots, X_{in}$ and regression parameters $\beta = \beta_1, \beta_2, \beta_3, \cdots, \beta_p$.
Starting R_i as the identity matrix and $\phi = 1$, the parameters β are estimated by solving equations as follows:

i.e in a normal case
$\mu_i = X_i\beta$ and $\frac{\partial \mu_i}{\partial \beta_t} = X_i, V_i = \hat{\phi} R_i$

$$\sum_{i=1}^{N} (X_i^T) R_i^{-1} (Y_i - \mu_i) = 0$$

$$\hat{\beta} = [\sum_{i=1}^{N} (X_i^T) R_i^{-1} X_i]^{-1} [\sum_{i=1}^{N} (X_{i'}^T) R_i^{-1} Y_i]$$

$$\beta \hat{t+1} = \hat{\beta} t - [\sum_{x=1}^{N} (\frac{\partial \mu_i}{\partial \beta_t})^T V_i^{-1} (\frac{\partial \mu_i}{\partial \beta_t})]^{-1} \sum_{x=1}^{N} (\frac{\partial \mu_i}{\partial \beta_t})^T V_i^{-1} (Y_i - \mu_i)$$

let $D_i = \frac{\partial \mu_i}{\partial \beta_t} and S_i = (Y_i - \hat{\mu_i})(Y_i - \hat{\mu_i})^T$ Then:

1. $\beta^{t+1} = \beta^t - [\sum_{x=1}^{N} (D_i)^T \hat{V_i^{-1}} (D_i)]^{-1} \sum_{x=1}^{N} (D_i)^T \hat{V_i^{-1}} (D_i)$
2. $var(\hat{\beta}) = N[\sum_{x=1}^{N} (D_i)^T \hat{V_i^{-1}} (D_i)]^{-1} \sum_{x=1}^{N} (D_i)^T \hat{V_i^{-1}} \hat{S_i} \hat{V_i^{-1}} (D_i) [\sum_{x=1}^{N} (D_i)^T \hat{V_i^{-1}} (D_i)]^{-1}$

More generally, because the solution only depends on the mean and variance of Y, these are quasi-likelihood estimates. The estimates from a GEE analysis are robust to miss-specification of the covariance matrix (Liang and Zeger, 1986), so, the regression parameter estimates are consistent even for independent covariance matrix. Upon convergence, in order to perform hypothesis tests and construct confidence intervals, it is of interest to obtain standard errors associated with the estimated regression coefficients. These standard errors are obtained as the square root of the diagonal elements of the matrix $V(\hat{\beta})$. The GEE provides two versions of these estimates:

1. Naive or "model based", $V(\hat{\beta}) = [\sum_{x=1}^{N} (D_i)^T \hat{V_i^{-1}} (D_i)]^{-1}$
2. Robust or “empirical” $V(\hat{\beta}) = M_0^{-1} M_1 M_0^{-1}$

Where:
$M_0 = \sum_{x=1}^{N} (D_i)^T \hat{V_i^{-1}} (D_i)$
$M_1 = \sum_{x=1}^{N} (D_i)^T \hat{V_i^{-1}} (Y - \hat{\mu_i})(Y - \hat{\mu_i})^T \hat{V_i^{-1}} (D_i)$

Here, $\hat{V_i}$ denotes $(Y - \hat{\mu_i})(Y - \hat{\mu_i})^T$

In the more general case, the robust or “sandwich” estimator, provides a consistent estimator of $V(\hat{\beta})$ (even if the working correlation structure is not the true correlation of Y_i).

Variable Selection Technique for GEE

In both GEE and GLMM, to select the significant covariates, avoid non-significant variables one by one starting from the most non-significant terms and finally the two models are compared using generalized Wald test for GEE and likelihood ratio test for GLMM. (Patetta, 2002)

Model Comparison Technique for GEE

Quasi Information Criterion (QIC): The quasi-likelihood counterpart to the AIC is the QIC, or the "quasi-likelihood under the independence model information criterion" (Pan, 2001). The QIC was derived from the AIC and they are conceptually the same. The quasi-likelihood function takes the following form: (McCullagh and Nelder, 1989)

$Q(\mu) = \int_y^\mu \frac{y-t}{\phi v(t)} dt$. Where $\mu = E(Y)$ and $Var(Y) = \phi V(\mu)$ with ϕ being the dispersion parameter.
An equation for the QIC is:
QIC = $-2Q(\hat{\mu}, I) + 2\text{trace}[\Omega_I^{-1}\hat{V}_R]$

Where I represents the independent correlation structure and R is the specified working correlation structure. The p-dimensional matrices Ω_I and $\hat{V}_R$ are variance estimators of the regression coefficients under the correlation structure I and R respectively. The QIC value is computed based on the quasi-likelihood estimate $\hat{\mu}$ and is used to select the appropriate working correlation structure for the model. However, Hin and Wang (2009) proposed using half of the second term in QIC is appropriate for the selection of the working correlation structure in GEE. This statistic is called the Correlation Information Criterion (CIC).
CIC = $\text{trace}[\Omega_I^{-1}\hat{V}_R]$
The first term in QIC, which is based on the quasi-likelihood, is free from both the working correlation structure as well as the true correlation structure, so it would not be informative in the selection of the covariance structure. Moreover, the form of quasi-likelihood is constructed under the assumption of the independent observations, although the parameters are estimated under the hypothesized working correlation structure. On the other hand, the second term in QIC contains information about the hypothesized correlation structure via the sandwich variance estimator.

The generalized Wald test: is used to compare models with different subsets of the regression parameters. That is, one can use the generalized Wald tests to test the joint null hypothesis that a set of regression parameters βs are equal to zero (Hedeker and Gibbons, 2006). In general, for any matrix L a test for hypothesis can be written as follows:

$H_0 : L\beta = 0$ versus $H_1 : L\beta \neq 0$

Where L is a $p \times q$ indicator matrix of ones and zeros. Here, p is equal to the number of parameters in the full model (including the intercept) and q equals the number of parameters in the generalized Wald test (that is, the difference in parameters between the full and reduced model). The Wald statistic is a quadratic form defined as follows:
$W^2_{stat} = \hat{\beta}^t L^t (LVar(\beta)L^t)^{-1} L\hat{\beta}$
And is distributed as χ^2 with q degrees of freedom under the null hypothesis.

3.2.3 Generalized Linear Mixed Models (GLMM)

Consider again a longitudinal study with T time points and let Y_{it} and X_{it} denote the same response and predictors/covariates as in the GEE above in (3.7). The GLMM is specified by:

$$Y_{it}|X_{it}, Z_{it}, b_i \sim f(\hat{\mu}_{it}).$$
$$g(\mu_{it}) = X^T_{it}\hat{\beta} + Z^T_{it}b_i, b_i \sim N(0, \Sigma_b), 1 \leq i \leq n, 1 \leq t \leq T \quad \textbf{(12)}$$

where $N(\mu, N(\mu, \Sigma)$ denotes a multivariate normal with mean μ and variance Σ, Z_{it} is a vector of predictors/ covariates (often set equal to X_{it}), and $g(\mu)$ is the appropriate link function for the type of response Y_{it}. The vector of latent variables, b_i , is called the random effects, denoting individual differences from the population mean b_i , which is known as the fixed effects. Although $\hat{\beta}$ is typically assumed to follow a multivariate normal as in (3.3), other types of distributions may also be considered.

Unlike the GEE, the GLMM accommodates correlated responses Y_{it} by directly modeling their joint distribution. Latent variables b_i are generally employed to model the correlated responses. Thus, although Y_{it} is still modeled for each time point t, by including the random effect b_i in the specification of the conditional distribution of Y_{it} given b_i(X_{it} and Z_{it}), the GLMM in (3.12) allows the resulting Y_{it} 's to be correlated (conditional on X_{it} and Z_{it} only). This approach allows one to specify multivariate distributions using familiar univariate distributions such as the Bernoulli (for binary responses) and the Poisson (for count responses).

Suppose that Y_{it} is binary and is modeled by a GLMM for binary response as follows:

$$Y_{it}|X_{it}, Z_{it}, b_i \sim Bernoulli(\hat{\mu}_{it})$$

$$logit(\hat{\mu}_{it}) = X_{it}^T\hat{\beta} + Z_{it}^T b_i \sim N(0, \Sigma_b), 1 \leq i \leq n, 1 \leq t \leq T \quad \textbf{(13)}$$

We note that models (3.8) and (3.13) look the same, except for the additional random effect in (3.13).

Method of Estimation and Statistical Inference for GLMM

Maximum likelihood (ML) by Laplace approximation technique is used to estimate the parameters. ML estimates standard deviations of the random effects assuming that the fixed-effect estimates are correct. The following derivations are done with respect to ML. Such likelihood may involve high-dimensional integrals that cannot be evaluated analytically so that much software are able to solve such complex manipulation using iteration technique. The likelihood of the data expressed as a function of unknown parameters is:

$$L(\beta, \alpha, Y) = \pi_{i=1}^{m} \int \pi_{i=1}^{ni} f(Y_{ij}|\beta, b_i) f(b_i|\alpha)\, db_i$$

It is the integral over the unobserved random effects of the joint distribution of the data and random effects. With Gaussian data, the integral has a closed form solution and relatively simple methods exist for maximizing the likelihood or restricted likelihood. With non-linear models, numerical techniques are needed. We consider the random effects as no missing data so that the 'complete' data for a unit is (Y_{i1}, b_i).

Denote $L = log(L)\ and\ \mu_{ij} = g^{-1}(X_{ij}\beta + Z_{ij}b_i)$ the score equation for β and b are:

$$\frac{\partial L}{\partial \beta} = S(\beta, \alpha|yb) = \sum_{i=1}^{m} \sum_{i=1}^{ni} X_{ij}(Y_{ij} - \mu_{ij}) = 0$$

The score equation for G is:

$$S(\beta, \alpha|yb) = \frac{1}{2} G^{-1}\{E(b_i b_i^t|y)\}G^{-1} - \frac{m}{2} G^{-1}$$

Where, G is variance covariance matrix for random effect. Hereby α denotes the unknown parameter in the density. These are solved using the E-M algorithm. In the estimation step, the expectations are evaluated using current parameter values and this may involve multivariable integration of large dimension. This will usually be done by Monte-Carlo integration.

Model Comparison Technique for GLMM

Akaike's Information Criterion (AIC) is used which is a measure of goodness of fit of an estimated statistical model. It is a tool for model selection. The AIC penalizes the likelihood by the number of covariance parameters in the model, therefore:
$AIC = -2Log(L) + 2p$
Where, L is the maximized value likelihood function for the estimated model and p is the number of parameters in the model. The model with the lowest AIC value is preferred.
Likelihood Ratio Test (LRT): It is constructed by comparing the maximized log likelihoods for the full and reduced models respectively and the test statistic is defined as:
$T^2_{LR} = -2ln(\frac{L_{ML}(\hat{\alpha_{ml,0}})}{L_{ML}(\hat{\alpha_{ml}})})$

Where $\hat{\alpha_{ml,0}}$ and $\hat{\alpha_{ml}}$ are respective maximum likelihood estimates which maximize the likelihood functions of the reduced and full model. The asymptotic null distribution of the likelihood ratio test statistic is a chi-square distribution with degrees of freedom equal to the difference between the numbers of parameters in the two models.

Model Checking Technique for GLMM

In GLMM, it is assumed that the random effects are normally distributed and uncorrelated with the error term. Residual plots can be used visually to check normality of these effects and to identify any outlying effect categories. Examining the plot of the standardized residuals versus fitted values by any covariates of interest can give a better feeling (Verbeke and Molenberghs, 2009).

3.3 Ethical Considerations

A permission to undertake the study has been obtained from Ethical Clearance at Strathmore University Institutional Ethics Review Committee (SU-IERC).

4 Data Analysis and Results

4.1 Results

A total of 193 HIV positive patients were included to this study. Some patients were on differentiated care others were not. The total number of observations was 480 for 8 variables. The baseline characteristics of patient observations are displayed in table 4.1 below. Among these observations 61(69.3%) males were not on differentiated care, 27(30.7%) were on differentiated care. 275(70.2%) observations were females not on differentiated care while 117(29.8%) were on differentiated care. 180(62.9%) of observations without TB were not on differentiated care while 106(37.1%) without TB were on differentiated care. 156(80.4%) observations not on differentiated care had TB, while 38(19.6%) observations with TB were on differentiated care. 245(70.4%) observations not on differentiated care were virally suppressed, while 103(29.6%) observations on differentiated care were virally suppressed. 91(68.9%) observations that were not virally suppressed were not on differentiated care, while 41(31.1%) observations that were not virally suppressed were on differentiated care. 143(72.6%) observations not on differentiated care were on CD4 category 1, while 54(27.4%) observations in CD4 category 1 were on differentiated care. 90(68.7%) observations on CD4 category 2 were not on differentiated care while 41(31.3%) observations were on differentiated care. 58(66.7%) observations on CD4 category 3 were not on differentiated care while 29(33.3%) observations were on differentiated care. 45(66.7%) observations on CD4 category 4 were not on differentiated care while 20(33.3%) observations were on differentiated care. The mean age for the last visit not on differentiated care was 21.63, the confidence interval was [20.70, 22.56]. The mean age for the last visit on differentiated care was 20.22, the confidence interval was [19.01, 21.42].

4.1.1 Exploratory Data Analysis

From the scatter plot matrix given in Figure 4.1, the diagonal alignment of the variables shows that the correlation structure is the independence structure. With this structure the correlations between subsequent measurements are assumed to be zero or measurements are independent to each other within individuals.

As shown in Figure 4.2, increase in CD4 counts leads to the movement towards differentiated care. More CD4 counts in a HIV positive patient, implies that the patient is stable. Stable patients should be placed under differentiated care.

Table 4.1. Baseline demographic and clinical characteristics of HIV patients on ART data who were either initiated or not initiated on differentiated care

		Differentiated Care			
		No	Yes	Total	p-value
CD4Category	0-200	143(72.6%)	54(27.4%)	197(100.0%)	0.75
	200-400	90 (68.7%)	41 (31.3%)	131 (100.0%)	
	400-600	58 (66.7%)	29 (33.3%)	87 (100.0%)	
	> 600	45 (69.2%)	20 (30.8%)	65 (100.0%)	
Total		336(70.0%)	144(30.0%)	480(100.0%)	
VL Suppression Category	$< 1,000$ copies	245 (70.4%)	103 (29.6%)	348 (100.0%)	0.755
	$> 1,000$ copies	91 (68.9%)	41 (31.1%)	132 (100.0%)	
Total		336 (70.0%)	144 (30.0%)	480 (100.0%)	
TB HIV Coinfection	No	180 (62.9%)	106 (37.1%)	286 (100.0%)	< 0.001
	Yes	156 (80.4%)	38 (19.6%)	194 (100.0%)	
Total		336 (70.0%)	144 (30.0%)	480 (100.0%)	
Gender	Male	61 (69.3%)	27 (30.7%)	88 (100.0%)	0.877
	Female	275 (70.2%)	117 (29.8%)	392 (100.0%)	
Total		336 (70.0%)	144 (30.0%)	480 (100.0%)	
Mean age at last visit		21.63[20.70, 22.56]	20.22[19.01, 21.42]		

From Figure 4.3, it is evident that absence of TB-HIV co-infection leads to the movement towards differentiated care. The highest point of the curve, towards differentiated care, is where there is absence of TB-HIV co-infection. Absence of TB-HIV co-infection implies a stable patient. Stable patients qualify for differentiated care.

The horizontal line in Figure 4.4 suggests that gender does influence whether or not a HIV positive patient should be initiated to differentiated care.

From Figure 4.5, the horizontal line suggests that there is no significant relationship between viral loads suppression category and differentiated care.

4.1.2 Modeling Differentiated Care Using GEE and GLMM

Generalized Estimating Equations (GEE)

In this section the ART data is analyzed using the generalized estimating equation. The output was as shown in Table 4.2:

The resulting output from Table 4.2 shows that the working correlation structure is the independence structure.

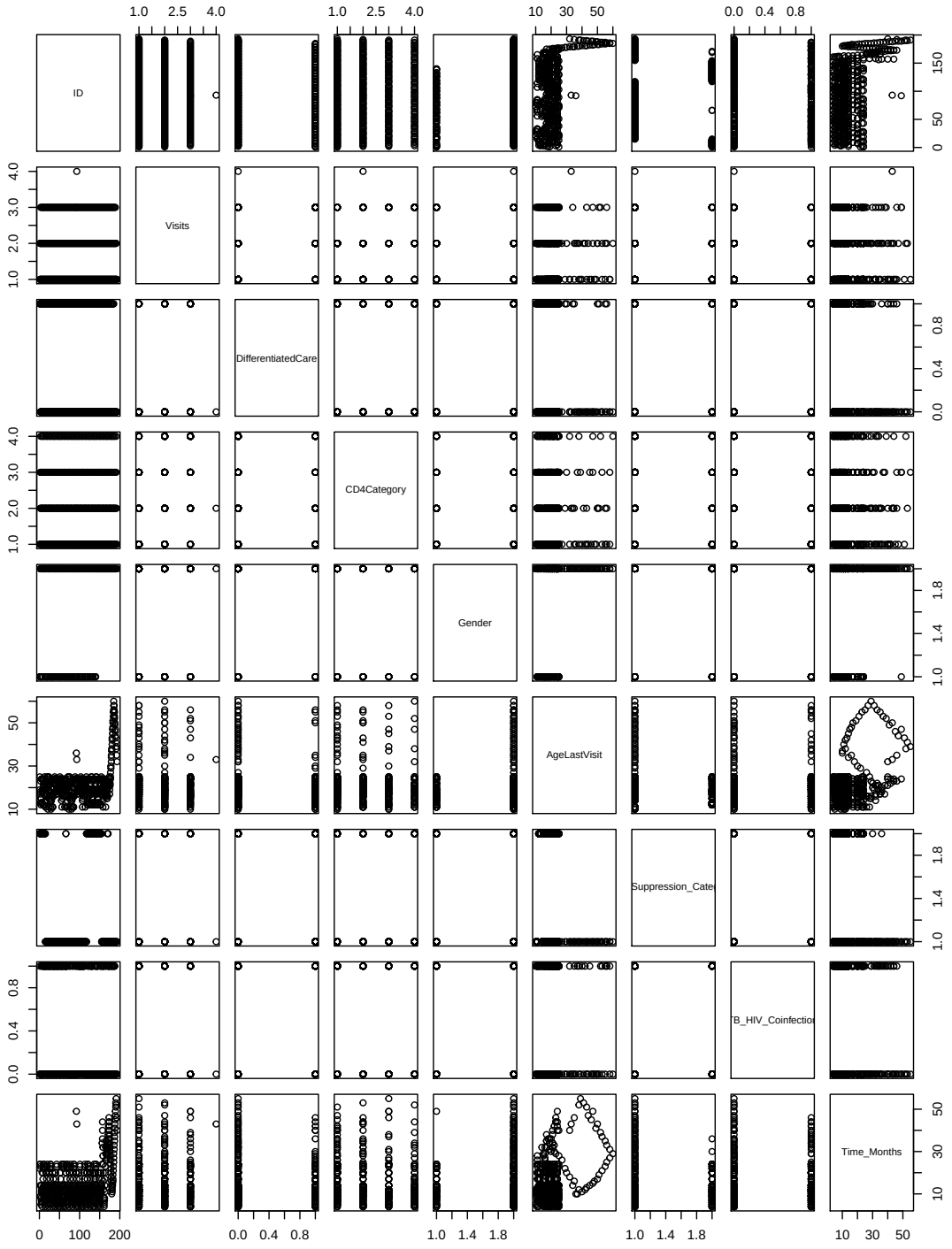

Figure 4.1. Scatter plots matrix of the variables

Looking at the p-values in Table 4.2 for each of the coefficients in the output, and comparing with the level of significance 0.05, only TB-HIV Coinfection is a significant predictor of whether or not a patient will need Differentiated Care (p-value = $0.00028 < 0.05$). As shown

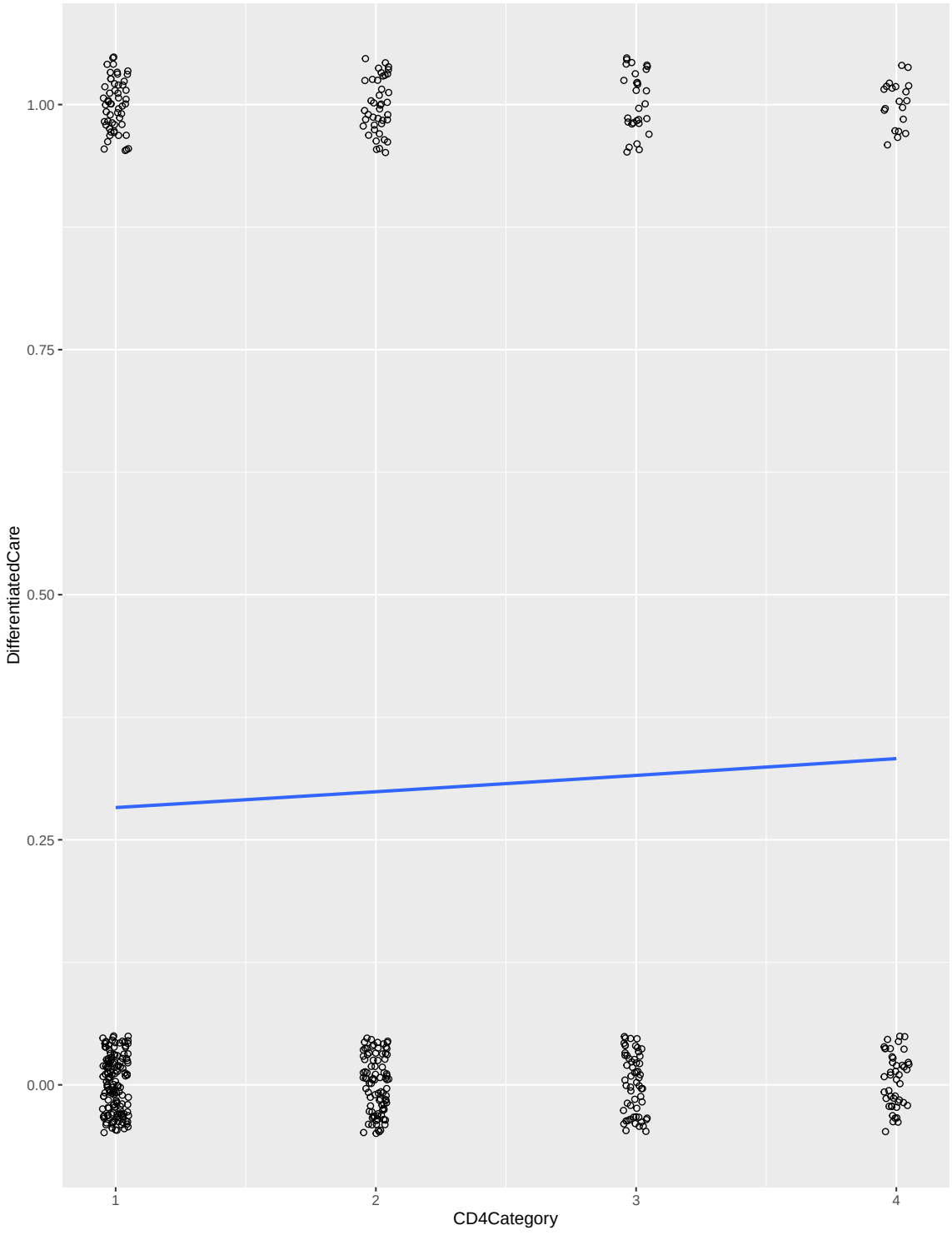

Figure 4.2. Plot of differentiated care vs CD4 Category

in table 2, the non-significant predictors were CD4 Category, Gender and VL Suppression Category.

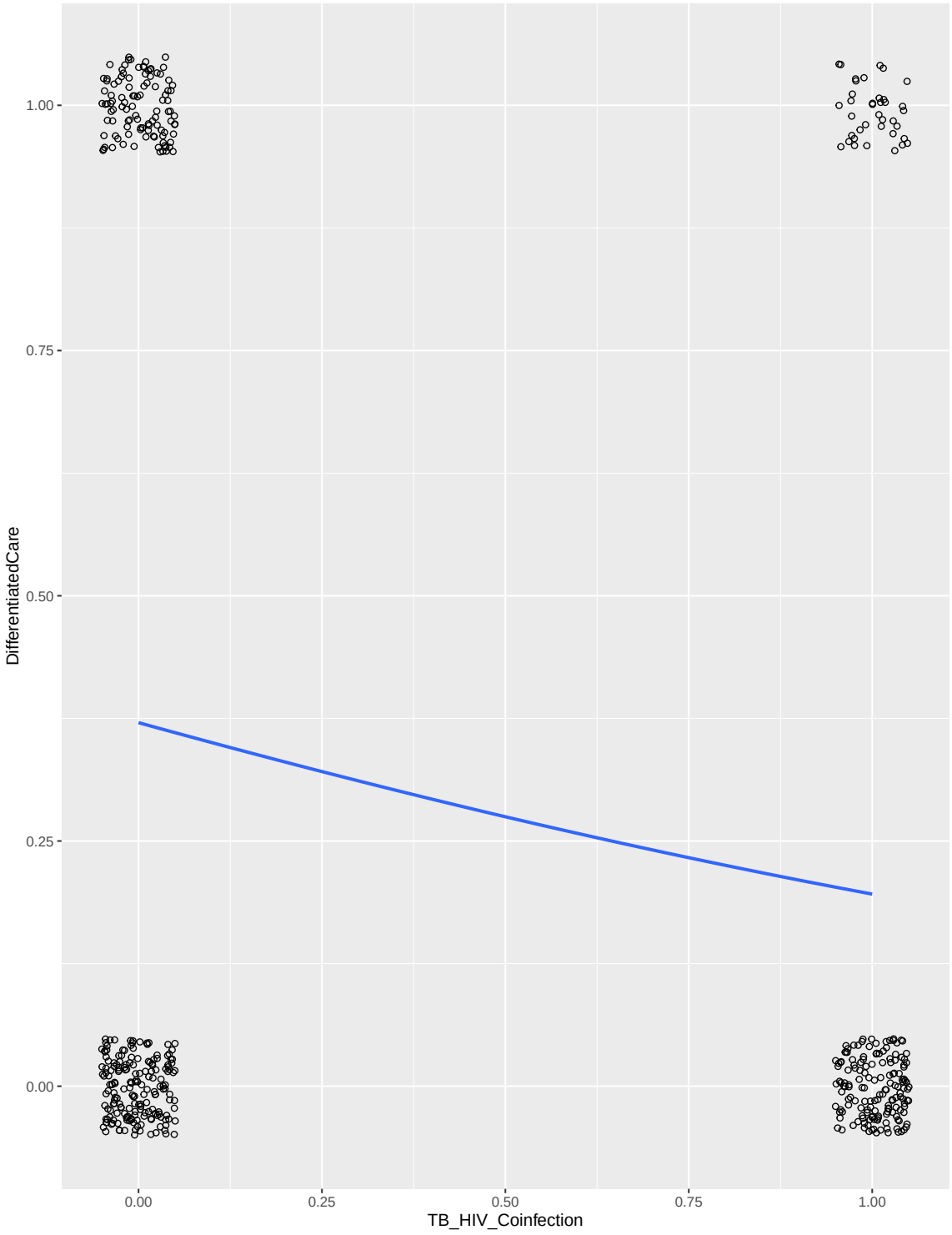

Figure 4.3. Plot of differentiated care vs TB-HIV co-infection

Interpretation of regression coefficients: The coefficient for TB-HIV Coinfection is -0.8789. This gives an odds ratio of 0.42 i.e. $(exp^{-0.8789})$. Now 100(1-0.42)% = 58%, thus a patient with TB-HIV Co-infection is 58% less likely to be put on differentiated care as

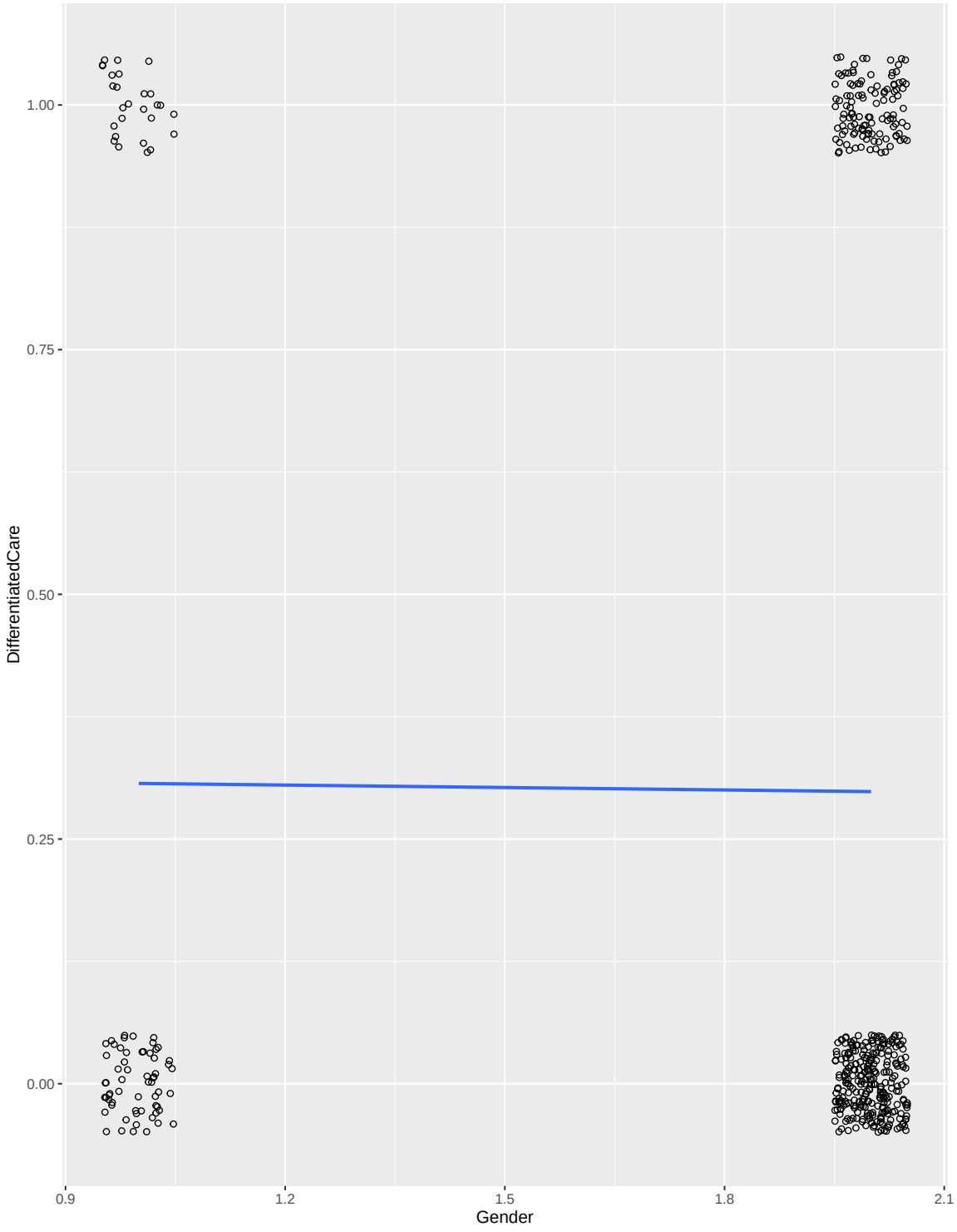

Figure 4.4. Plot of differentiated care vs gender

compared to a patient who is not TB-HIV co-infected. This patient is not likely to be stable. Differentiated care is for the stable patients.

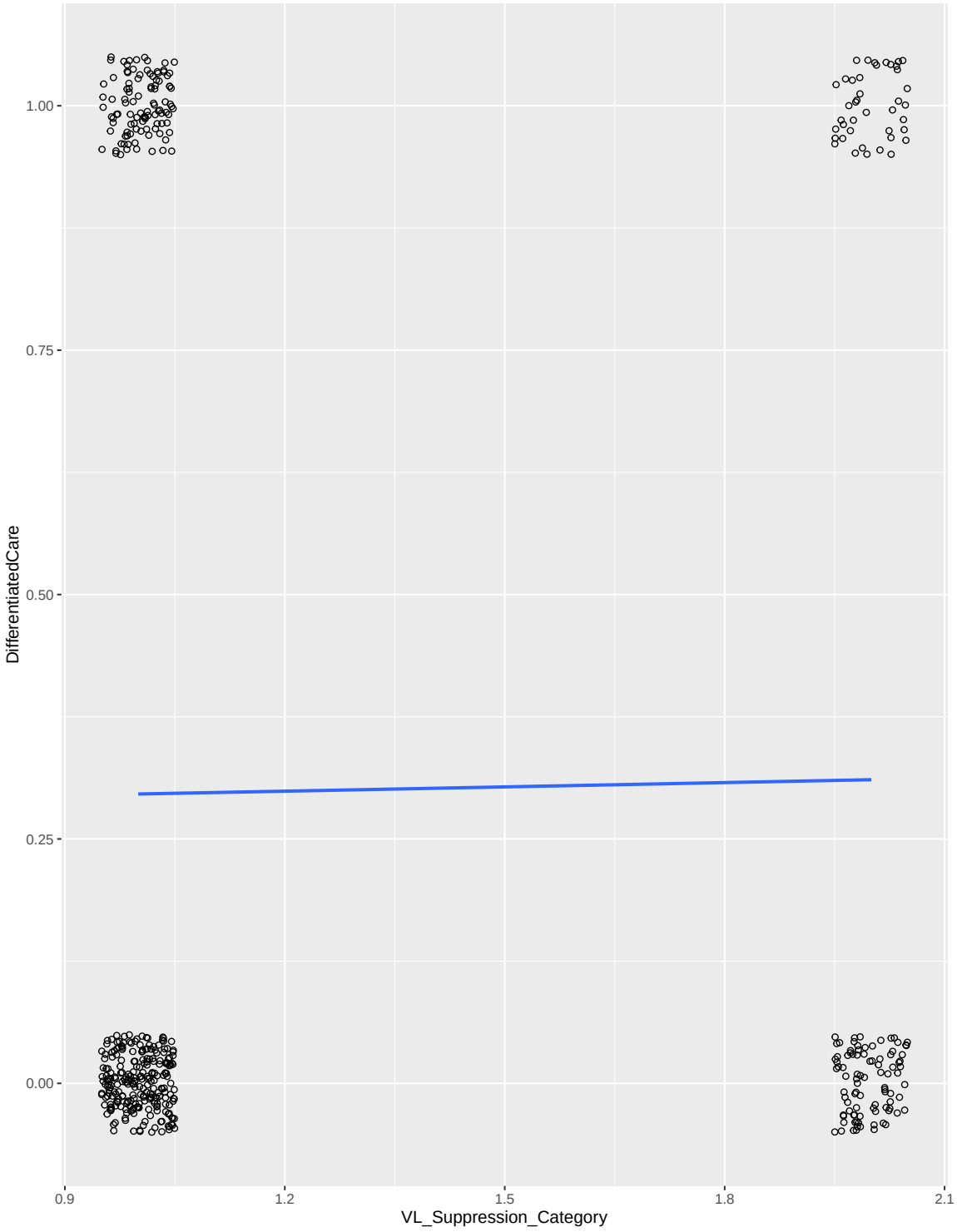

Figure 4.5. Plot of differentiated care vs viral loads suppression category

Generalized Linear Mixed Model (GLMM)

Table 4.3 shows the results from the GLMM.

Table 4.2. Resulting output from GEE

Coefficients:	estimate	san.se	wald	p
Inter:0	-0.62162125	0.3252790	3.652073873	0.0559994410
CD4Category2	0.07408740	0.2435099	0.092566877	0.7609389191
CD4Category3	0.23008731	0.2790967	0.679635344	0.4097123204
CD4Categor	0.12747292	0.3289662	0.150152795	0.6983893847
TB_HIV_Coinfection1	-0.87892649	0.2418493	13.207341572	0.0002788546
VL_Suppression_Category2	-0.02535246	0.2631146	0.009284325	0.9232384691
Scale is fixed				
Correlation Model:				
Correlation structure: independence				
Returned Error Value: 0				
Number of clusters: 193	Maximum cluster size: 4			

Table 4.3. Output from GLMM

AIC	BIC	logLik	deviance	df.resid		
573.4	606.8	-278.7	557.4	472		
Scaled residuals:						
Min	1Q	Median	3Q	Max		
-1.0927	-0.5463	-0.3728	0.8925	2.3468		
Random effects:						
Groups	Name	Variance	Std.Dev.			
ID	(Intercept)	1.163	1.078			
Number of obs: 480	groups:	ID, 193				
Fixed effects:						
Coefficients:	estimate	san.se	wald	p		
Inter:0	-0.85898	0.38197	-2.249	0.024525		
CD4Category2	0.06883	0.29756	0.231	0.817062		
CD4Category3	0.30553	0.33601	0.909	0.363189		
CD4Category4	0.22228	0.37674	0.590	0.555174		
Gender2	0.10179	0.32163	0.316	0.751642		
TB_HIV_Coinfection1	-1.06007	0.27602	-3.841	0.000123		
VL_Suppression_Category2	-0.03404	0.30881	-0.110	0.912218		
Correlation of Fixed Effects:						
	(Intr)	CD4Ct2	CD4Ct3	CD4Ct4	Gendr2	TB_HIV
CD4Categry2	-0.395					
CD4Categry3	-0.374	0.383				
CD4Categry4	-0.312	0.338	0.342			
Gender2	-0.732	0.047	0.067	0.024		
TB_HIV_Cnf1	-0.207	0.125	0.019	0.004	-0.047	
VL_Spprs_C2	-0.281	0.023	0.018	0.022	0.035	0.095

In longitudinal data analysis, what random effect should be included to the model in order to account for between individual variability is a critical issue. In this study ID was considered as a random effect and CD4 Category, TB-HIV Coinfection, Gender and VL Suppression Category as fixed factors. ID was used as a random effect since the patients in this study are a random sample of all the possible patients that can be included.

From the column standard deviation under random effects in Table 4.3, the standard deviation of 1.078 is a measure of how much variability in the dependent measure(differentiated care) there is due to ID.

From the output in Table 4.3, at 5% level of significance, the fixed effect TB-HIV coinfection significantly predicts Differentiated Care (p-value = 0.00012). In addition, the estimate of the variance for the random effect ID is 1.16. Since this differs from zero, it implies that there is variation in Differentiated Care for the different patients with TB-HIV Co-infection.

Interpretation of regression coefficients: The coefficient for TB-HIV Coinfection is -1.0601. This gives an odds ratio of 0.346 i.e. $(exp^{-0.8789})$. Now 100(1-0.346)% = 65.5%, thus a patient with TB-HIV Co-infection is 66% less likely to be put on differentiated care as compared to a patient who is not TB-HIV co-infected.This patient is not likely to be stable. Differentiated care is for the stable patients.

Significance of the model fit: The likelihood ratio test was used. We compare the GLMM model with a random effect for ID included to the binary logistic regression model with only the fixed factors. The logic of the likelihood ratio test is to compare the likelihood of two models with each other. First, the model without the factor that you're interested in (the null model), then the model with the factor that you're interested in.
1 - pchisq(568.33 - 557.41, 8 - 7) = 0.0009513161. The p-value for the test of the hypothesis that the GLMM model is not a significantly better fit than the binary logistic regression model is 0.00095. Thus at 5% level of significacnce we reject this hypothesis and conclude that the GLMM model is a significantly better fit.

We also compared the GLMM model with a random effect for ID included to the binary logistic regression model with only the fixed factors(GLM). The results of the GLM are shown in Table 4.4.

Comparing the output in Table 4.3 to Table 4.4, we find that the AIC for the GLMM is 573 while the AIC in the reduced model(GLM) is 582.33. The AIC of the model with the random effect(GLMM) was lower, hence the better model.

Table 4.4. Resulting output from GLM-Binary Logistic Regression

Deviance Residuals:				
Min	1Q	Median	3Q	Max
-1.0245	-0.9351	-0.6622	1.3839	1.8561
Coefficients:				
	Estimate	Std.Error	z value	Pr($>\|z\|$)
(Intercept)	-0.62162	0.29334	-2.119	0.0341
CD4Category2	0.07409	0.25287	0.293	0.7695
CD4Category3	0.23009	0.28357	0.811	0.4171
CD4Category4	0.12747	0.31852	0.400	0.6890
Gender2	0.02069	0.26174	0.079	0.9370
TB HIV Coinfection1	-0.87893	0.22120	-3.973	7.09e-05
VL Suppression Category2	-0.02535	0.22706	-0.112	0.9111
(Dispersion parameter for binomial family taken to be 1)				
Null deviance:	586.43	on 479 degrees of freedom		
Residual deviance:	568.33	on 473 degrees of freedom		
AIC: 582.33				
Number of Fisher Scoring iterations: 4				

5 Conclusions and Recommendations

The main objective of this research was to develop a data-driven longitudinal model with application to HIV differentiated care data. This objective was split into two specific objectives in order to adequately achieve the main objective. These specific objectives were as follows:

1. To model the association between longitudinally measured HIV variables and differentiated care using GLMM and GEE and to compare the two models.
2. To determine which statistical model is better using goodness of fit.
3. To identify factors which predict differentiated care.

In this study we evaluated the association between HIV differentiated care and its possible predictors using longitudinal models. Statistically two modeling approaches; the GEE and the GLMM which are extensions of the GLM were used for the analysis of ART data of patients who were either initiated or not initiated to differentiated care. We focused on interpretation and computation of model parameters. For parameter interpretation, we discussed differences between the GLMM and GEE when applied to model a binary response. Our binary response was differentiated care which is a new approach of managing HIV/AIDS.

The study also found TB-HIV co-infection to be the only significant predictor of differentiated care under both GEE and GLMM. In addition, from both models we saw that some of the fixed effect coefficients $(\beta_0, \beta_5, \beta_6)$ had opposite signs for the two models, this result is due to subject specific and population average interpretation for the two models. This supported the findings of (Fu, 2010; Renard, 2002). We also concluded that GLMM fits data better than GLM.

Since HIV/AIDS is a critical disease, modeling the HIV data helps to identify the factors that determine the success of ART so as to delay the quick progression of HIV. Thus further studies should be done in HIV research using these flexible statistical methodologies by including additional covariates like regimen, income status, weight, relationship status(single, married), number of years of being HIV positive in order to further improve the models' prediction performance. This would help in the monitoring and follow-up of the patients to ensure appropriate care is given.

Furthermore, though the choice between GEE and GLMM for longitudinal data can only be made on subject matter grounds, using GLMM is much emphasized than GEE for correlated data as GEE can only handle the within subject variations through the assigned working correlation structure where as GLMM in addition to within measurement variation, between individual variations can be accounted by incorporating the random effects. Due to that, GLMM fits a given data with a small disturbance than GEE. Although this research is motivated by HIV/AIDS studies, the basic concepts and methods developed here have much broader applications in management of other chronic diseases.

Bibliography

Archer, J. P. (2008). *The diversity of HIV-1*. PhD thesis, University of Manchester.

Boscardin, W. J., Taylor, J. M., and Law, N. (1998). Longitudinal models for aids marker data. *Statistical Methods in Medical Research*, 7(1):13–27.

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

Callaway, D. S., Ribeiro, R. M., and Nowak, M. A. (1999). Virus phenotype switching and disease progression in hiv-1 infection. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 266(1437):2523–2530.

Culshaw, R. V. (2006). Mathematical modeling of aids progression: limitations, expectations, and future directions. *Journal of American Physicians and Surgeons*, 11(4):101.

Degruttola, V., Lange, N., and Dafni, U. (1991). Modeling the progression of hiv infection. *Journal of the American Statistical Association*, 86(415):569–577.

Fu, S. (2010). *Regression approaches to estimation of relative risk: application to multiple sclerosis studies*. PhD thesis, University of British Columbia.

Gamerman, D. and Lopes, H. (1997). Markov chain monte carlo (texts in statistical science).

Grimsrud, A., Barnabas, R. V., Ehrenkranz, P., and Ford, N. (2017). Evidence for scale up: the differentiated care research agenda. *Journal of the International AIDS Society*, 20:22024.

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

Hin, L.-Y. and Wang, Y.-G. (2009). Working-correlation-structure identification in generalized estimating equations. *Statistics in medicine*, 28(4):642–658.

Hughes, M. D., Stein, D. S., Gundacker, H. M., Valentine, F. T., Phair, J. P., and Volberding, P. A. (1994). Within-subject variation in cd4 lymphocyte count in asymptomatic human immunodeficiency virus infection: implications for patient monitoring. *Journal of Infectious Diseases*, 169(1):28–36.

Lee, Y. and Nelder, J. A. (2001). Hierarchical generalised linear models: a synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, 88(4):987–1006.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Lin, G., Tu, J. X., Zhang, H., Hongyue, W., Hua, H., and Gunzler, D. (2016). Modern methods for longitudinal data analysis, capabilities, caveats and cautions. *Shanghai archives of psychiatry*, 28(5):293.

Lopez, W. (2011). Hiv/aids: A new era of treatment. *The York Scholar*, pages 11–17.

Lu, X. (2014). Statistical modeling and prediction of hiv/aids prognosis: Bayesian analyses of nonlinear dynamic mixtures.

Magder, L. S. and Zeger, S. L. (1996). A smooth nonparametric estimate of a mixing distribution using mixtures of gaussians. *Journal of the American Statistical Association*, 91(435):1141–1151.

McCullagh, P. (81). Nelder. 1989. generalized linear models. *London, chapman hall*.

McCullagh, P. and Nelder, J. (1989). Chapman and hall. *Generalized linear models*.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American statistical Association*, 92(437):162–170.

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

Organization, W. H. et al. (2016). *Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection: recommendations for a public health approach*. World Health Organization.

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

Phillips, A., Shroufi, A., Vojnov, L., Cohn, J., Roberts, T., Ellman, T., Bonner, K., Rousseau, C., Garnett, G., Cambiano, V., et al. (2015). Sustainable hiv treatment in africa through viral load-informed differentiated care. *Nature*, 528(7580):S68.

Renard, D. (2002). *Topics in modeling multilevel and longitudinal data*. PhD thesis.

Rosa, R. S., Santos, R. H., Brito, A. Y., and Guimarães, K. S. (2014). Insights on prediction of patients' response to anti-hiv therapies through machine learning. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 3697–3704. IEEE.

Taye, A. Z. (2012). Modeling the progression of hiv infection using longitudinally measured cd4 count for hiv positive patients following highly active antiretroviral therapy. Master's thesis, School of Graduate Studies, College of Natural Science, Jimma University.

Tunduny, T. K. (2017). *A HIV/AIDS viral load prediction system using artificial neural networks*. PhD thesis, Strathmore University.

Verbeke, G. and Molenberghs, G. (2009). *Linear mixed models for longitudinal data*. Springer Science & Business Media.

Wolfinger, R. and O'connell, M. (1993). Generalized linear mixed models a pseudo-likelihood approach. *Journal of statistical Computation and Simulation*, 48(3-4):233–243.

Wu, L. and Zhang, H. (2018). Mixed effects models with censored covariates, with applications in hiv/aids studies. *Journal of Probability and Statistics*, 2018.

Yu, T. and Wu, L. (2018). Robust modelling of the relationship between cd4 and viral load for complex aids data. *Journal of Applied Statistics*, 45(2):367–383.

Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects; a gibbs sampling approach. *Journal of the American statistical association*, 86(413):79–86.