A COMPARISON OF JOINT AND SEPARATE MODELS FOR REPEATED MEASUREMENTS AND TIME TO EVENT DATA: APPLICATION OF NUTRITION DATA FROM HIV POSITIVE PATIENTS AT QUEEN ELIZABETH CENTRAL HOSPITAL IN MALAWI

MSc. (BIOSTATISTICS) THESIS

ANDREW ACLAIN KUMITAWA

UNIVERSITY OF MALAWI

CHANCELLOR COLLEGE

OCTOBER, 2013

A COMPARISON OF JOINT AND SEPARATE MODELS FOR REPEATED MEASUREMENTS AND TIME-TO-EVENT DATA: APPLICATION OF NUTRITION DATA FROM HIV POSITIVE PATIENTS AT QUEEN ELIZABETH CENTRAL HOSPITAL IN MALAWI

MSc (BIOSTATISTICS)

By

ANDREW ACLAIN KUMITAWA

Bachelor of Science - University of Malawi

Thesis submitted to the Department of Mathematical Sciences, Faculty of Science, in partial fulfillment of the requirements for the degree of Master of Science (Biostatistics)

UNIVERSITY OF MALAWI

CHANCELLOR COLLEGE

OCTOBER 2013

DECLARATION

I, the undersigned hereby declare that this thesis is my own original work which has not been submitted to any other institution for similar purposes. Where other peoples' work has been used acknowledgements have been made.

ANDREW ACLAIN KUMITAWA

Name

_____

Signature

_____

Date

CERTIFICATE OF APPROVAL

The undersigned certify that this thesis represents the student's own work and has been submitted with our approval.

Signature:_____ Date:_____

Mavuto F Mukaka, MSc, PhD (Lecturer)

Supervisor

Signature:_____ Date:_____

Tsirizani Kaombe, MSc (Lecturer)

Programme Coordinator

DEDICATION

This thesis is dedicated to my beloved wife, Christina and my two children Emmanuel and Chimwemwe Kumitawa. The thesis is also dedicated to my late mother, Juliet Kumitawa, who died while I was still studying for the master's course.

ACKNOWLEDGEMENTS

I am thanking the Almighty God for making this work a success. May all the glory be to the Almighty God. I am very grateful to my supervisor Dr. M.F Mukaka for his support and assistance during the entire process of writing this thesis. Your time, resources and constructive criticism has made this work a success. Thanks to Dr B Ngwira for his valuable advice.

Special thanks should go to Prof Mark Manary for providing me with the dataset.

Thanks should also go to the coordinator of the programme, Mr. Tsirizani Kaombe and all staff members of Department of Mathematical Sciences for their Support.

Special thanks go to the Department of Community Health, College of Medicine, University of Malawi, for paying the tuition fees.

My special thanks should go to my wife, Christina and my two children Emmanuel and Chimwemwe Kumitawa. I am also indebted to all my classmates, dad, and my late mum, and my brother, Howard and sisters, Maggie and Chrissie, who provided me with moral support.

ABSTRACT

Sometimes clinical trials collect survival data, which have some variables measured longitudinally. This type of data is mostly analyzed using Cox proportional models with time dependent covariates. The longitudinal variables are treated as time dependent covariates. When there is association between a longitudinal variable and the time to event, estimates produced from separate models may be biased. The study uses Cox proportional models with time dependent covariates for survival data and linear mixed effects regression models for the longitudinal data. For the joint analysis, the joint modeling between repeated measurement and time to an event is used. The method is applied to data from a randomized clinical trial for the malnourished HIV positive patients who were on ART at Queen Elizabeth Central Hospital. One group received corn soya blend (CSB) and other group received ready to use therapeutic food (RUTF). Results from joint modeling showed that there is significant association between body mass index (BMI) and time to death of a patient, $p < 0.001$. Both joint model and Cox proportional model with time dependent covariates showed that the type of food did not have significant effect on the time to death of patients. Hemoglobin levels, sex of patient and use cotrimoxazole were significantly associated with time to death of malnourished HIV positive patients. It was also observed that some variables which were not significant in the separate models became significant in the joint model. This shows the importance of using joint models. Joint modeling of longitudinal and survival data gives unbiased estimates.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

ACRONYMS AND ABBREVIATIONS

AFT	Accelerated Failure Time
AIC	Akaike Information Criteria
AIDS	Acquired Immune Deficiency Syndrome
ANOVA	Analysis of Variance
BIC	Bayesian Information Criteria
BMI	Body Mass Index
CI	Confidence Interval
COEF	Coefficient
CPM	Covariance Pattern Model
CSB	Corn Soya Blend
DIC	Deviance Information Criteria
EM	Expectation Maximization
EXP	Exponential
GEE	Generalized Estimating Equations
GLMM	Generalized Linear Mixed Models
HIV	Human Immunodeficiency Virus
HR	Hazard Ratio

JLCM	Joint Latent Class Model
LME	Linear Mixed Effects Regression Models
LVCF	Last Value Carried Forward
MANOVA	Multivariate Analysis of Variance
MAR	Missing At Random
MCAR	Missing Completely At Random
MCMC	Markov Chain Monte Carlo
MI	Multiple Imputation
MLE	Maximum Likelihood Estimation
MNAR	Missing Not At Random
MRM	Mixed Effect Regression Model
NLME	Non Linear Mixed Effects Regression Model
PH	Proportional Hazard
RUTF	Ready to Use Fortified Food
SD	Standard Deviation
SE	Standard Error
SREM	Shared Random Effects Model
TB	Tuberculosis

CHAPTER 1 INTRODUCTION

This chapter introduces and states the problem being researched and explains why this topic was chosen and how data was collected.

1.1 Background

Clinical trials sometimes collect survival data. Survival data is defined as data, which its response of interest is time until some event occurs (Kalbfleisch & Prentince, 2002). Survival analysis is a statistical method used to analyze data when the outcome of interest is time to occurrence of an event. Survival analysis is also called time to event analysis. In medical field, time to event can be time until recurrence of tumor in a cancer study, time to death after surgery, or time until infection (Kalbfleisch & Prentince, 2002; Collet, 2003).

Standard statistical techniques cannot usually be applied to analyze survival data because the underlying distribution is rarely normal and the data is often censored (Bewick et al., 2004). A variable is said to be censored when there is a follow up time but the event has not yet occurred or is not known to have occurred (Bewick et al., 2004; Kalbfleisch & Prentince, 2002).

In clinical trials, longitudinal data are collected. Repeated measurement variables, which are variables collected repeatedly over a long period of time are analyzed using longitudinal data methods. There are several methods, which are used to analyze

longitudinal data. Some of them are univariate analysis of variance (ANOVA), multivariate analysis of variance (MANOVA), mixed effects regression models (MRM), covariance pattern models (CPM) and generalized estimating equations (GEE).

Several methods also exist for analyzing survival data. Among them are non parametric, semi parametric and parametric methods. Many textbooks have been written in order to address survival data analysis. These books include: Kalbfleisch and Prentince (2002), Collet (2003) and Machin, Cheung and Parmar (2006).
Kaplan and Meier (1958) proposed a non parametric method that is widely used as a starting point in the field of survival data analysis. Non-parametric methods are suited for the homogeneous samples.

Another method used to analyze survival is Cox proportional hazard model. The Cox proportional hazards model, which was proposed by Cox (1972), is now the most widely used approach for the analysis of survival data (Hosmer & Lemeshow, 1999). Despite the Cox proportional hazard model being the most widely used, in some situations, it may not be the appropriate method to use especially when proportion hazards (PH) assumptions do not hold. The extensions of Cox PH models such as stratified Cox model and Cox model with time dependent variables can be used for the analysis of survival data when PH assumptions fail to hold.

Survival data with baseline covariates and repeated measurements covariates can be analyzed using Cox proportional model with time dependent covariates (Collet, 2003). Effects of repeated measurements on the time to an event are assessed by

treating the repeated measurement as a time dependent covariate in a survival model (Nguti, Burzykowski, Rowlands & Janssen, 2005). The problems associated with this modeling approach were well described by Tsiatis, Degruttola and Wulfsohn (1995). If there is association between the repeated measurements data and time to an event, for example time to death, Cox model with time dependent models may not be appropriate approach of modeling this type of data (Nguti et al., 2005; McCrink, Marshall & Cairns, 2011). Time dependent covariates Cox models produce biased estimates when there is association between time to an event and repeated measurement variables (Nguti et al., 2005). In that case, joint modeling of survival and repeated measurement data becomes a better approach to use. Joint modeling of survival and longitudinal data may be used to analyze data, when both repeated measurements and time to an event are collected and these variables are associated. These two processes, namely repeated measurements and time to event, are associated through unobserved random effects (Tsiatis et al., 1995; McCrink et al., 2011). When there is an association between repeated measurements and time to an event, joint modeling gives better results (Henderson et al., 2000; Ibrahim, Chu & Chen, 2010). In fact Ibrahim et al. (2010) has demonstrated that joint modeling can improve the accuracy of the estimation for parameters in both models when the longitudinal measurements and survival times are highly correlated. In particular, Little and Rubin (2002) reported that joint modeling produces smaller standard error of estimates. With accurate estimates of parameters, the right conclusion on the effect of repeated measurement covariates on the survival of the individual can be made (McCrink et al., 2011). Nguti et al., (2005) reported that estimates from separate analysis (i.e. survival model and longitudinal data model) have been shown to be biased towards zero, thus

showing over estimated hazard ratios, and this bias can be reduced by using joint models.

Diggle, Sousa and Chetwynd (2008) reported several advantages of joint modeling of the repeated measurement and the time-to-event processes. They reported that the repeated measurements can be extrapolated from observed measurement times to the specific event time in a way that utilizes the entire measurement history. They also reported that the time to the event is allowed to depend on the true but unknown value of the repeated measurement, thus making adjustment of measurement error. This in turn leads to reduced bias of the parameter estimates of the Cox model. Also the repeated measurement process is adjusted for any loss of information arising from death or loss of individuals. When there is no association between longitudinal repeated measurements and event to survival, joint modeling reduces to separate survival data and longitudinal data methods (McCrink et al., 2011).

This study used secondary data that was collected in 2006. In the study, malnourished adults HIV positive patients were given one of the two food supplements (CSB or RUTF). Time to death of malnourished HIV positive patient receiving ART was the event of interest. However, weights of patients were also collected longitudinally. It is likely that body mass index (BMI) was associated to time of death of a malnourished HIV positive person.

Zechariah et al. (2006) reported in their paper entitled “Risk factors for high early mortality in patients on ART in rural district of Malawi” that many HIV infected patients in Malawi died within the first 3 months after the initiation of antiretroviral

therapy (ART). The aim of the study was to find factors associated with time to death of malnourished HIV positive patients receiving ART, but also to assess factors associated with the longitudinally collected body mass index (BMI). In adults, BMI is a measure used to indicate whether a person was underweight or not.

The data for this study was collected from the randomized clinical trial. In the clinical trial, data was collected from malnourished HIV positive at Queen Elizabeth Central Hospital in Blantyre, Malawi. Nutritional support in terms of food supplements were given to the patients. Nutritional support was identified as one of the most immediate and critical needs for patients living with HIV/AIDS (Manary et al., 2010; Ndekha et al., 2009).

1.2 Statement of the Problem

Wasting is a major problem in sub Saharan African among adults with advanced HIV infection and the prevalence of wasting ranges from 20% to 40% (Dannhauser et al., 1999; Van der Sande et al., 2004). Wasting is normally the result of inadequate nutrient intake because of anorexia, food insecurity associated with poverty, catabolic state induced by opportunistic infection or malignancy, or poor absorption of nutrients secondary to diarrhoea and malabsorption (Ndekha et al., 2009). Wasting is one of the risk factors of death among adults with advanced HIV infection in Sub Sahara. In Malawi, supplementary feeding together with treatment is advocated as the standard care of wasted adults with HIV in Malawi (Ndekha et al., 2009).

Corn soya blends (CSB) and ready to use fortified spreads (RUTF) are some of the supplementary foods given to HIV patients who are malnourished and receiving

antiretroviral therapy. Studies have shown than RUTF resulted in greater increase in BMI as compared to CSB (Manary et al., 2010; Ndekha et al., 2009). However it is not clear if RUTF and CSB have effect on the time to death of malnourished HIV infected patients who are on ART. Zachariah et al. (2006) reported that in Malawi mortality during the first 3 months of antiretroviral therapy is high, and a low BMI is associated with this early mortality. There is a need to assess the effects of CSB and RUTF on the time to death of malnourished HIV infected patients who are on ART.

When the aim of the study is to assess the effects of the repeated measurements on time to death of patient, Cox models with time dependent covariates are used (Sousa, 2011; Nguti et al., 2005). The effect of repeated measurements on the time to death has been assessed by treating the repeated measurement as a time dependent covariate in a survival model (Nguti et al., 2005). The effects of other covariates on the outcome variable such as body mass index have been analyzed using mixed effect regression model. The survival component and the longitudinal data components have been analyzed separately. It is therefore necessary to analyze survival data and longitudinal data simultaneously using joint models because there is association between lower body mass index and time to death of a patient (Manary et al., 2010; Zechariah et al., 2006). Joint modeling takes care of this association.

1.3 General Objective

The purpose of this study is to compare the results from joint and separate models for the repeated measurements and time to death for the malnourished HIV positive patients who are receiving antiretroviral therapy (ART) at Queen Elizabeth Central Hospital in Malawi.

1.3.1 Specific Objectives

- To model the effects of CSB and RUTF on the time to death of malnourished HIV infected patients who are on ART.
- To assess the relationship between body mass index (BMI) and time to death in malnourished HIV positive patients.
- To model jointly the body mass index and time to death for the malnourished HIV infected patients who are on ART.
- To model separately the repeated measurements and time to an event data.
- To compare the estimates from models produced by separate methods and joint modelling methods.

1.4 Significance of Study

Analysis of survival and longitudinal data poses a challenge when there is an association between time to an event variable and the repeated measurement variable. Using time dependent Cox model to analyze survival data with longitudinal variable may give biased results especially when there is association between the time to an event of interest and longitudinal variable. The statistical approach (joint modeling) used in this paper reduces bias and produces smaller standard errors (Henderson et al., 2000).

This paper will help to add knowledge in the field of nutrition especially among malnourished HIV positive patients. Few studies in the field of nutrition have used joint modeling of survival and longitudinal data approach. Therefore, this paper intends to add to the available work in the modeling of survival data when there is association between the survival event and repeated measurement data.

1.5 Structure of the thesis

This thesis is structured as follows: Chapter 2 reviewed literature on methods used to analyze survival data, longitudinal data and joint modeling. The methodology for this thesis is presented in chapter 3. Results for survival model, longitudinal model and joint modeling are presented in chapter 4. Discussion of results is presented in Chapter 5. Finally chapter 6 gives conclusions and recommendations.

CHAPTER 2 LITERATURE REVIEW

This chapter gives the literature review of survival analysis, longitudinal data analysis and joint modeling

.

2.1 Longitudinal Data Model

The longitudinal data are measurements collected repeatedly over a period of time from the same individual. The purpose of a longitudinal study is to show the effect or change of outcome variable over time and the factors which influence the change. In the subsequent sections, the common approaches for handling longitudinal data are reviewed. These include repeated measures ANOVA, MANOVA, mixed effects regression models and generalized estimating equations. This section uses information from the following books: Hedeker and Gibbons (2006), Faraway (2006) and Diggle, Heagerty, Liang and Zeger (2002).

2.1.1 Repeated Measure Analysis of Variance (ANOVA)

Repeated measure of analysis of variance (ANOVA) is the approach used to provide analysis of complete data. It works by regarding time as a factor on n levels in a hierarchical design with units as sub plots (Diggle et al., 2002). Diggle et al. (2002) described repeated measure ANOVA in the following way:

Consider the following model

$$y_{hij} = \beta_h + \gamma_{hj} + U_{hi} + Z_{hij} \qquad 2.1.1$$

where y_{hij} denotes jth observation from the ith unit within the hth treatment group;

$j = 1,2, \ldots, n; \quad i = 1,2, \ldots., m_h; h = 1,2, \ldots., g$. The term β_h represents main effects for treatments and γ_{hj} is an interaction between treatments and time with constraints that $\sum_{j=1}^{n} \gamma_{hj} = 0$, for all h. In equation 2.1.1, U_{hi} and Z_{hij} are mutually independent random effects for units and measurement error respectively, $E(Y_{hij}) = \beta_h + \gamma_{hj}$. Assuming that both U_{hi} and Z_{hij} are normally distributed with zero mean and variance $v^2 \; and \; \sigma^2$ respectively, then $Y_{hi} = Y_{hi1}, \ldots., Y_{hin}$ is multivariate normal, which has $V = v^2 I + \sigma^2 J$ as its covariance matrix, where I is identity matrix and J is a matrix all of whose elements are I. The model has constant correlation $\rho = \frac{v^2}{v^2 + \sigma^2}$, between any two observations on the same unit.

ANOVA requires that the data must be balanced. In addition to this, ANOVA makes an assumption of sphericity. Rabe-Hesketh and Skrondal (2008) defined sphericity as the assumption that all pair-wise differences between responses have the same variance. The assumption of sphericity is rarely met when analyzing longitudinal data. Because of this ANOVA is limited in its application. When the assumption of sphericity is violated, it can lead to skewed F-distributions. The advantage of ANOVA is that it takes into account the fact that subjects can have individual baseline observations but no subject-specific evolution in time. When they are missing data, ANOVA uses observations which have complete data only (Hedeker & Gibbons, 2006).

2.1.2 Multivariate Analysis of Variance (MANOVA)

Multivariate analysis of variance (MANOVA) is another approach used to analyze longitudinal datasets and has similar restrictions as the univariate ANOVA described above. This method treats the n repeated measurements as n x1 response variable. The

usual approach involves transforming observations into orthogonal polynomial coefficients. For one sample MANOVA, let $y_i = u + \varepsilon_i$, y_i is an $n\,x1$ response vector for the n repeated measurements, u is an $n\,x1$ mean vector for time points, ε_i is an $n\,x1$ vector of errors and $\varepsilon_i \sim N(0, \Sigma)$, y_i is normally distributed with mean μ and variance Σ. Under univaraite approach $\Sigma = \sigma_n^2 1_n 1'_n + \sigma_e^2 I_n$ and $u = u1 + \tau$. MANOVA does not handle missing values in data and in addition it assumes the variables to be measured at the same occasions. It is therefore not suitable for longitudinal datasets with non-responses (Hedeker & Gibbons, 2006).

2.1.3 Mixed Effects Regression Model (MRM)

Another approach used to analyze longitudinal data is the mixed regression models (MRM). This approach can be used for both categorical, continuous and count data. MRM gives unbiased results if missing data are assumed ignorable i.e. missing completely at random and missing at random. MRM allows the measurement occasions to vary among the individuals. The method handles both time invariant and time varying variables and is therefore a suitable method to analyze longitudinal data with non responses. When dependent or outcome variable is continuous and normally distributed, the MRM is referred as the linear mixed effects regression model. The linear mixed effects regression model approach is an extension to a class of regression models called generalized linear mixed effects regression models that is often useful for outcomes such as binary, count and ordinal data. The disadvantage of MRM is that the full-likelihood methods are more computationally complex than quasi-likelihood methods (Davis, 2002; Nakai & Ke, 2011).

2.1.3.1 Linear Mixed Effects Regression Model

Repeated measurements continuous outcome can be modeled using a linear mixed effects regression model (Hedeker & Gibbons, 2006; Molenberghs & Verbeke, 2005). The procedure is detailed below using the notations of Molenberghs and Verbeke (2005) and Hedeker and Gibbons (2006).

For any repeated measurement variable, for example body mass index (BMI), which is continuous variable. Let Y_{ij} represent the jth measurement of repeated measurement variable for example body mass index for the ith subject collected at time w_{ij} for $i = 1,2, \dots, n$ and $j = 1,2, \dots, p_i$. Total number of subject interviewed is n, and p_i is number of body mass index measurements collected from subject i . Consider $Y_i^T = (Y_{i1}, Y_{i2}, \dots , Y_{ip})$, in which Y_{i1} is the first measurement for subject i. It follows that $Y_i = X_{1i}\beta_M + Q_i r_i + \varepsilon_i$ 2.1.2

The term X_{1i} is a $p\ x\ s$ design matrix, Q_i is a $p\ x\ t$ design matrix and β_M is an $s\ x\ 1$ vector containing fixed effects. Random effects were depicted by r_i , which is a $t\ x\ 1$. The term r_i follows a normal distribution, $N(0, G)$ and r_i has a mean of zero and its variance covariance matrix $G = [vbc]$. In this case $vbc = Cov(r_{ib,}r_{ic})$, ε_i is a residual error vector. An assumption that r_i is independent of ε_i can be made. The residual errors have normal distribution with mean zero and its variance covariance is V_i . Nguti et al (2005) argues that Y_i is marginally normally distributed with $X_{1i}\beta_M$ and its variance covariance is

$$F_i = Q_i G Q_i^T + V_i \qquad 2.1.3$$

The linear mixed effect regression model has random and fixed effects. In equation 2.1.2, fixed effects were represented by $X_{1i}\beta_M$ and random effects by $Q_i r_i$. Variability within subjects is taken care by the term β_M . Variability between subjects

is modeled using random effects r_i. Linear mixed regression model can have either random intercept only, random intercept and slope or quadratic.

2.1.4 Generalized linear models

Generalized linear models (GLM) are applied when analyzing univariate discrete outcome variables, via known variances and link functions. Generalized linear models have three components. These are random component, systematic component, and link between the random and systematic components (Davis, 2002). The random component identifies the response variable y and assumes a specific probability distribution for y and the probability distribution belongs to exponential family (Davis, 2002).

For longitudinal data, the GLM is not sufficient to model discrete responses because of the dependency between observations within subjects. There are 3 main extensions of generalized linear models. These include marginal models, mixed effects models and transitional models. This section describes the extensions of GLM as discussed by Davis (2002).

2.1.4.1 Marginal Models

Let y_{ij} stands for the response at time j from subject i. Marginal expectation $\mu_{ij} = E(y_{ij})$ is modeled as a function of explanatory variables. The marginal expectation is the average response over the sub population that shares a common value of the covariate vector. Note that, this is what is modeled in a cross sectional study. Associations among repeated observations are modeled separately from the

marginal mean and variance of the response vector. The assumptions can be outlined as follows:

1. The marginal expectation μ_{ij} is related to the covariates through a known link function g:

   $g(\mu_{ij}) = x'_{ij}\beta_i$, where $x'_{ij} = x_{ij1}, \ldots, x_{ijp}$ is a vector of covariates specific to subject i at time j and β is a $p \times 1$ vector of regression parameters.
2. The marginal variance of y_{ij} is related to the marginal expectation μ_{ij} through $Var(y_{ij} = \emptyset V(\mu_{ij})$, where V is a known variance function and $\emptyset$ is a possibly unknown scale parameter.
3. The covariance between y_{ij} and y'_{ij} is a known function of μ_{ij}, μ'_{ij}, and a vector of unknown parameters α.

2.1.4.2 Random Effects Models

In random effects models, heterogeneity between individuals arising from unmeasured variables is accounted for by including subject specific random effects in the model. These random effects are assumed to account for all of the within subject correlation present in the data. Conditional on the values of the random effects, the responses are assumed to be independent.

The assumptions can be outlined as follows:

1. Given a vector bi of subject-specific effects for the ith subject, the conditional mean of y_{ij} satisfies the model $g\left(E\left[y_{ij}\middle|b_i\right]\right) = x'_{ij}\beta + z'_{ij}b_i$, where g is a known link function and z_{ij} is a vector of covariates for subject i at time j.
2. $y_{i1}, \ldots, y_{iti}$ are independent given b_i for each $i = 1, \ldots, n$.
3. $b_1, \ldots, b_n$ are independent and identically distributed with probability density function f.

2.1.4.3 Transition Models

In transition models for the analysis of repeated measurements, the observations $y_{i1}, \dots, y_{iti}$ from subject i are correlated because y_{ij} is explicitly influenced by the past values $y_{i1}, \dots, y_{i,j-1}$. The past outcomes are treated as additional predictor variables. The conditional expectation of the current response, given the past responses, is assumed to follow a generalized linear model. The linear predictor component of the model includes the original covariates as well as additional covariates that are known functions of past responses.

Thus, the general form of the model is

$$g\left(E\left[y_{ij}\left|y_{i1}, \dots, y_{i,j-1}\right.\right]\right) = x'_{ij}\beta + \sum_{r=1}^{s} \alpha_1 f_r\left(y_{i1}, \dots, y_{i,j-1}; \alpha_1, \dots, \alpha_s\right) \quad 2.1.4$$

Where $f_1, \dots, f_s$ are functions of previous observations and possibly of an unknown parameter vector $\alpha = (\alpha_1, \dots, \alpha_s)$. In addition, the conditional variance of y_{ij} given the past is proportional to a known function of the conditional mean i.e.

$Var\left(y_{ij}\left|y_{i1}, \dots, y_{i,j-1}\right.\right) = \emptyset V\left(E\left[y_{ij}\left|y_{i1}, \dots, y_{i,j-1}\right.\right]\right)$, where V is a known variance function and $\emptyset$ is an unknown scale parameter (Diggle et al., 2002).

2.1.4.5 Generalized Estimating Equations (GEE)

Generalized estimation equations (GEE) were proposed by Liang and Zeger (1986) based on concept of estimating equations. Generalized estimating equations (GEE) are generalization of generalized linear models (GLM). GEE support many different types of dependent variables. The method was developed to cater for categorical and counts responses, and can also be used to analyze continuous data (Diggle et al., 2002; Hedeker & Gibbon, 2006).

Let $Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{ip})$ be a vector of correlated responses for ith subject, $i = 1,2, \dots, n$.

Marginal expectation of response, $E(Y_{ij}) = \mu_{ij}$, and this depends on explanatory variable X_{ij} through a known link function, $g(\mu_{ij}) = \eta_{ij} = X_{ij}\ \beta$. Marginal variance of Y_{ij} depends on marginal mean according to $var(Y_{ij}) = v(\mu_{ij})\phi$, where $v(\mu_{ij})$ is known and ϕ may have to be estimated. Correlation between Y_{ij} and Y_{ik} is function of some additional parameter α, may also depend on μ_{ij} and μ_{ik}.

Estimate of β can be obtained as solution to the following generalized estimating equations $\sum_{i=1}^{n} D_i' V_i^{-1}(Y_i - \mu_i) = 0$, where $D_i = \frac{d(\mu_i)}{d\beta}$ and V_i is working covariance matrix, that is $V_i = Cov(Y_i)$, D_i is a function of β, V_i is a function of both $\beta\ and\ \alpha$.

Iterative two stage estimation procedure is required for generalized estimation equation.

1. Given current estimates of $\alpha\ and\ \phi$, an estimate of β is obtained as solution to the GEE.
2. Given current estimate of , estimate of $\beta\ and\ \phi$ are obtained based on standardized residuals $r_{ij} = (Y_{ij} - \hat{\mu}_{ij})\mathrm{v}\hat{\mu}_{ij}^{\frac{1}{2}}$ 2.1.5

If estimates of $\alpha\ and\ \phi$ are consistent, then the solution of generalized estimating equations $\hat{\beta}$ has following properties:

1. $\hat{\beta}$ is consistent estimator of β.
2. In large samples $\hat{\beta}$ has a multivariate normal distribution.
3. $Cov(\hat{\beta}) = F^{-1}GF^{-1}$, where $F = \sum_{i=1}^{n} D_i' V_i^{-1} D_i$, $G = \sum_{i=1}^{n} D_i' V_i^{-1} cov(Y_i) V_i^{-1} D_i$

Marginal distribution of Y_{ij} at each time point has to be specified. The GEE treated variance-covariance structure as a nuisance. In the GEE, unobserved variables are

dependent only on the covariates, as the result of this, the missing data structure for GEE is the covariate-dependent MCAR. Therefore GEE does not automatically provide unbiased estimates of parameters when data is missing at random (MAR). Some weighting will need to be done to obtain unbiased estimates (Lipsitz & Fitzmaurice, 2009). This is one of the shortfalls of GEE when there are some missing data.

The term marginal model refers to models for longitudinal data, which have random effects (Fitzmaurice et al., 2009). Specification of a GEE is similar to a GLM with a linear predictor, a link function and variance described as a function of the mean. An additional feature of GEE is the working correlation structure R, $n\,x\,n$ correlation matrix common for all subjects. It is important that choice of working correlation matrix should be consistent with the observed correlation matrix. However, choice of the correlation structure for the repeated measurements is not critical for GEE. This is because GEE provides estimated parameters and standard errors that are robust to misclassification of the variance covariance structure. The important thing is that the univariate analysis models at each time point should be specified correctly. GEE should be applied when the research interest is mainly on estimates and inference of the regression parameters, but is not suitable when modeling variance-covariance structures of longitudinal data.

In the frame work of GEE, there are two general approaches used to handle missing data. The first approach is to analyze multiple imputed data by generalized estimating equations. The second approach is the use of weighted estimating equations. This

approach is suitable when the missing data pattern is monotone, as a result of dropout. See Fitzmaurice et al., (2009) for more detail.

2.1.5 Covariance Patten Model (CPM)

Covariate Pattern model (CPM) can be regarded as an extension of MANOVA. CPM does not distinguish between subjects and within-subjects variance. CPM was first described by Jennrich and Schluchter (1986) (Hedeker & Gibbon, 2006). The regression model for CPM in matrix form can be written as $y_i = X_i\beta_i + e_i$, with $i = 1,2, \dots N$ for n individuals and $j = 1,2, \dots n_i$ observations for i individuals, y_i is an $n_i\ x\ 1$ vector for subject, β is a $p\ x\ 1$ vector of fixed regression parameters, the vector e_i is assumed to be normally distributed with zero mean and variance-covariance Σ_i, (Hedeker & Gibbon, 2006). CPM assumed that timing of measurements is fixed, this means that subjects are intended to be measured at the same finite number of occasions. CPM allows that individuals may have incomplete data.

There are different covariance patterns for covariance pattern model (CPM). These covariance patterns include independent covariance structure, exchangeable covariance structure, first order autoregressive structure, Toeplitz structure and unstructured form.

2.1.6 Missing data Mechanisms

Missing data are common in longitudinal data. One of the reasons is that studies take long and some of participants may drop or may be lost to follow up before an endpoint of interest is measured. This section discusses the classification of missing data and the mechanisms used to handle missing data.

2.1.6.1 Classification of Missing Data

Rubin (1976) has classified missing data in the following ways: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Let R_{ij} stand for an indicator variable taking value 1 if an individual i is observed at time j and 0 when an individual was not observed at time j. If subjects were supposed to be measured at n time points, then $n\ x\ 1$ complete dependent vector is $Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{in})'$. And Y_i is an $n\ x\ 1$ matrix of covariates X_i.

The $n\ x\ 1$ missing data indicator vector is $R'_i = R_{i1}, R_{i2}, \dots, R_{in}$, with $R_{ij} = 1$ if Y_{ij} is observed and $R_{ij} = 0$ if Y_{ij} is missing. Further divide the complete data variable vector Y into observed Y_i^o, and unobserved Y_i^m.

Rubin (1976) defined the terms as follows: data is said to be missing completely at random (MCAR) if the missing data occur totally at random. The missing data is not related to other observed or unobserved data. This is the most basic missing data mechanism and assumes missing data to occur for completely random reasons. The distribution of missing values R is thus assumed to be independent of both covariates and the dependent variable as $\mathrm{P}(R_i|Y_i^o, Y_i^m, \mathrm{X_i}) = \mathrm{P(R_i)}$ (Rubin, 1976). Shaffer (1997) presented a good summary on missing data mechanism.

The second mechanism of missing data to be discussed is missing at random (MAR). Data is said to be missing at random (MAR) if probability that responses or observations are missing depends on the set of observed responses, but is not related to the specific missing values that would have been obtained if there were no

missing data. MAR can be written as $\mathrm{P}(R_i|Y_i^o, Y_i^m, \mathrm{X_i}) = \mathrm{P}(R_i|Y_i^o, \mathrm{X_i})$(Rubin, 1976; Molenberghs & Fitzmaurice, 2009).

Missing completely at random (MCAR) and missing at random (MAR) are referred to as ignorable mechanisms. In order to obtain valid likelihood based estimates in the presence of some missing data, the data have to meet the following two conditions. The first condition is that the missing data should be missing at random (MAR). Secondly the parameters defining missing data process should not be related to the parameters to be estimated.

Finally, not missing at random (NMAR) is defined as the probability of missing responses, which depends on both the set of observed responses and the specific missing values that should have been obtained if there were no missing data. That is conditional distribution of R_i given Y_i^o, is related to Y_i^m and $\mathrm{P}(R_i|Y_i^o, Y_i^m, \mathrm{X_i})$ depends on at least some components of Y_i^m (Rubin, 1976; Molenberghs & Fitzmaurice, 2009). When data are missing not at random (MNAR), they are called non ignorable missing data.

2.1.7 Handling Missing Data

There are different approaches used to handle missing data. Some of these approaches include complete case analysis, last value carried forward, imputation methods, expectation maximization (EM), selection models, and pattern mixture models. This section discusses some of these approaches.

2.1.7.1 Complete Case Analysis

In the complete case analysis method, only subjects without missing observations are included in the analysis. That is, subjects with incomplete observations are discarded in the analysis of data. Advantage of complete case analysis method is that it can be used for any kind of statistical analysis, however it gives unbiased estimate of mean response trends only when the missingness is missing completely at random (MCAR) (Nakai & Ke, 2011). In the complete case analysis, the amount of data is reduced and this leads to the reduction statistical power (Diggle et al., 2002). When data is not completely missing at random, complete case analysis may give biased results. Carpenter and Kenward (2007) recommended that complete case analysis should not be used to address the problem of missing data.

2.1.7.2 Last observation carried forward (LOCF)

Last observation carried forward (LOCF) imputes values for missing data based on the last previous observed value. This method is usually used in the longitudinal data, in which data are observed or collected at several occasions. It imputes values equal to the last observed response for the variable for each unit (Diggle et al 2002). The disadvantage of LOCF method is that it may give biased results when the missing data is not missing at random. As the result, Carpenter and Kenward (2007) argued that this method should not be used when imputing missing data.

2.1.7.3 Expectation maximization (EM)

Expectation maximization (EM) is another method used to produce estimates of coefficients during data analysis. This method is based on Bayesian thinking. EM was

introduced by Dempster et al. (1977). EM uses maximum likelihood estimation (ML) to produce parameter estimates.

Likelihood methods handle problem of missing data by modeling and estimating parameters of joint distribution of Y_i, $f(Y_i|X_i,\gamma)$ (Molenberghs &Fitzmaurice, 2009). Maximum likelihood estimator (MLE) can be obtained by maximizing $f({Y_i^o}_i|X_i,\gamma)$. In this method, missing values are predicted by using observed data and the model of conditioned mean $E({Y_i^m}_i|Y_i^o,X_i,\gamma)$ (Molenberghs & Fitzmaurice, 2009).

EM works in two iterative stages. First stage is called expectation stage (E-step) and second stage is known as maximizing stage (M -step). Let θ^t be current estimate of parameter θ, then $W(\theta^t|\theta) = \int g(\theta|Y_i) f\left({Y_i^m}_i\middle|Y_i^o\right),\ (\theta^t = \theta) dY_i^m$, where $g(\theta|Y_i)$ is complete data log likelihood. M-step gets parameter estimates to maximize complete log likelihood from E-step. $W(\theta^{t+1}|\theta^t) \geq W(\theta^t|\theta)\ for\ all\ \theta$. E and M steps are iterated until iteration converges. The method assumes a large number of data so that the EM estimates can be approximately unbiased and normally distributed. In addition it assumes data to be ignorable, that is MCAR or MAR mechanism (Molenberghs & Fitzmaurice 2009).

.

2.1.7.4 Multiple imputation

Multiple imputation (MI) is another approach used to handle missing data. Multiple imputation produces M different datasets, in which each could have been the complete dataset if all values were observed. These M complete datasets are combined to obtain estimates and standard errors that reflect uncertainty in the missing data and the finite sample variation (Rubin, 1987). Multiple imputation

method is Bayesian based. MI involves 3 different stages; namely missing values are filled M times to generate M complete datasets; each of the M complete datasets is analyzed by using standard, compete, procedures; and results from the M analyses are combined to produce a single MI estimator and to draw inferences using Rubin's rule (Rubin, 1987).

In MI, missing data are substituted by their corresponding imputation samples, producing M completed data sets. Using the notation of Kenward and Carpenter (2009), let β_k be estimate of β and $V_{\boldsymbol{k}}$ be covariance matrix from the kth completed data set ($k = 1, \ldots, M$). The MI estimate of β is the simple average of the estimates

$$\hat{\beta}_{MI} = \frac{1}{M}\sum_{i=1}^{M}\tilde{\beta}_k \qquad 2.1.6$$

Rubin (1987) provides the following expression for the covariance matrix of $\hat{\beta}_{MI}$ that can be applied very generally and uses only complete-data quantities. Define

$W = \frac{1}{M}\sum_{i=1}^{M} V_k$ to be the average within-imputation covariance matrix, and

$B = \frac{1}{M-1}\sum_{i=1}^{M}(\tilde{\beta}_k - \hat{\beta}_{MI})(\tilde{\beta}_k - \hat{\beta}_{MI})'$ to be the between-imputation covariance matrix of β_k. Then, an estimate of the covariance matrix of β is given by

$$\hat{V}_{MI} = W + \left(\frac{M+1}{M}\right)B \qquad 2.1.7$$

Tests and confidence intervals are based on the approximate pivot $P = \left(\hat{\beta}_{MI} - \beta\right)V_{MI}^{-1}\left(\hat{\beta}_{MI} - \beta\right)$ (Rubin, 1987).

Multiple imputation produces unbiased estimates and variance if the data is missing completely at random (MCAR) or missing at random (MAR) (Little & Rubin, 2002; Diggle et al., 2002; Hedeker & Gibbon, 2006).

2.1.8 Missing Data Not At Random

The approaches described above make assumption that the data missing mechanism is MCAR or MAR, but in some cases the data may be missing not at random (MNAR). To handle MNAR, the missing data distribution must be taken into consideration when imputing the unobserved values. When handling data, which is not missing at random selection and pattern mixture models can be used.

2.1.8.1 Selection models

Selection model is made up of distributions for the complete data and missing data given the data itself. Therefore the joint distribution of the complete data Y_i and the missing data distribution R_i through models for marginal distribution of Y_i and the conditional of R_i given Y_i can be written as

$f(R_i\ Y_i|X_i\ ,\gamma,\phi) = f_Y(X_i,\gamma)f_{R|Y}(R_i|X_i,\gamma,\phi)$, where $\theta = (\gamma,\phi)$ (Little, 2009).

Selection models are extremely sensitive to the distributional shape that is chosen for the population (Schafer & Graham, 2002).

2.1.8.2 Pattern Mixture Models

Another approach used to model non ignorable missing data is pattern mixture model. This model groups the whole sample on basis of the missing data distribution. Pattern mixture models specify marginal distribution of R_i and conditional distribution of Y_i given R_i can be written as $f(R_i\ Y_i|X_i\ ,v,\delta) = f_R(R_i|X_i,\delta)f_{Y|R}(Y_i|X_i,R_i,v)$ 2.1.8. In equation 2.1.8, $\theta = (v,\delta)$. Unlike selection models, pattern mixture models are not sensitive to distribution of the population (Schafer & Graham, 2002).

2.2 Statistical Models for Survival Data

2.2.1 Introduction

There are many techniques used to analyze survival data. This section describes some of the techniques used to analyze survival data.

2.2.1.1 Survival Time Distribution

Using notation similar to that of Kalbfleisch and Prentince (2002), suppose T_i denote survival time of an ith individual ($i = 1,2,3, \ldots . n$) , that is taken as the minimum of true event time t*. The data that can be observed are $\{t_i, \delta_i\}, for\ i = 1,2,3, \ldots . n$ where $t_i = min(t_i^*, c_i)$. Further denote $\delta_i = I(t_i^* < c_i)$, where c_i is censoring time for the ith individual. Then T_i can be regarded as a random variable.

Let T be a non negative random variable representing the survival time. Survival time distribution can be described by one of the following three functions; survival function, hazard function, and probability density function. The definitions presented in this section are based on a book written by Kalbfleisch and Prentince (2002).

It has to be noted that survival function is defined by both discrete and continuous T. Both probability density and hazard functions are available for discrete and continuous T. Survival function $S(t)$ is defined for both discrete and continuous distributions as the probability that the survival time is greater than t. That is $S(t) = P(T > t), 0 < t < \infty$.

2.2.1.2 T discrete

If T is a discrete random variable taking ordered values $t_1 < t_2 < \cdots$, with associated probability function $f(t_i) = P(T = t_i),\ i = 1,2,3,\ldots$ then the survival function is expressed as $S(t) = \sum_{j\,|\,t_i > t} f(t_j)$.

The hazard function $h(t)$ is defined as the conditional probability of failure at time t_j given that the individual has survived up to time t_j.

$$h(t_j) = P(T = t_j \,|\, T \geq t_j) = \frac{f(t_j)}{S(t_j)}$$

$$= 1 - \frac{S(t_{j+1})}{S(t_j)} \qquad 2.2.1$$

2.2.1.3 T absolute continuous

If T is absolute continuous variable, then the probability density function of T is expressed as $f(t) = F'(t) = -S'(t), for\ t \geq 0$.

Hazard function gives instantaneous failure rate at t given that the subject has survived up to time t, mathematically; the hazard function is given by

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \,|\, T \geq t)}{\Delta t}\ . \qquad 2.2.2$$

Survival function and hazard function are related and their relationship is given by the following formula $h(t) = \frac{f(t)}{S(t)} = \frac{-dlogS\,(t)}{dt}$, where

$$S(t) = \exp[-\int_0^t h(u)du] = exp\,(-H(t))for\ t \geq 0.$$

$H(t) = \int_0^t h(u)du$ is called cumulative hazard function that can be obtained from survival function because of the following relationship $H(t) = -logS(t)$.

Probability density function for T may be written as

$$f(t) = h(t)\exp[-\int_0^t h(u)du] \ . \qquad 2.2.3$$

The remaining part of this section reviews some of the methods, which are used to analyze survival data.

2.2.2 The Kaplan Meier estimate of Survival Function

Kaplan-Meir estimate of survival function (Kaplan & Meier, 1958) is the estimator used by most statistical software packages (Hosmer & Lemeshow, 1999). The estimator uses information from all observations, both censored and uncensored, and considers survival at any point in time as series of steps defined by observed survival and censored time (Hosmer & Lemeshow, 1999; Collet, 2003).

Suppose that k individuals have experienced an event of interest, such as death in group of individuals. If we let $0 \leq t_{(1)} < \cdots < t_{(k)} < \infty$ be the observed events of interest i.e. death, ordered according to times the event of interest has occurred. Let k_j be the number of individuals who are at risk at $t_{(k)}$. The individuals at risk can be defined as the individuals who are alive and not censored just before $t_{(j)}$.

Furthermore let d_j be the number of events of interest (deaths) that have been observed at $t_{(j)}, j = 1, 2, \ldots, k$. Then Kaplan Meier estimator of $S(t)$ is defined by

$$\hat{S}(t) = \prod_{j:\, t_{(j)} < t} 1 - \frac{d_j}{k_j} \qquad 2.2.4$$

It should be noted that Kaplan Meier estimator changes its value when a death has happened. This estimator has discrete distribution (Hosmer & Lemeshow, 1999; Collet, 2003). Confidence intervals may be calculated by using Greenwood's formula, which was developed by Greenwood in 1922 (Hosmer & Lemeshow, 1999).

Comparison of two survival distributions can be done by using Log Rank Test. Non parametric methods such as Kaplan Meier are suited for the homogeneous samples, and their shortfall is that they cannot determine if variables or covariates are related to the survival times (Machin et al., 2006), thus, it is often difficult to control for potential confounders using these methods.

2.2.3 Cox Proportional Hazard Model

Cox proportion hazard model (Cox, 1972) has been widely used in analyzing survival data (Cox & Oakes, 1984; Hosmer & Lemeshow, 1999; Collet, 2003). The Cox proportional hazard model is given by the following:

$$h(t|z) = h_0(t)exp(\beta_i z_i + \beta_2 z_2 + \cdots + \beta_p z_p). \qquad 2.2.5$$

In the equation 2.2.1, z is explanatory vector, which does not change over time for any individual, $(\beta_i + \beta_2 + \cdots + \beta_p)$ is vector of regression coefficients and $h_0(t)$ is baseline hazard function. The hazard ratio (HR) for two individual with covariates values denoted z_1 and z_0 is expressed as

$$HR = \frac{h_0(t)\exp\left(z_1(\beta_i+\beta_2+\cdots+\beta_p)\right)}{h_0(t)\exp\left(z_0(\beta_i+\beta_2+\cdots+\beta_p)\right)}. \qquad 2.2.6$$

Cox proportion hazard model is time independent. The advantage of Cox model is that interpretation is easy and similar to that of the relative risk ratio (Hosmer & Lemeshow, 1999; Collet, 2003; Kalbfleish & Prentice, 2002).

2.2.4 Stratified Cox Model

Stratified Cox model stratifies predictors, which are not satisfying the proportional hazard assumptions (Hosmer & Lemeshow, 1999). Once the predictor has been identified, the data are grouped into subgroups, and then the Cox model is performed in each subgroup. Hosmer and Lemeshow (1999) describes stratified Cox model in

Chapter 7 of their book. The model is given by $h_{ik}(t) = h_{ok}(t)\exp(\beta^T X_{ik})$, where $k = 1, 2, \ldots . S$ represents the subgroup or stratum. The hazards for this model are non-proportional because the baseline hazards may be different between subgroups or strata. The coefficients β are assumed to be the same for each subgroup or stratum k. The partial likelihood function is obtained by multiplying partial likelihood for each stratum. The problem with this approach is that the effects of stratified predictors cannot be identified.

2.2.5 Cox Model with Time Dependent Variables

Sometimes values of covariates may change over time t. If this scenario arises, Cox proportional with time independent covariate may not be appropriate approach to use. This is because the proportion hazard makes assumption that the effects of any covariate in the model does not change at any point in time. Therefore this type of data, where covariates are changing with time, can be modeled using Cox proportion model with time dependent covariates. Cox model with time dependent variables has been discussed by Cox and Oakes (1984), Hosmer and Lemeshow (1999) and Collet (2003). In order to model time dependent effects $X(t)$, then $\beta X(t) = \beta(X)\, x\, p(t)$. Where $p(t)$ is function of time t.

If survival data has both time independent X_i and time dependent covariates $X_i(t)$. The Cox proportion model can be written as

$$h\big(t\big|x(t)\big) = h_o(t)\exp\big[\textstyle\sum_{i=1}^{k1}\beta_i\, x_i + \sum_{j=1}^{k2}\alpha_j\, x_j(t)\big] \qquad 2.2.7$$

At any time t, hazard ratio (HR) for two individuals with different covariates x and x' is given by

$$\hat{HR}(t) = \exp[\sum_{i=1}^{k1}\hat{\beta}(x_i^{'} - x_i) + \sum_{j=1}^{k2}\hat{\alpha}_j(x_j^{'}(t) - x_j(t))]$$

In the above formula the coefficient $\hat{\alpha}_j$ is not time-dependent. The term $\hat{\alpha}_j$ represents overall effect of $X_j(t)$ at all the times that covariate has been measured in the study.

Time dependent variables can be classified either as internal or external. An internal time-dependent variable is any variable, which can change the value of covariate over time and is related to the characteristics of the individual. For example hemoglobin level, blood pressure, body mass index and CD4 count. External time dependent variable is a variable whose value at a particular time does not require subjects to be under direct observations, that is, values change because of external characteristics to the individuals ,for example level of environmental degradation.

2.2.6 Parametric Proportional Hazards Models

Parametric proportional hazards models are also used to analyze survival data. They have got the same form as Cox proportional models. Hazard function at time t for a particular patient with a set of z covariates $(x_1, x_2, \dots, x_z)$ is given as

$$h(t|x) = h_o(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_z x_z) . \qquad 2.2.8$$

The distribution of $h_o(t)$ has to be specified in parametric proportion hazard models. The commonly applied parametric proportional hazards models are exponential, Weibull and Gompertz models. Weibull models are mostly used (Collet, 2003).

2.2.6.1 Weibull Proportional Hazard model

If survival times follow Weibull distribution with scale parameter λ and shape parameter γ such that survival and hazard function of a Weibull distribution are given by $S(t) = \exp -(\lambda t^{\gamma})$ and $h(t) = \lambda(t)^{\gamma-1}$ respectively. Both λ and γ are greater

than zero. When $\lambda > 1$ hazard rate increases and decreases when $\lambda < 1$. When $\gamma = 1$, the hazard rate does not change, i.e. it remains constant. When $\gamma = 1$, Weibull distribution reduces to exponential distribution.

Hazard function for Weibull proportional hazard model for a particular individual with z covariates $(x_1, x_2, \dots, x_z)$ is written as

$$h(t|x) = \lambda(t)^{\gamma-1} \exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_z x_z). \quad 2.2.9$$

2.2.6.2 Exponential Proportional Hazard Model

As already stated exponential proportional hazard model is a special case of Weibull model when γ = 1. The hazard function for exponential proportional hazard model is constant over time. Survival function is given as $S(t) = \exp - (\lambda t)$. Hazard function is expressed as $h(t) = \lambda$. Under the exponential proportion hazard models, the hazard function for exponential proportional hazard model for an individual is given by

$$h(t|x) = \lambda \exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_z x_z). \quad 2.2.10$$

2.2.6.3 Gompertz Proportional Hazard Model

Survival function of the Gompertz distribution is expressed as

$S(t) = \exp(\frac{\lambda}{\theta}(1 - e^{\theta t}))$ for $\theta \leq t \leq \infty$ and $\lambda > 0$. Its hazard function is expressed as $h(t) = \lambda \exp(\theta t)$, for $\theta \leq t \leq \infty$ and $\lambda > 0$. Parameter θ gives the shape of the hazard function. Gompertz hazard decreases or increases monotonically. The hazard function of an individual is expressed as

$$h(t|x) = \lambda \exp(\theta t) \exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_z x_z). \quad 2.2.11$$

2.2.7 Accelerated failure time (AFT) model

Parametric PH models are widely applicable in analyzing survival data. PH models assume a constant hazard over time. In practice, however the hazard function may not necessarily be a constant. It may either accelerate or decelerate over time. Some parametric models are available to handle these situations. There are relatively few probability distributions for the survival time that can be used with these models. In these situations, the accelerated failure time model (AFT) is an alternative to the PH model for the analysis of survival time data (Collet, 2003). These AFT models include log-logistic and log-normal AFT models. AFT models measure direct effect of the explanatory variables on the survival time instead of hazard. The accelerated failure time (AFT) has been described by Collet (2003) and Hosmer and Lemeshow (1999).

For a group of patients with covariate $X_1, X_2, \dots, X_p$, the model is written as $S(t|x) = S_0(t|\tau(x))$, where $S_0(t)$ is the baseline survival function, τ is an acceleration factor. Acceleration factor is ratio of survival times that correspond to any fixed value of $S(t)$. The acceleration factor is expressed in the formula below

$$\tau(x) = \exp(\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_p x_p). \qquad 2.2.12$$

In the accelerated failure time model, the covariate effects are thought to be constant and multiplicative on the time scale. Considering the relationship of survival and hazard function, the hazard function for an individual with covariate $X_1, X_2, \dots, X_p$ is expressed as $h(t|x) = \frac{1}{\tau(x)} h_0 t|\tau(x)$. The corresponding log-linear form of the accelerated failure time (AFT) model with respect to time is given by

$log\ Ti = u + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi} + \sigma\epsilon_i$, where u is intercept, α is scale parameter and $\sigma\epsilon_i$ is a random variable that is assumed to have a certain distribution.

2.2.7.1 Estimation of AFT model

AFT models use maximum likelihood method. The likelihood of n observed survival times, $t_1, t_2, \dots, t_n$ is given by $L(\alpha, \mu, \sigma) = \prod_{i=1}^{n}\{f_i(t_i)\}^{\delta_i}\{S_i(t_i)\}^{1-\delta_i}$, where $f_i(t_i)$ and $S_i(t_i)$ are the density and survival functions for the ith subject at t_i, and δ_i is an event indicator for the ith observation. The log-likelihood function is written as $L(\alpha, \mu, \sigma) = \sum_{i=1}^{n}\{\delta_i log\sigma t_i + \delta_i log f_{\epsilon i}(z_i) + (1-\delta_i) log S_{\epsilon i}(z_i,)\}$, where

$z_i = (\log(t_i) - \mu - \alpha_i x_{1i} - \alpha_i x_{2i} \dots - \alpha_i x_{pi})/\sigma$.

2.2.7.2 Log-logistic AFT model

The description of log logistic AFT model in this section was taken from Collet (2003). The survival and hazard function of log-logistic are given by

$S(t) = \frac{1}{1+e^{\theta}t^{k}}$, $h(t) = \frac{e^{\theta}kt^{k}}{1+e^{\theta}t^{k}}$, where θ and k are unknown parameters and $k > 0$. When $k \leq 1$, the hazard rate decreases monotonically and when $k > 1$ hazard rate increases from zero to a maximum value and then decreases to zero.

If survival times have a log logistic distribution with parameter θ and k, then the hazard function for the ith subject can be written as $h_i(t) = \frac{e^{\theta - klog\,\tau_i}kt^{k-1}}{1+e^{\theta - klog\,\tau_i}t^{k}}$,

where $\tau_i = exp(\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_p x_p)$ for individual i with p independent variables. If the baseline survival function is $S_0(t) = \frac{1}{1+e^{\theta}t^{k}}$, where θ and k are unknown parameters, then the baseline odds of surviving beyond time t are given by $\frac{S_0(t)}{1-S_0(t)} = \frac{1}{e^{\theta}t^{k}}$. The survival time for the ith individual also follows log logistic distribution, which is $S_i(t) = \frac{1}{1+e^{\theta - klog\,\tau_i}t^{k}}$. Therefore the odds of the ith individual surviving beyond time t is given by $\frac{S_i(t)}{1-S_i(t)} = \frac{1}{e^{log\,\tau_i - \theta}t^{-k}}$.

In a two group study the log (odds) of the ith individual surviving beyond time t are

$log\frac{S_i(t)}{1-S_i(t)} = \beta x_i klogt$, where x_1 is the value of a categorical variable, which can take a value of 1 in one group and 0 in the other group. If T_i has a log-logistic distribution, then ε_i has a logistic distribution. Therefore the survival function of logistic distribution is given by $S_{\varepsilon_i}(\varepsilon) = \frac{1}{1+\exp(\varepsilon)}$.

2.2.7.3 Log-normal AFT model

The description of log normal AFT model in this section was taken from Collet (2003). When the survival times are assumed to follow a log-normal distribution, baseline survival function is written as $S_0(t) = 1 - \varphi(\frac{logt-u}{\sigma})$ and the hazard function is given by $h_0(t) = \frac{\vartheta(\frac{logt}{\sigma})}{1-\varphi(\frac{logt}{\sigma})\sigma t}$, where u and σ are parameters, $\vartheta(x)$ is probability density function and $\varphi(x)$ is cumulative density function of the standard normal distribution. The survival function for the ith individual is $S_i(t) = S_i(t|\tau_i) = 1 - \varphi(\frac{logt-\alpha' x_i-u}{\sigma})$,

where $\tau_i = \exp(\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_p x_p)$. Therefore log survival time for the ith individual has normal $\mu + \alpha' x_i, \sigma$.

In a two group study, one can easily get $\varphi^{-1}[1 - s(t)] = 1/\sigma(logt - \alpha' x_i - u)$

where x_i is value of a categorical variable is 1 in one group and 0 in other group.

2.3 Joint Models for Longitudinal Measurements and Survival Data

Joint models of longitudinal measurements and event to time are commonly used especially when there is association between the two processes. Mainly joint models have been used in studies addressing AIDS, cancer issues and quality of life (Lim et al., 2013; Sweeting & Thomson, 2011; Rizopoulos, 2010). There are approaches that

are used to model survival and longitudinal data together. These approaches are two stage and likelihood based. This section reviews literature for the joint modeling methods.

Early discoveries in joint modeling focused on longitudinal data. The earlier work utilized longitudinal data model of the following form:

$$X_i(t) = f(t)^T \alpha_i \qquad 2.3.1$$

In the equation 2.3.1, $f(t)$ is vector of functions of t, and α_i is the linear function. Schluchter (1992) developed a log normal survival model. In fact the model developed was an extension of equation 2.1.2. The model was modeled in two stages. The author maintained regressions for individual subject during the first stage. In the second stage, an assumption that log survival time, true slope and intercept follow trivariate normal distribution. Maximum likelihood estimates were calculated using EM algorithm. Other authors who proposed a model in a two stage approach were Hogan and Laird (1997). Hogan and Laird (1997) modeled longitudinal observed response and survival time using the linear mixed effects model.

Pawitan and Self (1993) used joint models in which times to event were modeled using a parametric approach. Tsiatis et al. (1995) and Raboud et al. (1993) adopted the use of Cox proportion hazard model in modeling survival times. The Cox models were of the form:

$$h_i(t) = h_0(t)\exp(\gamma X_i(t) + \eta^T Z_i) \qquad 2.3.2.$$

In the equation 2.3.2, $X_i(t)$ is considered to be time dependent variable. In order to assess association, γ and η are estimated. Among the authors who used this approach are include Raboud et al. (1993) who focused on potential bias because of the use of

last value carried forward approach (LVCF) in providing missing data and its failure to account for measurement error. Raboud et al. (1993) concluded that their approach reduced bias as compared to naive approaches.

Self and Pawitan (1992) proposed another approach where inference was made on sub models for survival time and longitudinal model. The longitudinal model was similar to equation 2.3.1, while the hazard model was similar to equation 2.3.2. Self and Pawitan (1992) replaced $\gamma X_i(t)$ in equation 2.3.2 with $1 + \gamma X_i(t$). They further used a two stage approach to calculate estimates.

Tsiatis et al. (1995) used a two stage approach in their work. They used the linear mixed effect model to model longitudinal data sub model and Cox proportion hazard sub model to analyze event time data. They maximized partial likelihood in order to produce estimates. The two stage approach was further investigated by Bycott and Taylor (1998), Dafni and Tsiatis (1998) and Tsiatis, DeGruttola and Wulfsohn (1995). They all concluded that this approach reduce bias for the estimates γ and η shown in equation 2.3.2.

Methods based on likelihood were also investigated. These methods were based on specification of likelihood function for parameters in equation 2.3.2 and

$$Y_i(t_{ij}) = X_i(t_{ij}) + e_i(t_{ij}). \qquad 2.3.3$$

In equation 2.3.3, Y_i is observed longitudinal data, $e_i(t_{ij})$ is measurement error and is distributed normally and has zero mean and its variance is σ^2.

DeGruttola and Tu (1994) used a longitudinal data model shown in equation 2.3.1 and transformed event time i.e. lognormal model for survival time model. They used EM algorithm to maximize the log-likelihood. Wulfsohn and Tsiatis (1997) introduced the use Cox proportional hazards model and longitudinal variables when analyzing this kind of data. They used random effects for the longitudinal process. In this model, estimates that maximize joint likelihood of survival and longitudinal processes are calculated using EM algorithm. The model that was proposed by Wulfsohn and Tsiatis (1997) was extended by Zeng and Cai (2005) to include the covariates in the linear mixed effects random model in equation 2.3.1 in the longitudinal data. For the survival data, they used multiplicative hazard models. The relationship between the survival time and longitudinal processes is linked to the random effects.

Xu and Zeger (2001) introduced another concept in which general latent variable model was used to analyze for survival and longitudinal data simultaneously.
Henderson et al. (2000) proposed model for modeling the longitudinal and survival data. They linked survival and longitudinal data by using the latent stochastic process. Parameters were estimated using EM algorithm in which Gaussian Hermitte numerical integration was used during the E-step.

Lin et al. (2002) proposed latent class models for analyzing longitudinal and event to time data. If observed longitudinal trajectories depict heterogeneity in the observed longitudinal trajectories, the linear mixed effects models cannot fully measure the covariates. The latent class model provides way to handle additional heterogeneity to uncover distinct subpopulation (Song, 2013).

Faucett and Thomas (1996) modeled longitudinal data and event time by using Bayesian approach. Faucett and Thomas (1996) used Gibbs sampling approach to estimate the parameters. Xu and Zeger (2001) generalized this approach presented in model 2.3.4.

$$X_i(t) = f(t)^T \alpha_i + U_i(t) \qquad 2.3.4.$$

In the equation 2.3.4, $U_i(t)$ is stochastic process with zero mean. Wang and Taylor (2001) decided to include longitudinal model similar the one shown in equation 2.3.4 in the Bayesian framework. They used MCMC to analyze their data. Brown and Ibrahim (2003a) developed a semi-parametric Bayesian approach of the form of equation 2.1.2. Brown and Ibrahim (2003b) developed a method for analyzing survival time and longitudinal data, when a fraction of study participants has been cured. Also Law et al. (2002) proposed a method for analyzing longitudinal and survival time data when a fraction of study participants has been cured. There was a further development in 2008. Ye and Taylor (2008) proposed a joint model with a linear growth curve model with random intercept and slope for the longitudinal variable measurements.

The disadvantage of likelihood approach is that it is computational complex (Tsiatis & Davidian, 2004). Because of this problem Tsiatis and Davidian (2001) proposed a method, which is simple to implement. Basing on equations 2.3.1 and 2.3.2 $\gamma \; and \; \eta$ can be estimated easily. This approach uses conditional score. The concept of conditional score was pioneered by Stefanski and Carroll (1987) in order to analyze generalized linear models with a measurement error. The conditional score works by "treating α_i as nuisance parameters and conditioning on an appropriate sufficient statistics" (Tsiatis & Davidian, 2004).

2.4 Shared Random Effects Models

The Shared Random Effects Model (SREM) is direct extension of the idea of survival model with time dependent covariates by considering as covariates some characteristics of the mixed model defined for the longitudinal data (Faucett & Thomas, 1996; Henderson et al., 2000; Ibrahim et al., 2010; Wu et al., 2012). The characteristics are functions of the individual random effects of the mixed model that capture the individual deviations to the mean trajectory longitudinal data. Joint models that use shared random effects have two sub models namely longitudinal data sub models and survival data sub model.

2.4.1 Longitudinal data Sub model

For the Longitudinal sub model, assume that the repeated measures $Y_i(t_{ij})$ are the measures of true unobserved value for $j = 1,2, \dots, n_i$. The mean change overtime of $Y_i^*(t_{ij})$ can be modeled by taking into account the correlation within the repeated measures of a same subject.

$$Y_i(t_{ij}) = Y_i^*(t_{ij}) + \epsilon_i(t_{ij}) \ = X_{Li}(t_{ij})^T\beta + Z_i(t_{ij})^T b_i + \epsilon_i(t_{ij}) \qquad 2.4.1$$

Where $X_{Li}(t_{ij})$ and $Z_i(t_{ij})$ are p vector and q vector of time dependent covariates associated with the p vector of fixed effects β and q vector of Gaussian random effects b_i with mean 0 and variance-covariance matrix β. The design matrices X_{Li} and Z_i will be used for the row vectors $X_{Li}(t_{ij})^T$ and $Z_i(t_{ij})^T$ respectively for $j = 1,2, \dots, n_i$. In equation 2.4.1, the fixed part of $X_{Li}\beta$ represents the mean trajectory of the repeated measurements over time, while $Z_i b_i$ defines the individual deviation with relative to the mean trajectory. The vector of measurements is $\epsilon_i = (\epsilon_i(t_{i1}), .., \epsilon_i(t_{in_i}))$. Further assume that ϵ_i is independent and follow

multivariate Gaussian distribution of mean 0 and diagonal variance-covariance matrix $\Sigma_i = \sigma^2 I_{n_i}$, ϵ_i and b_i are independent.

2.4.2 Survival Sub model

The risk of event can be modeled using any survival model but proportional hazard models are mostly considered and defined as follows:

$$h_i t(X_{Si}, b_i) = h_0(t) e^{X_{Si}^T \gamma + h(b_i,t)^T \eta} \qquad 2.4.2$$

where $h_0(t)$ is the hazard function for baseline and γ defines association between p vector of covariates of X_{Si} (that can be time dependent) and the survival time. The function $h(b_i, t)$ is a multivariate function of the random effects b_i defined in (2.4.1) and is associated with the vector of parameter η. The association between the longitudinal and survival processes is measured by coefficients η, and $h(b_i, t)$ defines the nature of the dependence between the two processes.

2.4.3 Maximum Likelihood Estimation

Shared random effect models (SREM) can be estimated by considering the joint likelihood from the longitudinal and survival sub models.

Let θ be the whole vector of parameters defined in (2.4.1) and (2.4.2). The log likelihood of the observed data can be written as:

$$l(\theta) = \log[\prod_{i}^{N} (\int_{b_i} f_Y(Y_i | X_{Li}, b_i; \theta) f_T(T_i | X_{si}, b_i; \theta) f_b(b_i; \theta) db_i)]$$

$$l(\theta) = \sum_{i=1}^{N} \log\left(\int_{b_i} f_Y(Y_i (Y_i | X_{Li}, b_i; \theta) h_i (T_i | X_{si}, b_i; \theta)^{E_i} S_i(T_i | X_{si}, b_i; \theta) f_b(b_i; \theta) db_i\right) \quad 2.4.3$$

Where f_b and f_Y are multivariate Gaussian density functions of b and Y with respectively mean 0 and $X_{Li}\beta + Z_i b_i$, and variance-covariance matrix B and Σ_i,

$h_i(T_i|X_{si}, b_i; \theta)$ is the hazard function defined in (2.4.2) and recorded at the observed time. $S_i(T_i|X_{si}, b_i; \theta) = e^{-\int_0^{S_i} h_i(T_i|X_{si}, b_i; \theta)dt}$ is the derived survival function.

The maximum likelihood estimates can be calculated by iterative algorithms such as EM or Newton-Raphson algorithm (Rizopoulos, 2010). Zeng and Cai (2005) have shown that this estimator has good asymptotic properties. Guo and Carlin (2004) used a Bayesian approach to estimate these joint models.

2.4.3.1 Convergence problems

Joint models that use shared random effects have convergence problems. For example, equation (2.4.3) involves two integrals that do not have analytic solutions. The two integrals are usually approximated by numerical integration with Gauss-Hermite and Gauss-Kronrod quadratures (Henderson et al., 2000; Rizopoulos, 2010). The numerical approximations of the integrals, mostly the Gauss-Hermite for the random effects makes the calculations to be slow. In fact, the integral over the random effects is usually multidimensional with size q.

When q is less than 3, the Gauss-Hermite quadrature remains the standard method but in higher dimension settings when q is more than 3, alternative methods may be preferred to reduce the computational time. These methods include Laplace method, which was proposed by Rizopoulos et al. (2009) or a Monte Carlo method. Sene et al. (2014) noted that the structure of B does not intervene in the computational complexity, only the number q of random effects is limiting.

In order to improve the accuracy and reduce the number of nodes in Gaussian quadratures, adaptive versions have been proposed (Lessaffre & Spiessens, 2001).

The adaptive versions consist of centering and rescaling the integral around its modal value until the nodes are systematically placed at the optimum position. The challenge with this technique is that it is time consuming because it requires a subject and iteration specific optimization to define the optimum position. In the effort of retaining the same precision but simplifying the numerical aspect, Rizopoulos (2011) developed a pseudo-adaptive version in which the integral is centered and rescaled according to the posterior distribution of the random effects defined in the linear mixed model in (2.4.1) but estimated separately and once for all in a first step.

2.5 Other Types of Dependence

Other types of dependency can be assumed. Instead of considering strictly random effects shared between the two sub models, another approach such as correlated error terms can be considered.

2.5.1 Correlated error structures

Joint models with correlated error structures have been used by Wang and Taylor (2001) and Henderson et al. (2000). Wang and Taylor (2001) used an integrated Ornstein Uhlenbeck process. Originally Faucett and Thomas(1996) and Wolfsohn and Tsiatis(1997) used a linear mixed model with only a random intercept and a random slope but any function of time can be considered in $X_{Li}(t_{ij})$ and $Z_i(t_{ij})$ in (2.4.1) to capture the best trajectory of the repeated measures.

Henderson et al. (2000) used latent Gaussian stochastic process shared by the longitudinal and time to event process. Verbeke et al. (2010) noted that there is conflict for information between the random effects structure and measurements error structure that assumes correlated errors. This is because both structures aim at

modeling the marginal correlation in the data. Further Verbeke et al. (2010) advised that it is necessary to opt for either correlated error terms or an elaborate random effect structures (that uses for splines in design matrix Z_i). Both random effects and correlated error term should not be used at the same time. Tsiatis and Davidian (2004) provide more details on the differences between random effects and correlated error term.

2.5.2 Model Formulation

As it has already been said Henderson et al. (2000) proposed a method that uses correlated error structures. This sub section reviews the method for joint modeling of survival and longitudinal data proposed by Henderson et al. (2000) in their paper entitled "Joint modeling of longitudinal measurements and event time data".

Suppose that there are n individuals in a longitudinal study, studied for a period interval of (0, c]. The ith individual gives measurements y_{ij}, where $j = 1,2, \dots, n_i$ at times t_{ij}, $j = 1, \dots, n_i$. The realizations of counting process $\{N_i(u)\ for\ 0 \leq u \leq \tau\}$ for event time and predictable zero-one process $\{H_i(u)\ for\ 0 \leq u \leq \tau\}$ that shows if an individual is at risk of having the event of interest, in our case the event of interest is death.

Henderson et al. (2000) proposed a method for analyzing joint model on longitudinal measurement and time to an event. This method is the extension of the work of Wulfsohn and Tsiatis (1997). This method allows the survival and longitudinal data to be linked by latent stochastic process. Henderson et al. (2000) suggested latent bivariate Gaussian process, in which $W_i(t) = \{W_{1i}(t), W_{2i}(t)\}$. The repeated

measurement process depends on $W_{1i}(t)$ while survival process depends on $W_{2i}(t)$. The longitudinal measurement process takes the form:

$$Y_i = u_i(t_i) + W_i(t_i) + e_i \qquad 2.5.1$$

In the equation 2.5.1, e_i is an error term and is distributed as $N(0, R_i)$ and $Var(e_{ij}) = \sigma_e^2$.

In the equation 2.5.1, $u_i(t_i) = ((u_i(t_{i1}), \dots, u_i(t_{ini}))^T$ and $W_i(t_i) = ((W_i(t_{i1}), \dots, W_i(t_{ini}))^T$. In fact $u_i(t_i)$ is described as linear model, for example $u_i(t_i) = X_{1i}\beta_M$.

For the latent process $W_{1i}(t)$, Henderson et al. (2000) consider $W_{1i}(t) = U_{1i} + U_{2i}(t)$ where (U_{1i}, U_{2i}) is a bivariate normal random vector, which has zero mean and variance covariance $G_i = \begin{pmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \sigma_{12}^2 & \sigma_2^2 \end{pmatrix}$.

It can be noted that model 2.5.1 and 2.1.2 are similar. In fact $u_i(t_i)$ in model 2.5.1 is $X_{1i}\beta_M$ in model 2.1.2, while $W_i(t_i)$ in 2.5.1 corresponds to $Q_i s_i$ in (2.1.2) with $s_i = (U_{1i}, U_{2i})^T$.

In joint model, the time to event process is modeled by using Cox proportional hazard model

$$h_i(t) = h_0(t)\alpha_0(\mathrm{t})\exp(x_{2i}^T\beta_s + W_{2i}(t)). \qquad 2.5.2$$

In the equation 2.5.2, $\alpha_0(\mathrm{t})$ is left unspecified in order to avoid the impact of the parametric assumptions. The longitudinal measurement and time to an event are assumed to be conditionally independent given $W_i(t)$. In order to create association between two processes, $W_{2i}(t)$ is taken to be related to some components of $W_{1i}(t)$. This has been achieved by using the following general equation $W_{2i}(t) = \gamma_1 U_{1i} + \gamma_2 U_{2i} + \gamma_3 W_{1i}(t)$. For instance, joint model with $W_{2i}(t) = \gamma_1 U_{1i} + \gamma_2 U_{2i}$ would

allow both random intercept U_{1i} and slope $\gamma_2 U_{2i}$ involved in equation 2.5.1 to affect the risk event.

To fit a model with random intercept only, the following formula: $Y_{ij} = X_{i1}^{'}\beta_1 + U_{oi} + Z_{ij}$ for longitudinal sub model is used. The formula for hazard sub model is $h_i(t) = h_0(\text{t})\exp\{X_{i2}^{'}\beta_2 + \gamma U_{oi}\}$ (Phillipson et al., 2012). Joint model with random intercept and slope is modeled using following; $Y_{ij} = X_{i1}^{'}\beta_1 + U_{oi} + U_{1i}t_{ij} + Z_{ij}$ for longitudinal sub model and $h_i(t) = h_0(\text{t})\exp\{X_{i2}^{'}\beta_2 + \gamma(U_{oi} + U_{1i}t)\}$ for time to event sub model. Joint model with quadratic random effects is modeled using following; $Y_{ij} = X_{i1}^{'}\beta_1 + U_{oi} + U_{1i}t_{ij} + Z_{ij}$ for longitudinal sub model and $h_i(t) = h_0(\text{t})\exp\{X_{i2}^{'}\beta_2 + \gamma(U_{oi} + U_{1i}t + U_{2i}t^2)\}$ for time to event sub model (Phillipson et al., 2012).

2.5.3 Likelihood Function

Marginal distribution for observed measurement is obtained by factorizing the likelihood for observed measurements and product of conditional distribution of event N given observed values of Y. Henderson et al. (2000) described the likelihood function as follows.

Let θ represent combined vector of unknown parameters. The likelihood $L = L(\theta, Y, N)$ can be expressed as

$$L = L_Y x L_{N|Y} = L_Y(\theta, Y) x\, E_{w_{2|Y}}[L_{N|W_2}(\theta, N|W_2)] \qquad 2.5.3$$

In equation 2.5.3, $L_Y(\theta, Y)$ is a standard form corresponding to marginal normal distribution of Y conditional likelihood for event data, $L_{N|W_2}(\theta, N|W_2)$ depicts any

likelihood contribution coming from achieved number of measures before any failure (Henderson et al., 2000). Let $A_0(t) = \int_0^t \alpha_0\,(u)du$ to represent cumulative baseline intensity, $L_{N|W_2}$ can be written as

$L_{N|W_2}(\theta, N|W_2) =$

$\prod_i[\prod_t[\exp\{\, X_{2i}\,(t)'\beta_2 + W_{2i}(t)\alpha_0(t)]^{\Delta N_i(t)} x \exp[-\int_0^T H_i \exp\{\, X_{2i}(t)'\beta_2 + W_{2i}(t)dA_0(t).$

To come up with L requires an expectation with respect to distribution of infinite dimensions process W_2 given longitudinal measurements Y.

2.5.4 EM Estimation

For the EM algorithm, Henderson et al. (2000) extended the EM algorithm presented by Wulfsohn and Tsiatis (1997). The procedure works by iterating up to when convergence is reached. For the E step, consider random effects $U = (U_1, U_2, U_3)^T$ as missing data. The expected value can be determined conditional on observed data (Y, N) for all $h(u)$ appearing in the (θ, Y, N, U). This is the complete data likelihood. The conditional expectation can be expressed as

$$E[h(u)|Y, N] = \{\int h(u)f(N|U)f(U|Y)du\}|f(N|Y) \qquad 2.5.4$$

Where $f(N|Y) = \int (N|U)f(U|Y)\,du$ 2.5.5

In the equation 2.5.4 and 2.5.5, $f(N|U)$ is the contributions of the ith subject to the event time of complete likelihood and $f(U|Y)$ is the conditional of random effects given longitudinal data. The term U is low dimension, therefore is approximated by using Gauss-Hermite quadrature. The Gauss-Hermite quadrature is also used to evaluate the final log likelihood, $\log(L(\theta, Y, N) = \log(E_U[L(\theta, Y, N, U)|Y, N]$. In the second step called maximization, the complete data likelihood is maximized. The $h(u)$ function is replaced by its expectation.

2.5.5 Joint Latent Class models

Apart from joint models with shared random effects. The alternative approach is joint latent class models (JLCM), which relies on a different idea. JLCM assumes that the population is heterogeneous and therefore can be divided into a finite number of homogeneous subgroups or classes. Each class or subgroup is characterized by a specific trajectory of the repeated measurement variable and a specific risk of event (Lin et al., 2002; Proust-Lima & Taylor, 2009). Proust-Lima et al. (2012) argue that the latent class structure can be seen as a latent stratification, which entirely captures the dependency between the longitudinal and survival processes.

2.6 Extension of Joint Models

In the previous sections, the discussion has focused on joint models based on a Cox model for right censored survival data and a LME model for longitudinal data. Other extensions of joint models for survival data and longitudinal data can also be considered. For example, for survival data, instead of using Cox proportion hazard model, other form of models can be used. These forms include accelerated failure time (AFT) models, models for interval censored data and models for recurrent events.

For longitudinal data, nonlinear, generalized linear mixed models, semi parametric or nonparametric mixed models can be used. Wu et al. (2012) argued that whether a person uses different survival models and longitudinal models, basic ideas and approaches for inference are the same. This subsection reviews some of the extensions of joint models.

2.6.1 Joint Models Based on an Linear Mixed Effect Model and an Accelerated Failure Time Model

In joint modeling of longitudinal and survival data, the AFT model can be used to model the survival process. This review focuses on an AFT model with measurement errors in time dependent covariates. For longitudinal data, the linear mixed effect models can be used. This review is based on the work of Tseng et al. (2005). A semi parametric AFT model can be expressed in the similar way as the Cox model:

$$h_i(t) = h_o[\int_0^t \exp\{-z_i^*(u)\beta\}du]\, exp\{-z_i^*(t)\beta\}$$

where $h_i(t)$ is the hazard function of the ith individual at time *t*, $h_0(t)$ is the baseline hazard function, and $z_i^*(t)$ is the unobserved true covariate value at time t for the observed measurements $z_i(t)$. Tseng et al. (2005) proposed a likelihood method using an EM algorithm.

The likelihood function for all observed data is be written as

$$L(\theta) = \prod_i^n \int f(t_i, \delta_i | z_i^*, h_0, \beta) f(z_i^* | w_i, \alpha, \sigma^2) f(w_i | W) dw_i$$

where $f(t_i, \delta_i | z_i^*, h_0, \beta) = \left[h_0\{\phi(t_i; \theta, w_i)\} \frac{\partial \phi(t_i; z_i^{**}, \beta)}{\partial (t_i)}\right]^{\delta_i} \exp\{-\int_0^{t_i; z_i^{**}, \beta} h_0(u)du$

where z_i^{**} denotes the covariate history and ϕ is a known function.

Wu et al. (2012) noted that handling the AFT structure in the joint modeling setting is more difficult than for the Cox model because $f(t_i, \delta_i | z_i^*, h_0, \beta)$ is more complicated and the baseline function $h_0\phi(t_i;\ z_i^{**}, \beta)$ involves unknown quantities. Further, Wu et al. (2012) observed that the point mass function with masses assigned to all uncensored survival times t_i cannot be used for the baseline hazard function h_0. In order to avoid this problem, Tseng et al. (2005) assumed the baseline hazard function

h_0 to be a step function that takes constant values between two consecutive failure times. Monte Carlo EM algorithm was used to obtain the MLEs. In the E step, Monte Carlo method was used to approximate the conditional expectations. The M step involves more complicated computations due to the complicated baseline hazard h_0.

2.6.2. Joint Models with Interval Censored Survival Data

The previous sections have so far focused on right censored survival data. In some cases events are known to occur over certain time intervals. This kind of survival data is called interval censored. To simplify things, it will be assumed that all individuals were assessed at the same times.

Let S_i be the time to an event i.e. survival time for individual i, with observed value s_i. Let $r_i = (r_{i1}, \dots, r_{im})^T$ denote the vector of event indicators such that $r_{ij} = 1$ if individual i has an event occurred from time t_{j-1} to time t_j, and let $r_{ij} = 0$ otherwise for $i = 1,2, \dots, n;\ j\ 1,2, \dots, m$. Assume that $r_{i1} = 0$ for all i.

Let $p_{ij} = P(t_{j-1} \leq S_i < t_j)$ and let $\pi_{ij} = P(t_{j-1} \leq S_i < t_j \mid S_i \geq t_{j-1}) = 1 - P(S_i \geq t_j \mid S_i \geq t_{j-1})$.

Then, it follows that $p_{ij} = (1 - \pi_{i1}), (1 - \pi_{i1}) \dots (1 - \pi_{i,j-1})\pi_{ij}$. The probability function for the event indicator vector r_i can be written as

$$f(r_i) = \prod_i^m p_{ij}^{r_{ij}} = \prod_i^m \pi_{ij}^{r_{ij}} (1 - \pi_{ij})^{1-r_{ij}} \qquad 2.6.1$$

It can be noted that equation 2.6.1 is the probability function for a Bernoulli distribution. Further the observed error prone covariate value z can be introduced and z_i^* is its true value. Again assume

$\log\{-\log(1-\pi_{ij})\} = \beta^T z_i^* + \gamma_j$, where β and $\gamma = (\gamma_1, \gamma_2, \ldots \gamma_m)^T$ are unknown parameters. Then the probability function of r_i can be written as $f(r_i \,|z_i^*, \beta, \gamma)$. Denote θ as the collection of all parameters in all models, the likelihood for observed data can be expressed as:

$$L_0(\theta) = \prod_{i=1}^{n} [\int f(z_i \,|w_i, \alpha, \sigma) f(r_i \,|w_i, \beta, \gamma)\, f(w_i \,|W) dw_i] \qquad 2.6.2$$

In the equation 2.6.2, $f(z_i \,|w_i, \alpha, \sigma)$ is the conditional probability density function, given the random effects w_i and $f(w_i \,|W)$ be the marginal probability density function for w_i with covariance matrix W.

Maximum likelihood estimators (MLE) of parameters θ can be calculated by maximizing the observed data likelihood $L_0(\theta)$.

Evaluating $L_0(\theta)$ can be difficult because it involves an evaluation of intractable and possibly high-dimensional integral (Wu et al., 2012). Monte Carlo EM algorithms can be used.

2.6.3. Generalized Linear Mixed Models and Nonlinear Mixed Effects Models for Longitudinal Data

So far, this study has focused on LME models for modeling the longitudinal data. It is also possible to consider other types of models for longitudinal data. Wu et al. (2008) and Wu et al. (2010) considered nonlinear mixed effects (NLME) models for modeling the longitudinal data in joint models. When the longitudinal data is not normally distributed, generalized linear mixed models (GLMMs) can be used. GLMMs are nonlinear models and also empirical models.

When dealing with longitudinal models that are nonlinear, both two-stage and likelihood approaches for joint models may still be applied (Wu et al., 2012). The

major challenge with this type of models is that computation is more demanding because of the nature of nonlinearity of the longitudinal models.

2.6.4. Joint Models with Missing Data

Survival models with measurement errors in time dependent covariates have received much attention in the joint models literature. Another common situation is longitudinal models with informative dropouts, in which survival models can be used to model the dropout process. In both cases the focus is on creating association between longitudinal and survival processes. Joint models have also been considered in which the focus is on more efficient inference of the survival model by using longitudinal data as auxiliary information (Xu & Zeger, 2001; Faucett et al., 2002; Hogan & Laird, 1997) or assume that the longitudinal process and the survival process to be governed by a common latent process (Henderson et al., 2000).

When missing data are non ignorable, missing data process is normally included in inferential procedures. It is easy to incorporate missing data mechanisms in joint model inference that use likelihood methods (Wu et al., 2012). However the computation becomes more challenging. Wu et al. (2008) and Wu et al. (2010) considered the missing data problems for joint models by using Monte Carlo EM algorithms and Laplace approximations.

2.6.5 Models with Longitudinal data and Competing Risks

Standard methods for joint modeling of longitudinal and survival data allow for one event with a single mode of failure and an assumption of independent censoring. When there are several reasons why an event can happen, or other informative censoring happens, it is known as competing risks (Williamson et al., 2008). The

standard methods are not applicable to survival data with competing risks or informative censoring (Elashoff et al., 2007).

Elashoff et al. (2007) proposed a method for analyzing longitudinal measurements and competing risks failure times that allow for more than one distinct failure type. The method handles informative censoring by treating it as a competing risk in the model. It can also be used to model non ignorable missingness after event times. The longitudinal data is modeled using linear mixed effects and a mixture sub model is used to analyze competing risks survival data. The mixture model for competing risks enables one to evaluate the effects of some factors on both the marginal probabilities of occurrence of the risks and the conditional cause specific hazards. Parameters were estimated using an EM algorithm in both sub models.

Williamson et al. (2008) proposed a method that uses cause specific hazards sub model to allow for competing risks with a separate latent association between longitudinal measurements and each cause of failure. The joint analysis longitudinal measurements and competing risks failure time data is more challenging as compared to joint analysis of longitudinal measurements and survival data with a single failure type (Elashoff et al., 2007).

2.6.6 Joint Models with Multivariate Longitudinal data Outcome

Joint models can also be extended to multivariate cases. The longitudinal processes and event processes can be modeled simultaneously. Computation for joint models of this type is more challenging as compared to univariate cases (Xu & Zeger, 2001; Song et al., 2002).

Chi and Ibrahim (2006) proposed a likelihood approach that extends both longitudinal and survival components to multi-dimensional. A multivariate mixed effects model is used to capture dependence among longitudinal data over time and also dependence between different variables. For the survival component of the joint model, a shared frailty was introduced in order to induce correlation between failure times. The marginal univariate survival model is then applied to each marginal survival function. The multivariate survival model has a proportional hazards structure for the population hazard when the baseline covariates are specified through a specific mechanism. This method is also capable of modeling survival functions that have different cure rate structures.

Rizopoulos and Ghosh (2011) proposed a semi parametric multivariate joint model, which relates multiple longitudinal outcomes to time to event. In order to allow for greater flexibility, key components of the model were modeled non-parametrically. For the subject specific longitudinal evolutions a spline based approach was used. Baseline risk function was assumed to be piecewise constant. Distribution of the latent terms was modeled using a Dirichlet process prior formulation.

Baghfalaki et al. (2014) proposed a method for analyzing multivariate longitudinal data comprising of mixed continuous and ordinal responses and a time to event variable. The association structure between longitudinal mixed data and time to event data was modeled using a multivariate zero-mean Gaussian process. Discrete ordinal data was modeled by making an assumption that a continuous latent variable follows the logistic distribution. Continuous data was modeled by using a Gaussian mixed effects model. Baghfalaki et al. (2014) used an accelerated failure time model for the

event time variable. Parameters were estimated by a Bayesian approach that uses Markov Chain Monte Carlo.

2.7 Model Selection

In some cases, it becomes necessary to compare models, which are not nested. Models, which are nested, can be compared using the likelihood ratio test. Models that are not nested can be compared using approaches such as the Akaike information criterion (AIC) and Bayesian Information Criteria (BIC). AIC is defined as

$$AIC = -2l + 2(k + w) \qquad 2.7.1$$

where l is the log-likelihood, k is the number of covariates in the model and w is the number of model specific ancillary parameters. The term $2(k + w)$ in equation 2.7.1 can be thought of as a penalty for including extra predictors in the model. Smaller values of AIC suggest a better model (Hedeker & Gibbon, 2006). Another approach for model selection is BIC. BIC may be written as

$BIC = -2l + (k + w)\ln(n)$, where $\ln(n)$ is the log of the sample size n. The model is considered to be better if it has smaller the value of BIC (Hedeker & Gibbons, 2006).

The deviance information criterion (DIC) is considered as a hierarchical modeling generalization of the AIC and BIC. DIC is useful in Bayesian model selection problems, in which posterior distributions of the models are obtained by Markov chain Monte Carlo (MCMC) simulation. The problem of using AIC, BIC and DIC is that there are no formal statistical tests to compare different AIC values, different BIC values, and different DIC values respectively. Just like AIC and BIC, DIC is asymptotic approximation as the sample size becomes large. This approach is valid

when posterior distributions are approximately multivariate normal. AIC has been used in this thesis in order to compare models. AIC penalizes the number of parameters less strongly than does the Bayesian information criterion (BIC).

CHAPTER 3 METHODOLOGY

3.1 Introduction

Chapter 3 describes the methodology used in this study. In particular, study design, data collection and data analysis and likelihood function for joint model are described.

3.2 Study design

The study used secondary data, which was collected at Queen Elizabeth Central Hospital, in Malawi. The study design used was prospective cohort study, and participants were followed for a period of 14 weeks. The study participants were randomized into two groups using block randomization. The first group received corn soya blends (CSB), and other group received ready to use therapeutic food (RUTF). In total there were 491 participants, of these 246 received CSB and 245 received RUTF. In this study, weight of patients were measured at fixed times (4 times) for a total duration of 14 weeks (3 and half months). Ethical approval was received from the College of Medicine Research Ethical Committee (COMREC).

3.2.1 Participants and Duration

The study registered male and female who were HIV positive and were at least 18 years old. Participants were excluded if they were pregnant women, mothers who were breastfeeding and were participating in another supplementary feeding program. The study took place in the year 2006.

3.3 Nutritional Value of Food Supplements

The nutritional contents of the CSB and RUTF supplementary foods are given in Table 1. Nutritional contents given to patients in the CSB and RUTF supplementary foods were almost similar.

Table 1: Nutritional contents available in Corn Soya Blends and Fortified

	Ready to use therapeutic food (RUTF)	Corn-soy blended flour (CSB)	Estimated Average Requirements	
	(245 g/day)	(374 g/day)	Women	Men
Energy (kJ)	5694	5694	13252	13252
Protein (g)	35.5	50	46	56
Fat (g)	91	26.2	-	-
Calcium (mg)	830	258	1000	1000
Phosphorus (mg)	700	1050	580	580
Magnesium (mg)	240	500	255	330
Potassium (mg)	2880	1700	4700	4700
Selenium (µg)	78	22	45	45
Zinc (mg)	8	8	8	11
Copper (mg)	0.9	2.9	0.9	0.9
Iron (mg)	8	16	18	8
Vitamin A (µg)	710	1040	700	700
Vitamin C (mg)	90	26	60	75
Vitamin D (µg)	5	5	6	5
Vitamin E (mg)	52	32.5	12	12
Niacin (mg)	14	13	11	12
Folic acid (µg)	400	153	320	320
Thiamine (mg)	1.1	1.3	0.9	1.0
Riboflavin (mg)	1.3	0.8	0.9	1.1
Vitamin B-6 (mg)	1.3	1	1.1	1.1
Vitamin B-12 (µg)	1.4	0.5	2.0	2.0

Source: Ndekha et al (2009)

3.4 Data Description

The dataset had the following variable: type of food supplement given to the participants, sex of participant, TB status of patient, whether participant was receiving cotrimoxazole or not, age of participant in years, CD 4 count of participant, and hemoglobin level of participant. The categorical variables were coded as follows: type of food supplement was coded as 1 if participant was receiving CSB and 0 if receiving RUTF. Sex of participants was coded 0 if participant was female and 1 when participant was male. If participant had TB it was coded 1 and 0 otherwise. If participant was receiving cotrimoxazole, it was coded 1 and 0 otherwise. Body mass index was a continuous variable and was measured in kg / m^2. Body mass index was the only repeated measurement variable. Each participant was expected to have 4 visits. Hemoglobin level of participants was a continuous variable and was measured in mg/dl. CD 4 lymphocytes count of the patients was also measured. Survival time was measured in weeks.

3.4.1 Missing Data

As it has already been stated, this study used secondary data. The study used a complete dataset.

3.5 Data Analysis

3.5.1 Exploratory Data Analysis

The exploratory data analysis was done using statistical package called R, version 2.15.2. For categorical variables such as sex, TB status, proportions were used to summarize the categorical variables. Mean and standard deviation were used if the variable was continuous and normally distributed. Median and inter quartile range

were used if the variables were skewed. Confidence intervals were calculated at 95% where appropriate. Hypotheses were tested at 5% level of significance.

For each visit, the BMI for the participants in 2 groups were compared using a t test at 5% level of significance. Histograms for age, BMI, CD 4 count and hemoglobin level were plotted. Also graph subject specific evolutions in time for BMI were plotted for the 2 groups receiving food supplement, namely CSB and RUTF. Kaplan Meir graph was plotted in order to assess the survival of the 2 groups receiving different food supplements.

3.6 Model Fitting

3.6.1 Model for Survival Analysis

This model was fitted using survival package in R. The package survival is able to fit Cox proportion model with either time independent covariate model or time dependent covariate model. Cox proportion model with time dependent model was fitted. The dependent time covariate Cox model has been described in section 2.2.5. The model has age, sex, CD4 count and hemoglobin level as baseline time independent covariates and body mass index was fitted as a time dependent covariate.

3.6.2 Model for Longitudinal Data

This longitudinal model was fitted using R package called “nlme”. The linear mixed effect regression model was fitted. In the model body mass index was outcome variable, while age, sex, CD4 count hemoglobin level were independent variables. The linear mixed effects regression model has been discussed already in section 2.1.1. The model with intercept and random slope were used.

3.6.3 Model fitting for Joint Modeling

Joint modeling analysis was done in R. Statistical package joineR used to analyze data in this study was developed based on the work of Henderson et al. (2000) and Wulfsohn and Tsiatis (1997). JoineR package provide a function for fitting Wulfsoln and Tsiatis models called joint. This function permits the user to choose from three models for the joint random effects namely; random intercept; random intercept and slope; and quadratic random effects (Phillipson et al., 2012). Other function in joineR is jointdata, which supplies the data for analysis. Surv object provide survival data and long provide longitudinal data.

3. 7 Confidence Intervals and Standard Errors for Joint Model

The joineR uses bootstrap methods to calculate confidence intervals and standard errors. Phillipson et al. (2012) described bootstrap is a general computational tool that can be used to assign measures of accuracy to statistical estimates. Phillipson et al. (2012) further described how confidence intervals and standard errors were calculated in joiner using bootstrap method. The method works by generating N independent bootstrap samples $\{W^{*1}\}$, $\{W^{*2}\}$,. . . ,$\{W^{*N}\}$. Every independent sample contains n data values drawn randomly with replacement from the original data {W}. In this case, original data comprises of both longitudinal and survival outcomes including survival time, censoring indicator, longitudinal measurements and treatment type (Phillipson et al., 2012).

Standard error of an estimate θ is obtained by calculating sample standard deviations of the N bootstrap samples. Confidence intervals are calculated as follows: Let $\theta^{*(1)} < . . . < \theta^{*(N)}$ represent the ordered bootstrap replications of θ. The 95% confidence

interval for θ is approximated by ($\theta^{*(0.05N)}$, $\theta^{*(0.95N)}$) (Phillipson et al., 2012). Joiner uses joint.se function to calculate confidence intervals and standard errors.

The package joiner used for joint modeling survival and longitudinal data does not produce the p values. Therefore in order to assess whether the variable is significant or not, confidence intervals are used. If the confidence interval range contains 1, the variable is significant, if the confidence interval range does not contain 1, then the variable is not significant.

CHAPTER 4 RESULTS

Chapter 4 contains an analysis of the data collected and interpretations of the findings of this study.

4.1 Exploratory Data Analysis

The study interviewed 491 HIV positive patients who were starting antiretroviral therapy (ART). Out of these, 294 (59.8%) were female and 197 (40.1%) were male. The study participants were randomized into 2 groups, 246 patients were given corn soya blends (CSB) and 245 patients were given ready to use therapeutic food (RUTF). The median age was 33.8 years with inter-quartile range of (28.2 - 41.7 years), median age for women was 31.7 years with inter-quartile range of (26.9- 37.9 years) and for men was 37.4 years with inter-quartile range of (32.0- 45.0 years). Median age for group getting CSB was 34.0 years, for group receiving RUTF was 33.1 years. More than two thirds (68.8%) of the participants received cotrimoxazole. For group receiving CSB, 67.1% were receiving cotrimoxazole, 70.6% of participants in RUTF group received cotrimoxazole. The mean hemoglobin level was 9.7 mg/dl with standard deviation of 2.1 mg/dl. Among patients receiving CSB, mean hemoglobin level was 9.8 mg/dl and standard deviation was 2.2mg/dl. Mean hemoglobin level for RUTF group was 9.5mg/dl and standard deviation was 2.0 mg/dl. Median CD4 count for the participants was 90; with inter quartile range of (33-184). Patients who were receiving CSB had median CD4 count of 91, with inter quartile range of (33-185);

those who were receiving RUTF had median CD4 count of 88; with inter quartile range of (34-182). Almost one fifth of the participants (21.8%) had tuberculosis (TB). Out of 491 patients, who participated in this study, 27.7% died. Among participants receiving CSB, 26.0% died and 29.4% of participants receiving RUTF died. Among female 24.5% died, CI (19.7%, 29.8%); while among male 32.5% died, CI (26.0%, 39.5%), Refer Table 2.

Table 2: Descriptive Results for Participants According Food Supplement Groups

Food Supplement	Group receiving CSB(n= 246)	Group receiving RUTF(n=245)	All participants (n= 491)
Sex			
Female	142 (57.3%)	152 (62.0%)	294(59.9%)
Male	104 (42.3%)	93 (38.0%)	197(40.1%)
Age in years- median(inter quartile range)	34.0(29.9- 42.1)	33.1(28.8- 41.3)	33.8(28.2-41.7)
Male	38.6(32.2-45.0)	36.9(31.6-45.4)	37.4(32.0-45.0)
Female	31.3(26.5-38.9)	31.9(27.1-37.5)	31.7(26.9-37.9)
No of patients Died	64(26.0%)	72 (29.4%)	136(27.7%)
Female who died	42(27.6%)	30(21.1%)	72 (24.5%)
Male who died	32 (32.3%)	32(32.7%)	64 (32.5%)
Number of patients on Cotrimoxazole	165 (67.1%)	173(70.6%)	338(68.8%)
Hemoglobin- mean(sd)	9.8 (2.2)	9.5 (2.0)	9.7(2.1)
CD 4 Count of participants median(inter quartile range)	91(33 – 185)	88 (34 – 182)	90 (33 – 184)
No of patients with TB	60(24.5%)	47(19.7%)	107(21.8%)

Body mass index (BMI) was calculated for all the visits. See Table 3 that gives values of the BMI for the study participants. At the enrolment time, mean BMI was 16.5 kg/m^2, with standard deviation of 1.4 kg/m^2. Mean BMI for group receiving CSB was 16.5 kg/m^2 and 16.5 kg/m^2 for group receiving RUTF. There was no significant difference in BMI for patients who were receiving CSB and RUTF, t test = 0.36, p value = 0.717. Second visit took place 2 weeks later, at that time the mean BMI for all the study participants was 17.0 kg/m^2, with standard deviation of 1.6 kg/m^2. Mean BMI for both groups was at 17.0 kg/m^2. No significant difference was detected during second visit, t test = 0.01, p value = 0.992. During the third visit, BMI for group receiving CSB was 17.5 kg/m^2 and BMI of group receiving RUTF was 17.7 kg/m^2. There was no significant difference in the BMI of the two groups, t test = 1.33, p value = 0.184. During forth visit, BMI for the group receiving CSB was 17.8 kg/m^2 and BMI for the group receiving RUTF was 18.3 kg/m^2. There was a significant difference in BMI of the two groups, t test = 2.38, p value = 0.018. At the visit 5 (14 weeks after the initiation of treatment) BMI for group receiving CSB was 18.4 kg/m^2 and that of group receiving RUTF was 19.0 kg/m^2. There was a significant difference of BMI between the two groups of patients, t test 2.98, p value = 0.003. This means that group receiving RUTF had higher BMI than group receiving CSB after 10 weeks of treatment.

Table 3: BMI for Patients Receiving CSB and RUTF at Different Times of Follow up

BMI	All participants	CSB	RUTF	T statistic	P value
Visit 0 mean(sd)	16.7(1.4)	16.5(1.4)	16.5(1.5)	0.36	0.717
Visit 1 mean(sd)	17.0(1.6)	17.0(1.5)	17.0(1.6)	0.01	0.992
Visit 2 mean(sd)	17.6(1.8)	17.5(1.7)	17.7(1.8)	1.33	0.184
Visit 3 mean(sd)	18.1(2.0)	17.8(1.8)	18.3(2.1)	2.38	0.018
Visit 4 mean(sd)	18.7(2.0)	18.4(1.8)	19.0(2.1)	2.98	0.003

NOTE: 1. t test compares BMI for participants who were receiving CSB and RUTF

2. BMI was measured in kg/m^2

4.1.1 Distribution of Variables

Body mass index (BMI) and Hemoglobin level were normally distributed. Age was skewed to the left, while CD4 count was highly skewed to the left, Refer Figure 1.

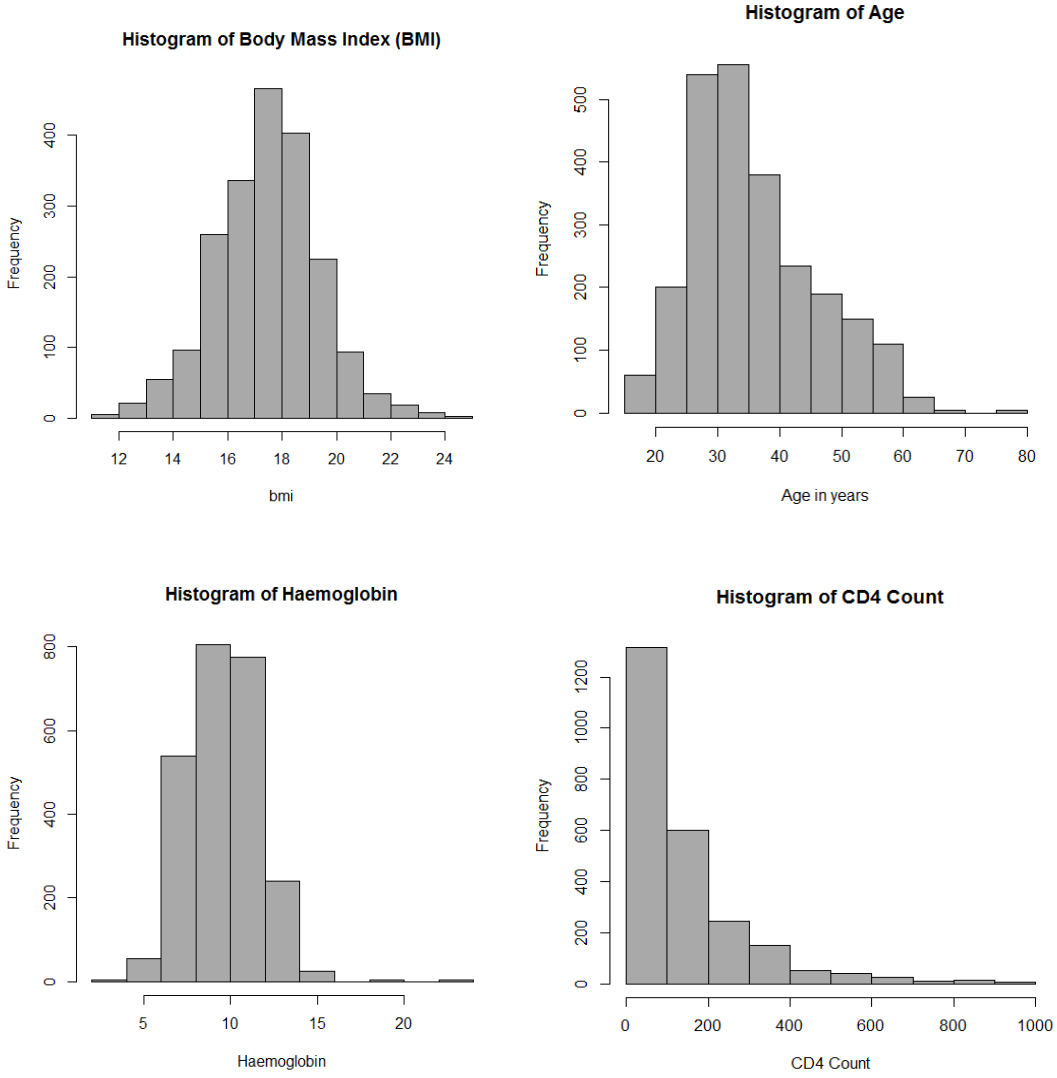

Figure 1: Distributions of BMI, Age, Hemoglobin Level and CD4 Count

Patients who were receiving CSB and RUTF had different variability in their longitudinal profiles for the body mass index (BMI), refer Figure 2.

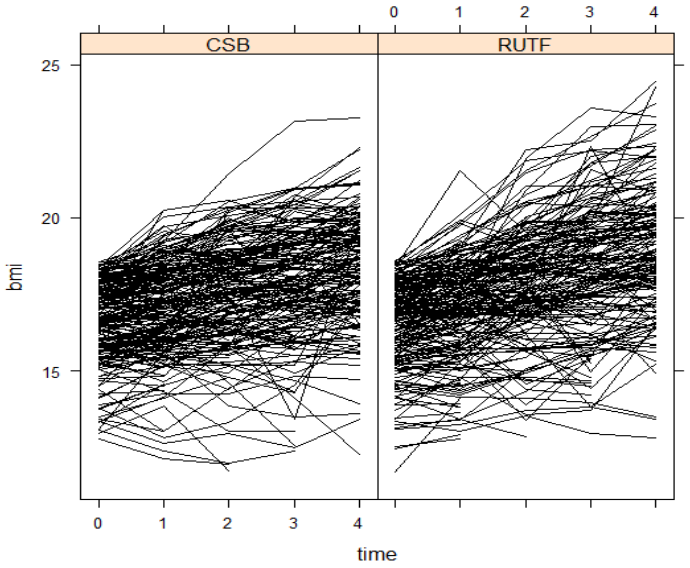

Figure 2: Subject Specific evolutions in time of Body Mass Index for CSB and RUTF

There was no difference in the survival of patients in the first 2 weeks. After week 2, the group that was receiving CSB has slightly higher survival than the group receiving RUTF, refer Figure 3. But there was no significant difference in the survival between patients who were receiving CSB and RUTF, using log rank test, Chi square= 0.9, 1 degree of freedom, p value = 0.342

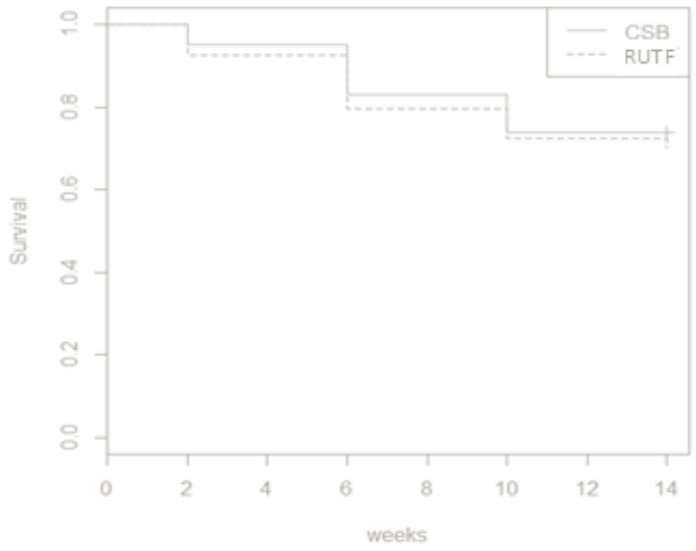

Figure 3: Kaplan Meier Graph for Participants Receiving CSB and RUTF

4.2 Survival Data Analysis

4.2.1 Time Dependent Cox Model

Results from Cox proportional hazards models with time dependent covariates are presented in Table 4. The group that was given RUTF had a higher risk of death, HR = 1.2, 95% CI (0.953, 1.511), however, this was not significant, p value = 0.120. Body Mass Index (BMI) had effect on the survival of patients, HR = 0.636, 95% CI (0.592, 0.683), p value < 0.00001. This means that as BMI increases by 1kg/m^2, the risk of death decreases by 36%. Male patients had a higher risk of death as compared to female patients, HR= 1.4, 95% CI (1.116, 1.830), p value = 0.005. This means that male had 1.4 times hazard of death as compared to women. Age of a patient had an

effect on the time to death, HR = 1.02, 95% (1.003, 1.026), p value = 0.016. As the age of person increased by 1 year, the hazard of death increased by 2 %. CD4 count had no effect on the survival of patients, HR = 1.0, 95% CI (0.999, 1.001), p value = 0.235. Hemoglobin level of patient had effect on survival, HR = 0.80, 95% CI (0.753, 0.854), p value <0.0001, this means as the hemoglobin level of patient increases by 1 mg/dl, the hazard of death for that patient decreases by 20%. TB status of a patient had no effect on survival of patients, HR = 1.08, 95% CI (0.968, 1.189), p value = 0.182. Receiving cotrimoxazole, had effect on survival of patients, HR = 0.36, 95% CI (0.287, 0.457) p value < 0.00001. Patients who were receiving cotrimoxazole had a lower risk of death compared to patients who were not receiving cotrimoxazole. Receiving cotrimoxazole reduces the hazard of death by 64%. Wald test had a value of 275.5 and p value <0.0001. Also both Likelihood test and score (Log rank) test had p value < 0.0001. All the three tests are significant, providing evidence that at least one of the coefficients is significantly associated with the time to death of the patient.

Table 4: Results from Cox Model with Time Dependent Covariates

Parameter	Exp (coefficient)	Std Error	95% Confidence Interval	Z	P value
Food type(RUTF)	1.201	0.118	(0.953, 1.511)	1.556	0.1197
BMI	0.636	0.037	(0.592, 0.683)	-12.394	< 0.0001
Sex (Male)	1.429	0.126	(1.116, 1.830)	2.829	0.0047
Age	1.014	0.006	(1.003, 1.026)	2.414	0.0158
CD4 Count	1.000	0.0003	(0.999, 1.001)	1.189	0.2346
Hemoglobin	0.802	0.032	(0.753, 0.854)	-6.845	< 0.0001
TB	1.073	0.523	(0.968, 1.189)	1.335	0.1819
Cotrimoxazole (Yes)	0.363	0.119	(0.287, 0.457)	-8.555	< 0.0001
Wald Test	275.3	P value < 0.0001			
Likelihood Test	282.8	P value < 0.0001			
Score(Log rank)Test	294.0	P value < 0.0001			

4.3 Longitudinal Data Analysis

In order to assess the effects of BMI over a long period of time, mixed effects model was fitted. Dependent variable was BMI. The mixed effect model was fitted with random intercept and slope. Table 5 shows the results of this model. In the model, type of food supplements has a coefficient of 0.160, which means that mean BMI of group getting RUTF is 0.160 higher than BMI of group getting CSB. But these results were not significant, p value = 0.2600. Sex of a person did not have effect in the changes of BMI, p value = 0.2655. CD4 count did not have significant effect on BMI,

p value = 0.5298. Hemoglobin level had an effect in the changes of BMI. Increase in BMI by 1kg/m^2 increases hemoglobin level of a patient by 0.0915, p value = 0.0085. Receiving cotrimoxazole was not significant, p value = 0.9204. TB status of a person did not have significant effect on the changes of BMI, p value = 0.0787.

Table 5: Results from Longitudinal Data

Parameter	Estimate	Std Error	95% CI	P value
Intercept	14.956	0.430	(14.113, 15.802)	<0.00001
Time	0.458	0.014	(0.430, 0.488)	<0.00001
Food type	0.1601	0.142	(-0.119, 0.439)	0.2600
Sex	-0.1702	0.153	(-0.470, 0.130)	0.2655
CD4 Count	0.0003	0.0005	(-0.0006, 0.0012)	0.5298
TB	-0.1264	0.07174	(-0.267,0.015)	0.0787
Cotrimoxazole	0.0155	0.1550	(-0.289,0.320)	0.9204
Hemoglobin	0.0915	0.034	(0.023, 0.159)	0.0085
Age	0.0183	0.007	(0.004, 0.032)	0.0103
AIC	6543.135			
BIC	6604.816			
Log likelihood	-3260.567			

Trajectories for the body mass index for the patients who were censored and those who died were plotted. The trajectories are shown in Figure 4.

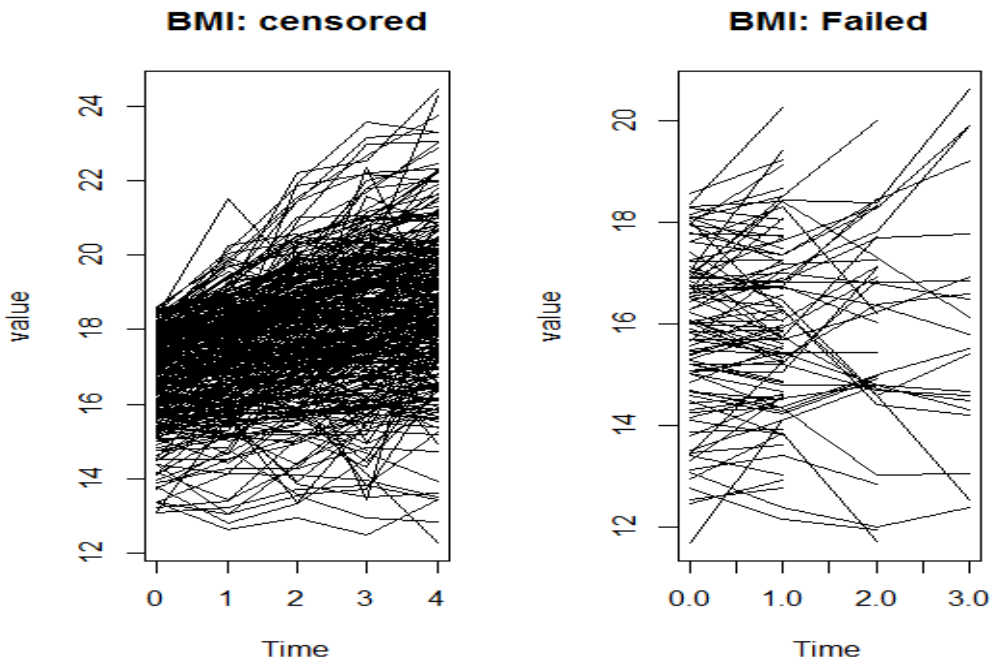

Figure 4: Trajectories for Body Mass Index

4.4 Joint Modeling Results

Table 6 presents results for joint modeling. Type of food did not have effect on survival of patient. This had a coefficient of 0.222 and a 95% confidence interval of (13.707, 15.135). Sex of patient had significant effect on survival of patient, coefficient = 0.507, 95% CI (0.150, 0.515). The hazard ratio was $e^{0.507} = 1.66$. This means male patients were at higher risk of dying than female patients. Age of patient did not have significant effect on survival of participants, 95% CI (-0.012, 0.018). Those patients who were receiving cotrimoxazole had lower risk of death as compared to patients who were not receiving, coefficient = - 0.922, 95% CI (-1.391, -0.648). The hazard ratio (HR) was $e^{-0.922} = 0.398$. Receiving

cotrimoxazole had an effect on time to death of patients. Patients who were receiving cotrimoxazole had lower risk of death than patients who were not receiving.

Patients who had higher levels of hemoglobin had lower risk of death as compared to patients with lower levels of hemoglobin, coefficient = - 0.274, 95% CI (-0.403, -0.201). The hazard ratio (HR) was $e^{-0.274} = 0.761$. Hemoglobin level had an effect on time to death of patients.

Table 6 also presents results longitudinal sub model. Type of food did not have significant effect on BMI, 95% CI (-0.266, 0.218). CD4 count of a patient has effect on BMI of the patient. Increase in BMI by 1 kg/m^2, increases CD4 count by 0.001. This was significant, 95% CI (0.0002, 0.0022). Receiving cotrimoxazole had no effect on the BMI of the patient. Hemoglobin level had effect on the BMI of a patient. In fact increase of BMI by 1 kg/m^2, increases hemoglobin level of patients by 0.12.

In Table 6, the joint model the latent association, γ is -0.178. The latent association quantifies the effect of longitudinal outcome to the risk of death. In our case the latent association measures the effect of body mass index (BMI) to the time to death. There is significant association between BMI and survival of a patient, 95% CI (-0.241, -0.141). The hazard ratio for increase the relation of BMI and survival of a patient is 0.84($e^{-0.178}$). Body Mass Index (BMI) had effect on the survival of patients HR = 0.84.This means that as BMI increases by 1 kg/m^2, the risk of death decreases by 16%. Variance for random intercept was 1.76; variance for random slope was 0.21; and residual variance was 0.35.

Table 6: Results from Joint Model

Longitudinal Model	Parameter		Estimate - coefficient	Std Error	95% CI
	Intercept		14.492	0.376	(13.707, 15.135)*
	Time		0.403	0.026	(0.339, 0.450)*
	Food type		-0.003	0.124	(-0.266, 0.218)
	Sex		-0.039	0.135	(-0.327, 0.227)
	Age		0.022	0.006	(0.006, 0.034)*
	CD4 Count		0.001	0.001	(0.0002, 0.002)*
	Cotrimoxazole		0.060	0.132	(-0.349, 0.282)
	Hemoglobin		0.119	0.034	(0.068, 0.187)*
	TB		-0.143	0.075	(-0.349, -0.023)*
Survival Model	Parameter	Exp(coef)	Estimate - coefficient	Std Error	95% CI
	Food type	1.245	0.223	0.152	(-0.043, 0.515)
	Sex	1.660	0.507	0.193	(0.150, 0.890)*
	Age	1.004	0.004	0.008	(-0.012, 0.018)
	CD4 Count	0.997	-0.003	0.001	(-0.002, 0.0003)
	Cotrimoxazole	0.398	-0.922	0.168	(-1.391, -0.648)*
	Hemoglobin	0.761	-0.274	0.046	(-0.403, -0.201)*
	TB	1.109	0.103	0.080	(-0.081, 0.257)
			Estimate - coefficient	Std Error	95% CI
Latent Association			-0.178	0.026	(-0.241, -0.141)
Variance for random intercept U_o			1.764	0.144	(1.455, 2.001)
Variance for random slope U_1			0.201	0.0181	(0.163,0.234)
Residual variance			0.348	0.035	(0.291, 0.415)
AIC			6443.14		
BIC			6503.82		
Log likelihood			-3307.66		

Note: *shows that results were significant

4.5 Comparison of Joint Models and Separate Models

Table 4 and Table 5 show results produced by separate models. Table 6 gives the results from joint modeling. For instance, it was observed that some variables like CD4 count and TB status did not have significant effect on body mass index (BMI) in the separate model. However in joint model CD4 count and TB status had significant effect on body mass index. It was also found age of patient had significant effect on the survival of the person in the separate model; however in the joint model age was not significant on the survival of a person. In the longitudinal, variables like age; Hemoglobin; had effect on the body mass index of patients in both separate and joint model. Also intercept and time were significant both in separate and joint models. Type of food supplement, sex of person did not have significant effect on the body mass index (BMI) in both separate and joint models.

In terms, of standard errors, most of variables in the joint model had smaller standard errors as compared to the same variables in joint model. For instance, type of food supplements had a standard error of 0.142 in separate model; and in joint model it had standard error of 0.124. Also CD4 count, sex, TB status, use of cotrimoxazole had smaller standard errors in joint model than in separate models, Refer Tables 5 and 6. Most of variable in joint models in the longitudinal sub model had shorter range of confidence intervals, Refer Tables 5 and 6. Narrow confidence intervals are more desirable that wider confidence intervals.

The models were compared using log likelihood. Model produced by separate longitudinal model had a log likelihood of -3260.57 and that of joint model was -3140.66. Model produced by longitudinal model has AIC of 6543.135 and that of

joint model had an AIC of 6443.14. Also separate longitudinal model had a BIC of 6604.812 as compared to 6503.82, which was BIC for the joint model. This suggests that joint model had a better fit than separate model.

4.6 Assessment of Model Assumptions

The Cox proportional hazards model assumes that $\beta(t) = \beta$. The hypothesis $H_o\beta(t) = \beta$ can be tested using the scaled Schonfeld residuals by an approximate score test (Grambsch and Therneau 1994). The Schonfeld's global test was used to test proportion hazard assumptions. The "rho" estimates the correlation coefficient between survival time and the scaled Schoenfeld residuals. The high p-value of more than 0.05 of the score test implies no evidence against the assumption of proportional hazards. A p-value of less than 0.05 of the test statistic (i.e., the model chi square, sometime referred to as Wald chi-square in some packages) indicates a good model fit. Under such a condition, the analyst concludes that the current model can reject the null hypothesis that all the regression coefficients equal zero, and equivalently, at least one coefficient that is not equal to zero. This means that there is no evidence against the assumption of proportional hazard, p value = 0.360, Refer Table 7. Hence, the model is acceptable.

Table 7 : Schoenfeld's Global Test Results

Variable	rho	Chi square	P value
Food type	0.026	0.259	0.6174
Body mass index	-0.091	2.857	0.0909
Sex	-0.115	3.738	0.0532
Age	0.010	0.025	0.8752
CD4 lymphocyte count	-0.040	0.527	0.4678
Hemoglobin level	-0.006	0.015	0.9041
Tuberculosis	0.030	0.266	0.6060
Cotrimoxazole	-0.033	0.327	0.5676
GLOBAL	NA	8.802	0.3593

Graphs for all the explanatory variables against survival time were fitted. All the graphs show that the fitted lines (slopes) for the scaled Schoenfeld residuals for all covariates are not significantly different from zero, See Figure 5. There the assumptions of proportional hazards have been met. This is in line with results obtained from the Schoenfeld global test.

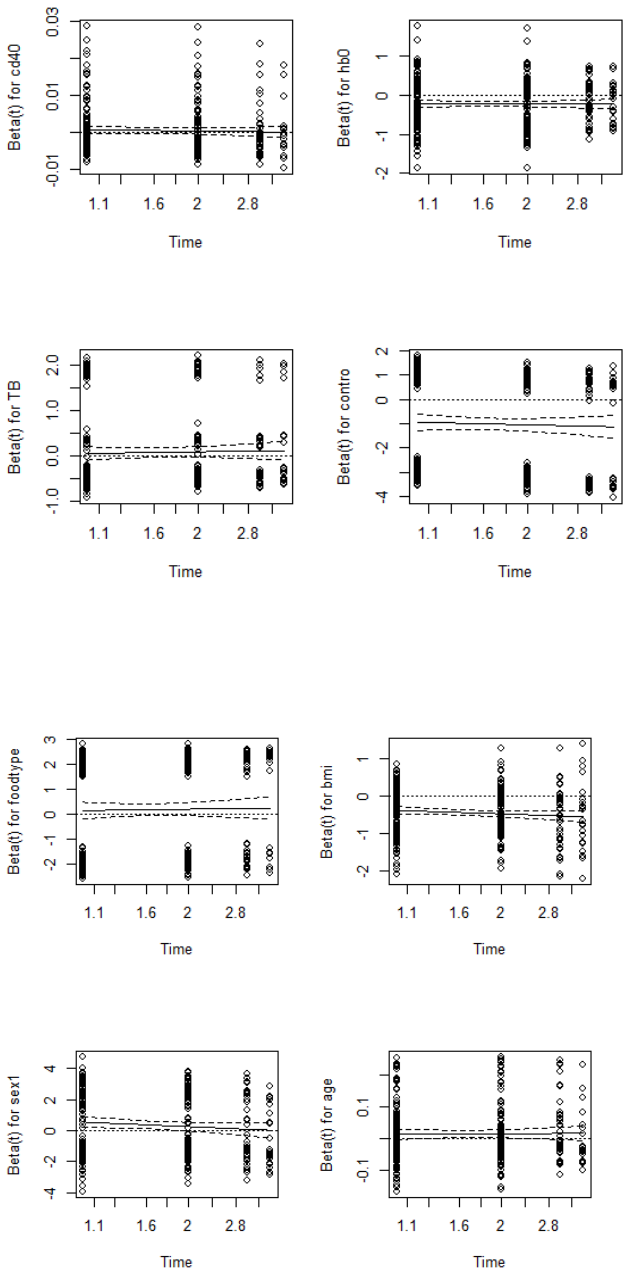

Figure 5: Presents Graphs for scaled Schoenfied Residuals

Figure 6 show homogeneity plots of residuals and Q-Q plots of residuals and of random effects for the mixed effects random model. The plots of residuals show that the assumptions of homogeneity and normality of the residuals have been met in the mixed random effects model.

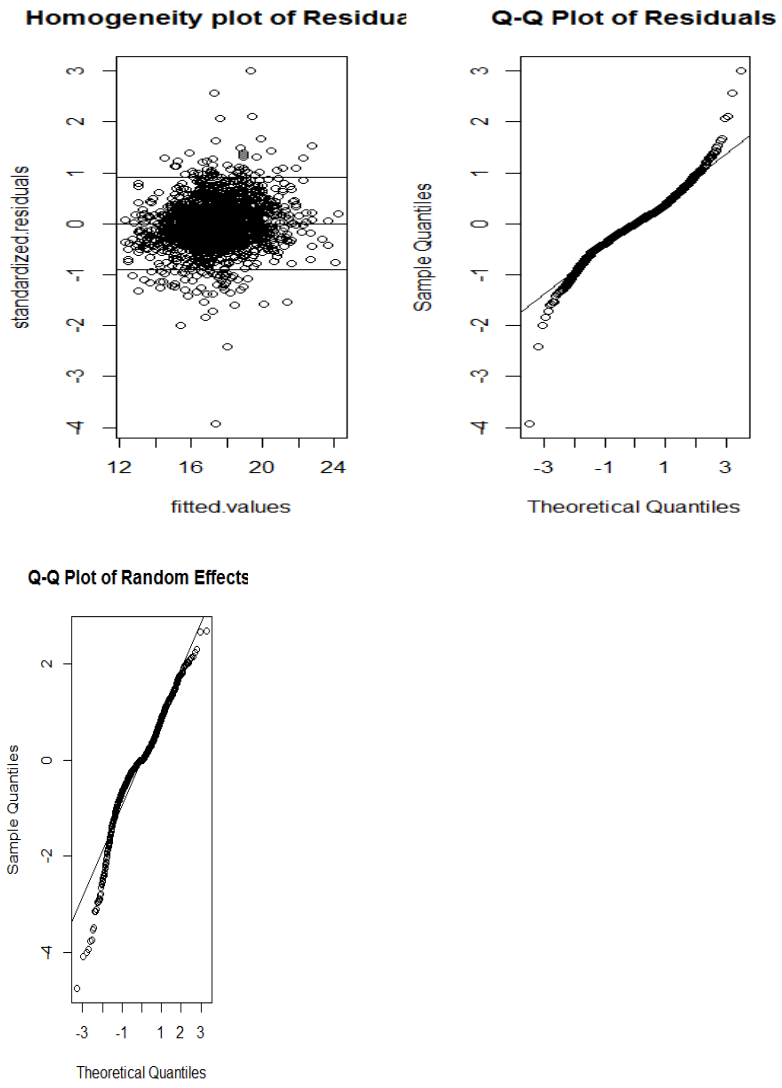

Figure 6: Homogeneity Plots of Residuals and of Random Effects for Mixed Effects Random Model

The Q-Q plot for longitudinal process of the joint model shows that the assumptions of normality of the random effects have been satisfied, See Figure 6.

CHAPTER 5 DISCUSSION

Two groups of HIV positive patient who were receiving different food supplements (CSB and RUTF) were compared using joint modeling approach and separate models. It has been that shown the repeated measurement variable, body mass index is significantly associated with time to an event, that is, survival of the patient. It has also been shown that separate and joint models give different results. It is likely that the results of joint modeling are more valid than single models, and that joint model gives smaller standard errors of the estimates (Henderson et al., 2000; Ibrahim, 2010; Nguti et al., 2005).

Corn Soya Blend and Fortified food did not have effect on the survival of the malnourished HIV positive patients. Men who are HIV positive and malnourished are at risk of dying as compared to women who are also HIV positive and malnourished. Body mass index of HIV positive patients who are malnourished has significant effect on the survival of patients. The lower the body mass index the higher the risk of death. Use of cotrimoxazole had an effect on the survival of patients. Patients who were not using cotrimoxazole had a higher risk of dying than patients who were using cotrimoxazole.

Comparison of Cox proportional time dependent covariate and the join models reveals some interesting features. In particular, body mass index, sex of person, hemoglobin level, and use of cotrimoxazole have effect on the survival of patient both in Cox proportional time dependent covariate and the joint models. Age of patient had

significant effect on the survival of the person in the Cox proportional time dependent covariate only and in joint model it was not significant. It has to be noted that joint models are known to produce unbiased estimated (Nguti et al., 2005), however in the joint model age was not significant on the survival of a person. Another difference noted, was that most of variables in the joint model had smaller standard errors of the estimates. This is an advantage of joint models over independent models (McCrink et al., 2011).

Comparison of longitudinal process and joint model reveals that some variables which were significant in the separate longitudinal model were not significant in joint model; and some variables which were not significant in the separate longitudinal model became significant in joint model. It was observed that variables like CD4 count and TB status did not significantly affect body mass index (BMI) in the separate model. However in joint model CD4 count and TB status had significant effect on body mass index. In the longitudinal sub model, variables like age and hemoglobin level had significant effect on the body mass index of patients in both separate and joint model. Also intercept and time were significant both in separate and joint models. Type of food supplement, sex of person did not have significant effect on the body mass index (BMI) in both separate and joint models.

In terms of standard errors, most of variables in the joint model had smaller standard errors as compared to the same variables in separate models. For instance, type of food supplement, CD4 count, sex, TB status, and use of cotrimoxazole had smaller standard errors in joint model than in separate models. This is in agreement with what other researchers found. For instance, Henderson et al. (2000) reported that joint

models produce smaller standard errors than separate models. Also in their study Nguti et al. (2005) reported that standard errors produced by joint models were smaller than standard errors in separate models. Nguti et al. (2005) further argued that the smaller the standard errors the better the results. Most of variable in joint models in the longitudinal sub model had shorter range of confidence intervals. Narrow confidence intervals are more desirable that wider confidence intervals.

CHAPTER 6 CONCLUSIONS

This chapter summarizes the study, gives some recommendations for analyzing survival data which has repeated measurements variables, and the limitations of the study.

6.1 Conclusion

Type of food supplement (CSB and RUTF) did not have effect on the survival of patients. The relationship between body mass index and survival of person living with HIV has been established. Body mass index has been shown to have significant effect on the time to death of malnourished patients. When time to an event and the repeated measurement variable are associated, separate models may produce biased results as compared to joint models. Joint models, on average may produce smaller standard errors. This research has shown joint models give better results than separate models, when there is association between the repeated measurement and time to an event variable.

6.2 Recommendations

When one has survival data with repeated measurement variable, and time to event is associated with repeated measurement variable, it is recommended that joint modeling of longitudinal and time to event data should be used.

Patients who are HIV positive and malnourished should be given food supplements in order to improve their body mass index. Patients who are HIV positive and

malnourished must be given treatment that increases levels of hemoglobin if they have low hemoglobin levels. Patients who are receiving ART and are malnourished must be given cotrimoxazole in order to reduce the risk of death.

6.3 Limitations of Study

The study used joint modeling of longitudinal and time to the event data. Longitudinal sub model used linear mixed effect regression model and the survival sub model used Cox proportional hazard model. The study used joint models with correlated error structures. The study did not use flexible models that use semi parametric or non parametric approach.

REFERENCES

Bangfalaki, T., Ganjali, M. & Berridge, D. (2014). Joint modeling of multivariate longitudinal mixed measurements and time to event data using a Bayesian approach. *Journal of Applied Statistics*, doi: 10.1080/02664763.2014.898132.

Bewick, V., Cheek, L. & Ball, J. (2004). Statistics review 12: Survival Analysis. *Crit Care,* 8(5), 389-294, doi: 10.1186/cc2955.

Brown, E. R., & Ibrahim, J. G. (2003a). A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics,* 59, 221-228.

Brown, E. R., & Ibrahim, J. G. (2003b). Bayesian approaches to joint cure rate and longitudinal models with applications to cancer vaccine trials. *Biometrics,* 59, 686-693.

Bycott, P., & Taylor, J. (1998). A comparison of smoothing techniques for CD4 data measured with error in a time-dependent Cox proportional hazards model. *Statistics in Medicine*, 17(18), 2061–2077.

Carpenter, J. R., & Kenward, M. G. (2007). *Missing data in clinical trials: A practical guide*. UK National Health Service, National Centre for Research on Methodology.

Chi, Y.Y. & Ibrahim, J. G. (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics*, 62(2), 432–45.

Collett, D. (2003). *Modelling Survival Data in Medical Research.* London: Chapman and Hall.

Cox, D. R. (1972). Regression models and life tables (with Discussion). *J. Roy. Statist. Soc.Ser. B* 34, 187-200.

Cox, D.R. & Oakes, D. (1984). *Analysis of survival data*. London: Chapman & Hall.

Dafni, U., & Tsiatis, A. (1998). Evaluating surrogate markers of clinical outcome when measured with error. *Biometrics,* 54(4), 1445–1462.

Dannhauser, A., van Staden, A.M., van der Ryst, E., Nel, M., Marais, N.,…& Erasmus, E. (1999). Nutritional status of HIV-1 seropositive patients in the Free State Province of South Africa: Anthropometric and dietary profile. *Eur J Clin Nutr,* 53,165-73.

Davis, C.S. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. New York: Springer-Verlag.

DeGruttola, V., & Tu, X. M. (1994). Modeling progression of CD-4 lymphocyte count and its relationship to survival time. *Biometrics,* 50, 1003-1014.

Dempster, A., Laird, N. & Rubin, D. (1977). Maximum likelihood from incomplete data via EM algorithm. *Journal of Royal Statistical Society Series B,* 39(1), 1-38.

Diggle, P. J., Heagerty, P., Liang, L. K., & Zeger, S. L. (2002). *Analysis of longitudinal data* (2nd ed.). New York: Oxford University Press.

Diggle, P.J., Sousa, I., Chetwynd, A.G. (2008). Joint modelling of repeated measurements and time-to-event outcomes: The fourth Armitage lecture. *Statistics in Medicine,* 27, 2981-2998.

Elashoff, R.M., Li, G., & Li, N. (2007). An approach to joint analysis of longitudinal measurements and competing risks failure time data. *Statistics in Medicine,* 26, 2813-2835.

Elashoff, R., Li, G., & Li, N.(2008). A joint model for longitudinal measurements and survival data in the presence of multiple failure types. *Biometrics,* 64,762-771.

Faucett, C.L., & Thomas, D.C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: A Gibbs sampling approach. *Statistics in Medicine*, 15, 1663-1685.

Gruttola, V. D., & Tu, X. M. (1994). Modelling progression of cd4 lymphocyte count and its relationship to survival time. *Biometrics*, 50 (4), 1003-1014.

Guo, X. & Carlin, B. (2004). Separate and joint modelling of longitudinal and time to event data using standard computer packages. *The American Statistician,* 58, 16–24.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. New Jersey: John Wiley & Sons, Inc.

Henderson, R., Diggle, P., Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics,* 1, 465–480.

Hogan, J.W., & Laird, N.M., (1997). Mixture model for the joint distribution of repeated measures and event times. *Statistics in Medicine*, 16, 239-257.

Hosmer, D.W., & Lemeshow, S. (1999). *Applied Survival Analysis Regression Modeling of Time to Event data*. New York: John Wiley & Sons, Inc.

Ibrahim, J.G., Chu, H., & Chen, L.M. (2010). Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology*, 28, 2796-2801.

Jennrich, R.I. & Schluchter, M.D.(1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 42, 805–820.

Kalbfleisch, J. D., & Prentice, R.L. (2002). *The statistical analysis of failure time data.* New York: Wiley.

Kaplan, E., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of American Statistical Association*, 53, 457- 481.

Kenward, M., & Carpenter, J. (2009). Multiple Imputation. In G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs (Eds.). *Handbooks of modern statistical methods. Longitudidal Data Analysis*. (pp. 447- 499). New York: CRC Press.

Klein, J.P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics*, 48, 795-806.

Laird, N.M., & Ware, J.H. (1982). Random effects models for longitudinal data. *Biometrics*, 38, 963-974.

Law, N. J., Taylor, J. M. G., & Sandler, H. M. (2002). The joint modeling of a longitudinal disease progression marker and the failure time process in the presence of cure. *Biostatistics* 3, 547-563.

Lesaffre, E., & Spiessens, B. (2001). On the effect of the number of quadrature points in logistic random-effects model: An example. *Applied Statistics*, 50, 325–335.

Lim, H.J, Mondal, P., & Skinner, S. (2013). Joint modeling of longitudinal and event time data: Application to HIV study. *Journal of Statistics and Informatics*, 1, 1.

Lin, H.Q., Turnbull, B.W., McCulloch, C.E. & Slate, E.H. (2002). Latent class models for joint analysis of longitudinal biomarker and event process data: application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association*, 97, 53-65.

Little, R. (2009). Selection and pattern-mixture models. In G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs (Eds.), *Handbooks of modern statistical methods. Longitudinal Data Analysis*. (pp. 409-431). New York: CRC Press.

Little, R.J.A. & Rubin, D.B. (2002). *Statistical Analysis with Missing Data (2nd ed.),* New York: Wiley.

Lipsitz, S., & Fitzmaurice, G. (2009). Generalised estimating equations for longitudinal data analysis. In G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs (Eds.), *Handbooks of modern statistical methods: Longitudinal Data Analysis*. (pp. 43-78). New York, USA: CRC Press.

Machin, D., Cheung, Y.B. & Parmar, M. (2006). *Survival Analysis: A Practical Approach.* West Sussex, England: John Wiley & Sons Ltd.

Manary, M., Ndekha,M., & van Oosterhout, J. (2010). Supplementary feeding in the care of the wasted HIV infected patient. *Malawi Medical Journal*, 22 (2), 46-49.

McCrink, L., Marshall, A., & Cairns, K. (2011). Joint Modelling of Longitudinal and Survival Data: A Comparison of Joint and Independent Models. Int. Statistical Inst.: Proc. 58th World Statistical Congress, 2011, 4971-4976, Dublin (Session CPS044) http://2011.isiproceedings.org/papers/950990.pdf. Accessed on January 2013.

Molenberghs, G., & Verbeke, G. (2005). Models for Discrete Longitudinal Data. Springer Science + Business Media, Inc.

Molenberghs, G., & Fitzmaurice, G. (2009). Incomplete data: Introduction and overview. In G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs (Eds.). *Handbooks of modern statistical methods. Longitudinal Data Analysis*. (pp. 395-408). New York, USA: CRC Press.

Molenberghs, G., Verbeke, G., & Kenward, M.(2009). Sensitivity analysis for incomplete data. In G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs (Eds.). *Handbooks of modern statistical methods. Longitudinal Data Analysis.* (pp. 501-551). New York, USA: CRC Press.

Nakai, M., & Ke, W. (2011). Review of the methods for handling missing data in longitudinal data analysis. *Journal of Math. Analysis*, 5 (1), 1-13.

Ndekha, M., Oosterhout, J., Zijlstra, E., Manary, M., Saloojee, H., & Manary, M. (2009). Supplementary feeding with either ready-to-use fortified spread or corn-soy blend in wasted adults starting antiretroviral therapy in Malawi: Randomised, investigator blinded, controlled trial. BMJ, 338:b1867 doi:10.1136/bmj.b1867.

Nguti, R., Burzykowski, T., Rowlands, D., & Janssen, P. (2005). Joint modelling of repeated measurements and event time: Application to performance traits and survival of lambs bred in sub-humid tropics. *Genet. Sel. Evol*, 37, 175–197.

Pawitan, Y. & Self, S. (1993). Modeling disease marker processes in AIDS. *Journal of American Statistical Association*, 88, 719-726.

Phillipson, P., Sousa, I., & Diggle, P. (2012). Joiner: Joint modeling of repeated measurement and time to event data. http://www.cran.r study.org/web/packages/joiner/vignettes/joiner.pdf , accessed on 10 December 2012.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & the R Core team (2009), nlme: Linear and Nonlinear Mixed Effects Models. http://www.cran.r-project.org/web/packages/nlme, accessed on 10 December 2012.

Prentice, R.L. (1982). Covariate Measurement Errors and Parameter-Estimation in a Failure Time Regression Model . *Biometrika*, 69 (2), 331-342.

Proust-Lima, C., Joly, P. & Jacqmin-Gadda, H.(2009). Joint modelling of multivariate longitudinal outcomes and a time-to-event: a nonlinear latent class approach. *Comput Stat Data Anal*, 53(4), 1142–1154.

Proust-Lima, C., Sene, M., Taylor, J.M.G. & Jacqmin-Gadda, H. (2012). Joint latent class models for longitudinal and time-to-event data: A review. *Statistical methods in medical research.*

Rabe-Hesketh, S., & Skrondal, A., (2008). *Multilevel and Longitudinal Modeling Using Stata* (2nd ed.). Texas: Stata Press.

Raboud, J., Reid, N., Coates, R. A., & Farewell, V.T. (1993). Estimating risks of progressing to AIDS when covariates are measured with error. *J. Roy. Statist. Soc. Ser. A* 156, 396-406.

Rizopuulos, D. (2010). JM: An R Package for the joint Modeling of Longitudinal and Time to event Data. *Journal of Statistical Software,* 35, 9.

Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3), 819–829.

Rizopoulos, D. & Ghosh, P. (2011). A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in medicine*, 30(12), 1366–1380.

Rizopoulos, D., Verbeke, G. & Lesaffre, E. (2009). Fully exponential Laplace approximation for joint modeling of survival and longitudinal data. *Journal of the Royal Statistical Society Series B,* 71, 637-654.

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

Rubin, D. B. (1987). *Multiple Imputation fornon response surveys*. New York: John Wiley & Sons.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Schafer, J.L., & Graham, J.W. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods*, 7, 147-177.

Schluchter, M.D. (1992). Methods for the analysis of informatively censored longitudinal data. *Statistics in Medicine*, 11, 1861-1870.

Self, S., & Pawitan, Y. (1992). Modeling a marker of disease progression and onset of disease. In N. P. Jewell, K. Dietz and V. T.Farewell (Eds.). *AIDS Epidemiology: Methodological Issues*. (pp. 231-255). Boston: Birkhauser.

Sene, M., Bellera, C.A. & Proust-Lima, C. (2014). Shared random effect models for joint analysis of longitudinal and time to event data: application to the prediction of prostate cancer recurrence. *Journal de la Societe Francaise de Statistique*, 155(1).

Song, X., Davidian, M., & Tsiatis, A. (2002). A semiparametric likelihood approach to joint modelling of longitudinal and time-to-event data. *Biometrics*, 58, 742-753.

Sousa, I. (2011). A review on joint modelling of longitudinal measurements and Time to event. *REVSTAT*, 9 (1), 57-81.

Stefanski, L. A., & Carroll, R. J. (1987). Conditional scores and optimal scores in generalized linear measurement error models. *Biometrika* ,74, 703-716.

Sweeting, M.J., & Thompson, S.G. (2011). Joint modeling of longitudinal and time to event with application to predicting abdominal aortic aneurysm growth and rapture. *Biometrical Journal*, 53(5), 750-763.

Therneau, T., & Grambsch, P. (2000). *Modelling survival data: Extending the cox model*. New York: Springer-Verlag.

Tseng, Y.K., Hsieh, F., & Wang, J.L. (2005). Joint modelling of accelerated failure time and longitudinal data. *Biometrika*, 92(3), 587–603.

Tsiatis, A. A., & Davidian, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*, 88, 447-458.

Tsiatis, A. A., & Davidian, M. (2004). Joint modelling of longitudinal and time to event data. *Statistica Sinica*, 14 , 809-834.

Tsiatis, A.A., Degruttola, V., & Wulfsohn, M.S. (1995). Modeling the relationship of survival to longitudinal data measured with error: Application to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*, 90, 27-37.

Van der Sande, M.A.B., van der Loeff, M.F.S., Aveika, A.A., Sabally, S., Togun, T.,…, & Sarge-Njie, R. (2004). BMI at time of HIV diagnosis: a strong and independent predictor of survival. *J Acquir Immune Defic Syndr* ,37,1288-94.

Verbeke, G., Molenberhs, G. & Rizopoulos, D. (2010). Random effects models for longitudinal data. In K. van Montfort, J. Oud, and A. Satorra (Eds.). *Longitudinal research with latent variables*. (pp. 37-96). Berlin: Springer – Verlag. doi: 10.1007/978-3-642-22.

Wang, Y., & Taylor, J. M. G. (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *J. Amer. Statist. Assoc.* 96, 895-905.

Williamson, P.R., Kolamunnage-Dona, R., Philipson, P. & Marson, A.G. (2008). Joint modeling of longitudinal and competing risks data. *Statistics in Medicine*, 27, 6426-6438, doi:10.1002/sim.3451.

Wu, L., Hu,X.J.,& Wu, H. (2008). Joint inference for nonlinear mixed-effects models and time to event at the presence of missing data. *Biostatistics*, 9(2) 308–320.

Wu, L., Liu, W. & Hu, X.J.(2010). Joint inference on HIV viral dynamics and immune suppression in presence of measurement errors. *Biometrics*, 66(2), 327–335.

Wu, L., Liu, W., Yi, G.Y. & Huang, Y. (2012). Analysis of longitudinal and survival data: Joint modeling, inferences, and issues. *Journal of Probability and Statistics*. doi:10.1155/2012/640153.

Wulfson, M.S., & Tsiatis, A.A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, 53, 330-339.

Xu, J., & Zeger, S.L. (2001). Joint analysis of longitudinal data comprising repeated measures and times to events. *Journal of Royal Statistics Society, Series C*, 50, 375-387.

Ye, W., Lin, X.H., & Taylor, J.M.G. (2008*).* A penalized likelihood approach to joint modeling of longitudinal measurements and time-to-event data. *Statistics and Interface*, 1, 33-45.

Zachariah, R., Fitzgerald, M., Massauoi, M., Pasulani O., Arnould, L., Makombe, S., & Harries, A.D. (2006). Risk factors for high early mortality in patients on ART in rural district of Malawi. *AIDS*, 20, 2355-2360.

Zeng, D.L., & Cai, J.W. (2005). Simultaneous modeling of survival and longitudinal data with an application to repeated quality of life measures. *Lifetime Data Analysis*, 11, 151-174.

APPENDICES

Appendix A: Commands used for data analysis in R

```
# Data input
> library(statmod)
> library(foreign)
> library(joineR)
> data <- read.dta("C:\\Users\\User\\Desktop\\Data\\DataLong.dta")
> names(data)

> data.long <- data[, c(1,2,16)]
> data.surv <- UniqueVariables(data, var.col = c("time_to_effect",
"status_death"), id.col = "sfhiv")
> data.baseline <- UniqueVariables(data, var.col =
c(3,4,7,8,10,11,12), id.col = "sfhiv")
> data.jd <- jointdata(longitudinal = data.long, survival =
data.surv, baseline = data.baseline, id.col = "sfhiv", time.col =
"time")
> summary(data.jd)
> plot(data.jd)
> take <- data.jd$survival$sfhiv[data.jd$survival$status_death == 0]
> data.jd.cens <- subset(data.jd, take)
> take1 <- data.jd$survival$sfhiv[data.jd$survival$status_death == 1]
> data.jd.uncens <- subset(data.jd, take1)
> par(mfrow=c(1,2))
> plot(data.jd.cens, Y.col ="bmi", main = "BMI: censored")
> plot(data.jd.uncens, Y.col ="bmi", main = "BMI: Failed")
> jointplot(data.jd, Y.col = "bmi", Cens.col = "status_death", lag =
8,col1 = "black", col2 = "gray", ylab = "BMI")
> jointplot(data.jd, Y.col = "bmi", Cens.col = "status_death", lag =
3,col1 = "black", col2 = "gray", ylab = "BMI")
> jointplot(data.jd, Y.col = "bmi", Cens.col = "status_death", lag =
2,col1 = "black", col2 = "gray", ylab = "BMI")
> jointplot(data.jd, Y.col = "bmi", Cens.col = "status_death", lag =
3,col1 = "black", col2 = "gray", ylab = "BMI")

> model.jointrandom <- joint(data.jd, bmi ~1+time+ foodtype + sex1 +
age + cd40 + contro + hb0 +TB, Surv(time_to_effect, status_death) ~
foodtype + sex1 + age + cd40 + contro + hb0 +TB, model = "int")
> summary(model.jointrandom)


> names(model.jointrandom)

> summary(model.jointrandom,variance = FALSE)
> model.jointrandom.se <- jointSE(model.jointrandom, n.boot = 100)
> model.jointrandom.se

# Fitting LME

> fit1=lme(bmi ~ time + foodtype + sex1 +cd40+ TB+ contro+hb0+age,
random = ~1|sfhiv ,data = data, na.action =na.omit)
> summary(fit1)
```

```
> fit2=lme(bmi ~ time + foodtype+ sex1 +cd40 + TB+ contro+hb0+age,
random = ~time|sfhiv ,data = data,na.action=na.omit)
> summary(fit2)

> fit3=lme(bmi ~ time + foodtype + sex1 +cd40 + TB+ contro+hb0+age,
random = ~(time+I(time^2))|sfhiv ,data = data, na.action=na.omit)
> summary(fit3)

# comparing models

> anova(fit1, fit2)
> anova(fit1, fit3)
> anova(fit2, fit3)
> anova(fit2, fit1)
> qqnorm(fit1, ~ranef(.))
> qqnorm(fit2, ~ranef(.))
> qqnorm(fit3, ~ranef(.))

# Plotting Kaplan Meier Graph

>            fit4=survfit(Surv(time_to_effect,status_death)~foodtype,
type="kaplan-meier", data=c(data.jd$survival,data.jd$baseline))
>plot(fit4,lty=c(1,2),mark.time=TRUE,col=c("red","blue"),xlab="years"
, ylab="Survival")
>plot(fit4,lty=c(1,2),mark.time=FALSE,col=c("red","blue"),xlab="years
", ylab="Survival")
> legend(6,0.2,c("CSB", "RULTF"),lty=c(1,2),col=c("red","blue"))

# Kaplan Meier graph
>         survdiff(Surv(time_to_effect,status_death)         ~foodtype
,data=c(data.jd$survival,data.jd$baseline))
>           survdiff(Surv(time_to_effect,status_death)           ~sex1
,data=c(data.jd$survival,data.jd$baseline))
>          survdiff(Surv(time_to_effect,status_death)          ~contro
,data=c(data.jd$survival,data.jd$baseline))

# Fitting Cox proportion hazard model

>              fit5=coxph(Surv(time_to_effect,status_death)~foodtype,
data=c(data.jd$survival,data.jd$baseline))
> summary(fit5)
Call:

>  fit5=coxph(Surv(time_to_effect,status_death)~foodtype  +  sex1  +age
+contro+ cd40 +hb0, data=c(data.jd$survival,data.jd$baseline))
> summary(fit5)

> cox.zph(fit5, transform="identity")

> ran=random.effects(fit2)
> U=ran[,1]+ran[,2]
>            fit7=coxph(Surv(time_to_effect,status_death)~foodtype+U,
data=c(data.jd$survival,data.jd$baseline))
> summary(fit7)

> fit10 <- joint(data.jd, bmi ~1+time+ foodtype + sex1 + age + cd40 +
contro  +  hb0  +TB,  Surv(time_to_effect,  status_death)  ~  foodtype  +
sex1 + age + cd40 + contro + hb0 +TB, model = "intslope")
> summary(fit10)
> fit10.se<- jointSE(fit10,n.boot=100)
>  fit10.se
```

```
# Putting in a format for Cox model for Time dependent covariates

> datacox <- read.dta("C:\\Users\\User\\Desktop\\Data\\datacox.dta")
> sum(!is.na(datacox[,25:29]))
[1] 2022
> datacox.2 <- matrix(0, 2022, 28) # to hold new data set
>    colnames(datacox.2)    <-c('start',    'stop',    'event.time',
names(datacox)[1:24], 'bmi')

> row <-0 # set record counter to 0
> for (i in 1:nrow(datacox)) { # loop over individuals
+ for (j in 25:29) { # loop over 14 weeks
+ if (is.na(datacox[i, j])) next # skip missing data
+ else {
+ row <- row + 1 # increment row counter
+ start <- j - 25 # start time (previous week)
+ stop <- start + 1 # stop time (current week)
+ event.time <- if (stop == datacox[i, 1] && datacox[i, 2] ==1) 1
else 0
+ # construct record:
+ datacox.2[row,] <- c(start, stop, event.time, unlist(datacox[i,
c(1:24, j)]))
+ }}}
> datacox.2 <- as.data.frame(datacox.2)
> remove(i, j, row, start, stop, event.time) # clean up

# Fitting Cox model with Time dependent covariates

> mod.cox.4 <-coxph(Surv(start, stop, status) ~ foodtype + bmi +sex1+
age + cd40 + hb0 + TB + contro,data=datacox)
> summary(mod.cox.4)

# Model Building: Linear Mixed Effects Regression model

# Data input
>                              model.mixed                                <-
read.dta("C:\\Users\\User\\Desktop\\Data\\DataLong.dta")
> library(lattice)
> library(nlme)
> attach(model.mixed)

#Distribution for body mass index

> hist(bmi , col =" darkgray ")
>  xyplot(bmi ~time ,data , type ="l", xlim =c(1 ,3) , main ="A
Trajectory plot of Body Mass Index ")
> xyplot ( bmi ~ time |sfhiv ,model.mixed, type ="l", subset =(sfhiv
< 31) ,strip =FALSE , main =" Individual plots (for the first 30
Patients )")
> mod.mixed.1 <-lme( bmi~ foodtype + sex1 + time + cd40 + hb0+ TB +
contro,  random  =  ~  time  |sfhiv  ,  data  =  model.mixed  ,
na.action=na.omit)
> summary(mod.mixed.1)
> intervals(mod.mixed.1)

#Testing normality of random effects

> par(mfrow =c(1 ,2))
```

```
> fitted.values <- fitted(mod.mixed.1)
> standardized.residuals <-residuals(mod.mixed.1)
> plot(fitted.values , standardized.residuals , main ="Homogeneity
plot of Residuals")
>         abline(h=c(-1.96*         sd(standardized.residuals),0,1.96*
sd(standardized.residuals)))
> qqnorm(standardized.residuals , main ="Q-Q Plot of Residuals")
> abline(0, sd(standardized.residuals))
> eblups <-as.vector(unlist(ranef(mod.mixed.1)))
> qqnorm(eblups , main ="Q-Q Plot of Random Effects ")
> abline(0, sd(eblups))
```