MODELLING ASSOCIATION BETWEEN COMPETING RISKS SURVIVAL AND LONGITUDINAL OUTCOMES IN RANDOMIZED CONTROLLED TRIAL FOR MALARIA INTERVENTIONAL STUDIES USING JOINT MODELS

MASTER OF SCIENCE. (BIOSTATISTICS) THESIS

DANGER DIDJIER GWEDEZA MASANGWI

University of Malawi
Chancellor College

NOVEMBER, 2019

MODELLING ASSOCIATION BETWEEN COMPETING RISKS SURVIVAL AND LONGITUDINAL OUTCOMES IN RANDOMIZED CONTROLLED TRIAL FOR MALARIA INTERVENTIONAL STUDIES USING JOINT MODELS

M.ASTER OF SCIENCE BIOSTATISTICS

By

DANGER DIDJIER GWEDEZA MASANGWI
BSc.(Statistics) -

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

November, 2019

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.

DANGER DIDJIER GWEDEZA MASANGWI

Full Legal Name

Signature

Date

CERTIFICATE OF APPROVAL

The undersigned certify that this thesis represents the students own work and effort and has been submitted with our approval.

Signature:_____ Date: _____

Mavuto Mukaka, PhD (Professor)

Supervisor

Signature: _____ Date: _____

Adamson Muula, PhD (Professor)

Co-supervisor

DEDICATION

This work is dedicated to Jehovah, the giver of wisdom and understanding, My parents, Mr and Mrs Masangwi for the words of hope even in the hardest times. I, also, dedicate the work to DELTAS SSACAB for all academic support.

ACKNOWLEDGEMENT

With grateful heart, let me express gratitude for the success of this projcet to God, the giver of wisdom and understanding be glory. The presence and precious commitment of supervisors: Prof. Mavuto Mukaka, and Professor Adamson Muula have made this paper a reality. I treasure the support . I take you for tireless experts, committed to grooming the upcoming world of scientists. You are always an inspiration to the young faculty, continue with the spirit and the pace.

To DELTAS Sub-Saharan Africa Consortium for Advanced Biostatistics (S2ACAB), I am extremely thankful for the financial support both for academic and upkeep throughout the study of this programme. You have exposed me to the world of experts in Biostatistics in Africa and beyond.

Let me also extend a vote of thanks to Mr Tsiridzani Kaombe for the encouragement and motivation. To Elicy, I thank you for the encouragement and support throughout. Classmates, I thank you all, for your contribution in our time of study.

ABSTRACT

Biomedical studies may collect longitudinal and survival data in follow-up malaria studies. In randomized controlled trials in malaria interventional studies the longitudinal and survival data are analyzed separately (mixed-effect and Cox Models), yet the longitudinal outcomes may be important predictors in the survival outcomes. Standard methods for survival analysis, cannot be considered with such longitudinal outcomes. In such studies, survival process may also include multiple events (competing risks), implying that three blocks, survival, longitudinal and competing risks need to be considered in the analysis. In order to assess the association between the malaria longitudinal and the survival outcomes collected in biomedical studies, joint modelling framework was considered to combine the three blocks in the analysis. Joint models were also compared to separate models. Different survival outcomes observed were severe malaria (4.95%), withdrawal (10.89%) and censored (84.16 %). The time-dependent haemoglobin level and parasite count were not associated with the risks of severe malaria and withdrawal in the extended-time dependent Cox model. The true longitudinal markers parasite counts and haemoglobin levels were associated with the risk of severe malaria *($p < 0.0001$)* and *($p = 0.029$)* respectively and had no effects on the risk of withdrawal in the joint models as these markers change with time. Generally the separate models were the best fit to the malaria dataset than the joint models due to lack of association between the survival outcomes and longitudinal outcomes in the cause-specific time dependent hazard model.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS AND ACRONYMS

AIDS : Acquired Immuno Deficiency Syndrome

ANOVA : Analysis of Variance

AQ : Amodiaquine

AR : Auto-correlation

ART : Artesunate

BIC : Bayesian Information Criterion

CIC : Cumulative Incidence Curves

CPMs : Covariance Pattern Models

CQ : Chloroquine

CR : Competing risk

GEE : Generating Estimating Equlations

GLM : Generalized Linear Models

MANOVA: Multiple Analysis of Variance

MCAR : Missing Completely at Random

MNAR : Missing not at Random

MRM : Mixed-effects Regression Model

PANSS : Positive and Negative Symptom Rating Scale.

RMLE : Restricted Maximum Likelihood Estimation

SP : Sulfadoxine-Pyrimethamine

TB : Tuberclosis

CHAPTER 1

INTRODUCTION

1.1Background

Biomedical studies may collect repeated measurements of longitudinal data and time to event/events of interest data during follow-up. A typical example is the AIDS study where CD4 count and viral load are collected longitudinally and the time to AIDS or death is also monitored (Elashoff *et al.,* 2008). Another example is in cancer studies where the longitudinal data and time to event data are collected for each subject. The longitudinal data such as circulating tumor cells, immune response to a vaccine, a genetic biomarker, or a health outcome are recorded (Ibrahim *et al*, 2010). Yet another example is prostate cancer study, where patients are followed-up over time and during that period death or metastasis can occur. In malaria studies, randomised blinded trials are carried out to compare efficacy, and safety of drugs and resistance of parasites. During follow-up, one of the measures of interest, may be time to fever resolution, time to parasite clearance, with possible longitudinal covariates like white blood cell count or red blood cell counts and changes in haemoglobin levels.

In these follow-up studies various outcomes are possible. To analyze such data, there are methods for separate analysis of longitudinal data and survival data. For example, in survival data, survival methods correctly incorporate information from both censored and uncensored observations in estimating important model parameters. Early work in survival analysis dates back to 1958, where a non-parametric estimator of survival

function is proposed. Non-parametric methods such as Kaplan-Meier product-limit estimator and life tables are used to estimate the survival curves. Parametric methods such as Weibull, exponential and Log-normal and log–logistics are widely applied in survival data (Collet, 2003). In order to analyze the effects of covariates on time-to-event, methods such as Cox proportion hazard model, and extended Cox model for time-dependent covariates are used. The problem with Cox model for analysis of survival data is that it is only theoritically valid for exogenous time-varying covariates but not when studying biomarkers (endogeneous) and other patient parameters (Andrinopoulou, 2014). The reason behind this inadequacy of Cox model is that it assumes that from one visit to another, the marker's level remains constant and that a sudden change in the levels occurs when the patients come for a visit, and this leads to a crude appromiximation about the path of the biomarker. Ignoring these special characteristics and fitting the extended Cox model, would result in bias for the estimated effect of a biomarker (Andrinopoulou, 2014).

In cases, where there are more than one failures (competing risks), intepretation of survival probabilities has always been questionnable (Kleinbaum & Klein, 2005). There have always been problems in the analysis such as estimation of the relationship between covariates and rate of occurence of failures of specific types, study of interrelation between failure types under a specific set of study conditions and the estimation of failure rates for certain types of failure given the removal of the other failure types (Kalbfleisch & Prentice, 2002). Methods such as cause-specific Cox model and cumulative incidence curves have been developed and used in such situations. A common assumption in all these models is that censoring is noninformative for survival

data with a single failure type, which is no longer applicable in presence of informative censoring.

In longitudinal data, methods such as generalized mixed-effects regression models, Covariance pattern models, ANOVA, Generalized Estimating Equations (GEE) are used to analyze the repeated measurements with possible covariates (Hedeker & Gibbons, 2006). The impressive feature about these models is that of explicit account for the correlation within the measurements obtained from the same patients and can handle unequally spaced visit times (Andrinopoulou, 2014). A major challenge for analysis of longitudinal outcomes is the fact that measurements for the outcomes are incomplete (missing). Missing data in longitudinal studies arise from a variety of reasons. The main concern in longitudinal analysis with missing data arises when there is an association between the longitudinal profile and the missing process. Mechanisms such as Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR) are commonly encoutered in longitudinal studies (Andrinopoulou, 2014). However, in literature, methods for handling missing data in longitudinal settings are considered, including selection models, pattern mixture models and shared parameter models for discrete times (Molenberghs & Kenward, 2007). Joint distribution methods applicable for continuous time are applied to analyse the longitudinal and missing processes (MNAR) since in reality patients skip visits or dropout from the study.

In biomedical studies that collect longitudinal data such as malaria studies, covariates can be important predictors of survival or some other time-to-event. In survival data analysis, censoring is assumed to be noninformative, yet longitudinal response data is

affected by informative dropout, especially in cases with competing risks and also the inclusion of time-varying covariates in survival analysis. This suggests need to investigate the relationship between the longitudinal biomarkers and time to event of interest. Thus the aim of this project is to assess the relationship between longitudinal and competing risks survival malaria outcomes using joint models.

1.2 Problem Statement

Longitudinal and survival data collected together in biomedical studies are analyzed independently, regardless of any possibility of relationship between these data (Ibrahim *et al*.,2010). As such, the methodologies for analysis are not sufficient for they do not account for other parameters, and the association between the two types of the data. Using separate methods and ignoring special features of longitudinal and survival data, may lead to underuse of potential variable information and lead to biased results and conclusion about treatment effects (Sudell *et al*.,2016). Hence, need for joint modelling approach to model the association between the two data sets. Also, many studies on joint modelling of longitudinal and survival data in literature analyze and model the data with one failure event but not much in presence of competing risks (multiple failure events). To add to the body of knowledge in this subject area, with an application to malaria data, this project is undertaken.

1.3 Objective

In particular the aim of the study is to assess association between malaria longitudinal and competing risks survival outcomes using joint models.

1.3.1 Specific objectives

Specifically for this project, the emphasis was on the following objectives:

- Investigate association between baseline covariates and longitudinal and survival malaria outcomes using separate models, and competing risks joint models.
- Assess performance of joint models and separate models for malaria longitudinal and competing risks survival outcomes in randomized controlled trial.

CHAPTER 2

STATISTICAL BACKGROUND

In this chapter, an introduction of two aspects that help understand the joint modelling process of longitudinal and survival data are presented. In the first case, survival data and methods for analysis of such data are introduced. The second part of this chapter presents the longitudinal data and the statistical methods for the analysis of repeated measurements. These two blocks, lead to an introduction of joint models for analysis of longitudinal and survival data even in competing risks settings.

2.1 Survival Data

Survival analysis is a collection of statistical procedures for data analysis, for which the outcome variable of interest is time until an event occurs (Singh & Mukhopadhyay, 2011). Over the last few decades, since the World War II motivated the study in the reliability of the military equipment, the survival analysis has been a very important field of research . It studies the time until an event of particular interest occurs, and with it, it answers questions such as what kind of treatment is better for a certain illness, or what variables have an influence in the recovery of a patients (Hevia, 2014). Initial studies on survival analysis had an interest on death as an event of interest. In modern world, survival data extend to time until onset of a disease, time until stock market crash, time until an equipment failure in engineering, time until earth quake and so on (Smith, 2002). The event of interest is usually called *failure* and the variable is called *failure time or survival time.*

Survival analysis has become one of the most frequently used methods for analyzing data in disciplines ranging from medicine, epidemiology, and environmental health, to criminology, marketing, and astronomy (Lee & Go, 1997). In oncological studies, time from diagnosis to death from any reason, time to tumor recurrence and the time from diagnosis to tumor-related death are of interest (Zwiener, Blettner, & Hommel, 2011). In business analytics, important outcomes such as time until a warranty claim, time from initial sales contact to a sale and time from employee hire to either termination or quit are analysed using survival techniques. Another example in clinical trials is time to treatment failure in TB patients, time until AIDS for HIV patients and time until cardiovascular death after some treatment intervention, (Elashoff, Li, & Li, 2008).

More generally, biomedical studies have been a root that inspires the biostatistics field. They provide data with specific features that need special caution when doing the analysis, and they keep coming up with situations where new statistical tools have to be developed in order to be able to handle them (Hevia, 2014). Due to increased biomedical research, survival data is more prevalent hence need for dedicated statistical methods for analysis of such data for better analysis results and action. For example, in clinical trials, to compare survival times of patients who receive one or other of the treatment types, it is important to explore the relationship between the potential predictors to survival or hazard of an event of interest. The resulting estimates could be particularly useful in devising a treatment regimen and in counselling the patients about their prognosis (Collet, 2003).

As shown above, survival data are common and collected in various fields. In this study the focus is on application of survival analysis to biomedical fields. Another important

note on survival data is that the data are collected in follow-up studies, where patients are followed until time when an event of interest occurs. In such studies not all subjects have the event of interest observed, such individuals are said to be censored. The following subsection presents censoring concept as is used in biomedical studies that collect survival data.

2.1.1 Censoring in Surviva l Data

In follow-up studies, not all subjects experience the event of interest. This may be due to loss to follow-up, or the study ends before the event of interest is observed. For these reasons, part of the event of interest still remains unobserved, and as such the event time is said to be *censored.* For example, suppose that a patient is recruited to a clinical trial, where the outcome of interest is time to death from a certain cause. Suppose that the patient moves to a different location or country and is no longer traced. The only information available on the survival experience of this patient is the last date on which he or she was known to be alive, which may be the last visit to the clinic. In such cases, the death for this individual will not be observed and his or her survival time is censored.

Censoring is classified as left or right (*when survival time is less or greater than the observation time*). There is also interval censoring where time to event of interest is believed to occur between some time points. Censoring process can also be classified as informative or noninformative.

The distinction between informative censoring and noninformative censoring is that, the former occurs when subjects withdraw from the study with reasons related to the expected failure time while the latter, the reasons for drop-out are independent of the study. Censoring makes the survival data more skewed and as such standard statistical methods such as *t-test,* linear models and others are not appropriate for this type of data (Collet, 2003). This skewness behaviour of survival data leads to specific and dedicated methods for analysis of survival data.

In order to summarize survival data, it is required to define two important functions. The following section presents two central functions of interest in the analysis of survival data.

2.2 Important Functions in the Analysis of Survival Data

In summarizing the survival data, there are two functions of central interest namely, the survival function and hazard function. These functions are therefore defined in the next sub-section.

2.2.1 Survival function

Let T be a non-negative random variable for event time. In survival analysis , a subject i is represented by the pair $(T_i , \propto_i)$ where T_i is the observed event time for subject i and $T_i = (T, C_i)$ with C_i as the censoring time and $\propto_i$ as the censoring indictor with $\propto_i = 1$ if censored and $\propto_i = 0$ otherwise. The survival function is defined as the probability of surviving time t or probability that an event occurs after an instant t. Assuming that T is a continuous random variable, with F(t) as probability distribution of T, then the survival function is defined as:

$S(t) = P(T \geq t) = 1 - F(t) = 1 - \int_t^{\infty} f(u)\, du$ for $t \geq 0$

$f(.)$ is the corresponding probability density function. This function is therefore used to represent the probability that an individual survives from the time origin to some time beyond *t.*

$S(t)$ must be a decreasing function with $S(0) = 1$ and $\lim_{t\to\infty} S(t) = 0$. In general, the function $S(t)$ provides useful summary information such as the median or quantile survival times.

2.2.2 Hazard function

Another important function in survival analysis is the hazard function which expresses the risk or hazard of an event at some time t, and is obtained from probability that an individual fails at time t conditional on having survived to that time with T lying in the interval $[t, t + \delta t]$. The hazard function is defined as:

$\lambda(t) = \lim_{\delta t\to 0} \frac{P(t \leq T < t + \delta t | T \geq t)}{\delta t}$ for $t > 0$.

Note that $\lambda(t)\delta t$ is the approximate probability that an individual fails in the interval

$[t, t + \delta t]$ conditional on individual having survived to time *t.* The hazard function is also called the *instantaneous rate of failure* or *intensity rate*.

The hazard function can be expressed in terms of survival function, likewise survival function can be expressed in terms of the instantaneous rate of failure function. Using the standard results in probability theory, the conditional probability of an event A given B is written as:

$P(A|B) = \frac{P(A \cap B)}{P(B)}$.

The conditional probability in the hazard function can be written as:

$\frac{P(t \leq T < t+\delta t)}{P(T \geq t)} = \frac{F(t+\delta t)-F(t)}{S(t)}$.

The hazard function, will then be defined as:

$\lambda(t) = \lim_{\delta t \to 0} \left(\frac{F(t+\delta t)-F(t)}{\delta t}\right) \frac{1}{S(t)}$, but $\lim_{\delta t \to 0} \left(\frac{F(t+\delta t)-F(t)}{\delta t}\right) = F(t)' = f(t)$, thus

$\lambda(t) = \frac{f(t)}{S(t)}$.

From this relationship, it follows that:

$\lambda(t) = \frac{-d(\log(S(t))}{dt}$ and $S(t) = \exp(-\int_0^t \lambda(s)\, ds = \exp(-\Lambda(t))$, where $\Lambda(t)$ is the cumulative or integrated hazard.

2.3 Methods for Analysis of Survival data

The first step in the analysis of a set of survival data is to present the numerical or graphical summaries of the survival times for individuals in a particular group in terms of survival and hazard probabilities. Survival data are conveniently summarized through estimates of the survival function and hazard function. This is achieved through the use of non-parametric, semi-parametric and parametric methods for analysis of survival data. In the next sections, the focus is on describing these methods.

2.3.1 Non-parametric methods

These methods are also called *distribution-free*, because they do not require specific assumptions about the distribution of undelying survival times. The most well known functions for estimation of survival and hazard functions are *Kaplan-Meier* , *Nelson-Aelen* estimators and *life tables*. Other non-parametric methods for comparision of

groups of survival data will be introduced later. Here, Kaplain-Meier and Nelson-Aelen estimators of survival and hazard functions are discussed.

Kaplan-Meier Estimator

This was proposed by Kaplan and Meier (1958). It is also called *product-limit estimate* of the survival function. The derivation of Kaplan-Meier is done using the following steps: Suppose that there are n individuals with observed survival times $t_1 < t_2 < t_3, \ldots, < t_n$. Some of these observations may be right-censored and there may be more than an individual with the same survival observed times. Let r_i be those at risk of event at time t_i and d_i be those that have failed at t_i. The Kaplan-Meier estimator assumes that the distribution is discrete instead of continuous, with the events only occurring at these observed time points. The probability that an individual fails at t_i is denoted $\hat{\lambda}(t_i) = \frac{d_i}{r_i}$, as the estimated hazard at time t_i and the corresponding estimated survival probability is given by: $\hat{S}(t_i) = \frac{r_i - d_i}{r_i}$.

Under the assumptions that censoring is non-informative and that individuals fail independently, the probability of surviving at any time t can be written as the product of the conditional probabilities:

$$P(T > t) = P(T > t | T > t-1) P(T > t-1)$$

$$= P(T > t | T > t-1) P(T > t-1 | T > t-2) \ldots,$$

By repeating this method, the estimated survival function at any time t, is a Kaplan-Meier survival estimator:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(\frac{r_i - d_i}{r_i} \right)$$

A plot of the Kaplan-Meier is a step function in which the estimated survival probabilities are constant between adjacent event times and decrease at each event time.

Breslow and Crowley (1974) and Peterson (1977), proved the consistency of this estimator and had shown that $\sqrt{n}(\hat{S}(t) - S(T))$ converges in law to Gaussian process with mean 0 and variance-covariance structure (Hevia, 2014). The variance for $\hat{S}(t)$ is estimated using Green's formula,

$$Var\left(\hat{S}(t)\right) = \hat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{r_i(r_i - d_i)}$$

Nelson-Aalen estimator of survival function

This is an altenative estimator of survival function which is based on the individual event times. It is obtained from an estimate of the cumulative hazard function. It is also known as Alt-shuler's estimate. This was proposed by Bleslow (1972).

$$\hat{S}\,(t) = \exp\big(-\Lambda(t)\big) = \prod_{i=1}^{k} \exp\left(-\frac{d_i}{r_i}\right)$$

with d_i and r_i defined as in the derivation of Kaplan Meier estimator. The estimator has shown to perform better than Kaplan-Meier especially with small samples. However, the estimates are asymptotically equivalent particularly at the earlier survival times.

Having introduced the Kaplan-Meier and Nelson-Aelen methods for estimating the survival probabilities, the non-parametric methods for comparing the survivor probabilities in different groups of survival data are now considered.

2.3.2 Comparison of two groups of survival data

In clinical trials , it is common to randomise subjects to different treatments under study. Survival experiences of patients in different groups may really differ suggesting the need to consider the treatments, or the differences may not be there in such that the observed differences are merely due to chance variation. In order to help distinguish between these two explanations, non-parametric methods are applied. The basic approach to compare two groups of survival data is to plot the corresponding estimates of the two survival functions on the same axis, and the resulting plot can be informative. This idea, is presented by (Collet, 2003) shown in figure 1.

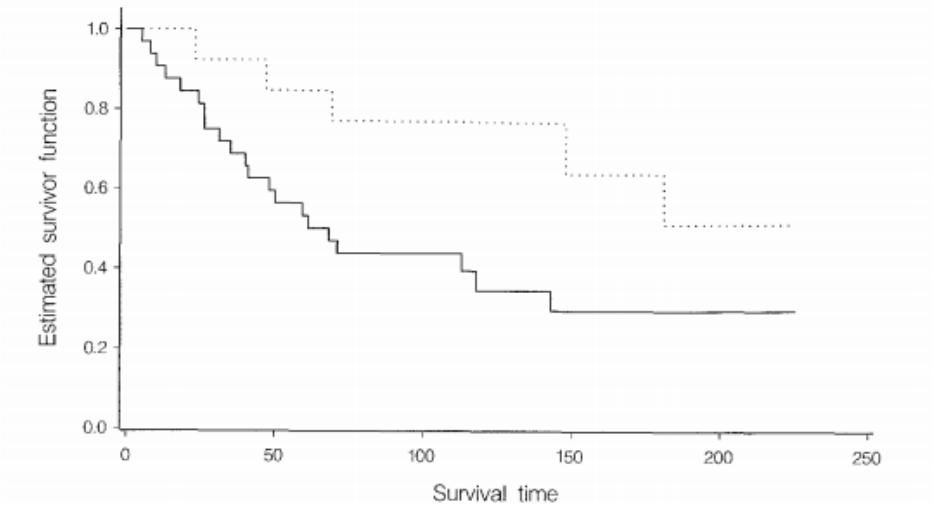

Figure 1*: Kaplan-Meier estimate of the survival functions comparing two groups of women turmours survival data: positively stained (___) and negatively stained (....).*

Figure 1 indicates that the survival function for women with negatively stained turmours is greater than that of women with positively stained turmors. This indicates that the result of staining may be a useful indicator in prognosis. However, the plot does not quantify the extent of between-group differences. As a result, non-parametric procedures such as *log-rank test* , *Wilcoxon test* , *Tarone-ware, Fleming's Harrington, Cox's F-test,* and *Gehan's Generalized Wilcoxon* are used. In this section, log-rank test and Wilcoxon test are considered.

The log-rank test

Suppose there are two groups of survival data namely group1 and group2. Let $t_1 < t_2 < t_3, \ldots, < t_n$ be ordered event times across the groups. At time t_i, denote number of individuals who fail in group1 be d_{1i} and number of individuals that fail in group2 be d_{2i}. Let n_{1i} and n_{2i} be individuals at risk at time t_i in group1 and group2 respectively. Then n_i is the total number of individuals at risk at time t_i. More information is summarized in Table 1.

Table 1: Number of events at the i-th event time in each of the two groups

Group	**Number of events at time t_i**	**Number surviving beyond t_i**	**Number at risk just before t_i**
1	d_{1i}	$n_{1i} - d_{1i}$	n_{1i}
II	d_{2i}	$n_{2i} - d_{2i}$	n_{2i}
Total	d_i	$n_i - d_i$	n_i

Fixing the marginal totals in the table, and under the hypothesis that the survival is independent of group, the four entries are determined by the value d_{1i} which has a *hypergeometric distribution*, with mean:

$$e_{1i} = \frac{n_{1i}d_{1i}}{n_i} \text{ and variance } v_{1i} = \frac{n_{1i}n_{2i}d_i(n_i - d_i)}{{n_i}^2(n_{i-1})}.$$

Combining the information in the table to get overall deviations and corresponding variance gives a statistics:

$$W_L = \frac{U_L^2}{V_L} \sim \chi_1^2 \text{ , where } U_L = \sum_{i=1}^{r}(d_{1i} - e_{1i}) \text{ and } V_L = \sum_{i=1}^{r} v_{1i}.$$

This method was proposed by Mantel and Haenszel (1959). A test based on this statistic is called Mantel-Cox or Peto-Mantel-Haenszel.

The statistic is called log-raank test since it is derived from the ranks of the survival times and the resulting rank statistic is based on logarithm of the Nelson-Aalen estimate of survival function. The statistic summarizes the extent to which the observed survival times deviate from the expected under the hypothesis of no group differences.

Wilcoxon Test

This is also called Bleslow test and is used to test the null hypothesis that there is no difference in survival functions of the two groups of the survival data. It is based on the statistic

$$U_W = \sum_{i=1}^{r} n_i(d_{1i} - e_{1i})$$

where the differences are weighted by n_i (total number of individual at risk at time t_i). The variance for this statistic is estimated by

$$Var(U_W) = \sum_{i=1}^{r} {n_i}^2 v_{1i} = V_W$$

with v_{1i} , d_{1i} , e_{1i} as defined in the log rank test above. The corresponding test statistic is then

$$W_W = \frac{{U_W}^2}{V_W} \sim \chi_1^2$$

In order to compare more than two groups of the survival function, extensions are made to both Wilcoxon and logrank tests.

The logrank test is more suitable when an altenative to null hypothesis of no difference in survival functions between groups is that the hazard for an individual in one group at any time is proportional to the hazard for a similar individual in the other group at the same time. This hypothesis is called *proportional hazard*, a useful assumption in modelling survival data. In case of other deviations from null hypothesis, Wilcoxon test

is more appropriate. Having looked at non-parametric procedures for analysis of survival data, next are semi-parametric methods.

2.4 Semi-Parametric Regression Methods for Censored Survival data

The non-parametric methods provided above, can be useful in analysis of single group or survival data, or making comparison for different groups of survival data. In medical studies that give rise to survival data , it is important for example to record demographic variables such as age, sex of the patient and other physiological variables such as blood volume, haemoglobin levels, and heart rate. It may also be important to record the lifestyle of the patients such as smoking , physical exercise and dietary behaviours. For example, in a clinical trial involving two treatments for prostate cancer, the primary aim is to compare the survival experience of patients in the treatment arms (Collet, 2003). However, variables such age of the patient, size of the turmour are recorded and may likely influence the survival times. As such it will be imperative to take account of these variables when making assessment of extent of any difference in the survival times. The variables such as age, turmour size and physiological variables are called *explanatory variables.*

In order to explore the relationship between the survival experience of patients and the explanatory variables, methods based on statistical modelling are applied. In survival analysis, the interest centers on risk or hazard of failure at any time *t.* For this reason, the modelling process focuses on hazard function.

There are two common broad regression models used to relate the predictors to the hazard function and these are: *Proportional hazards model/ Cox regression model* and *accelerated failure time model*.

The proportion hazard model is *a semi-parametric* model, because no assumptions are made regarding the nature of the baseline hazard function $\lambda_0(t)$ and also there is no assumption made regarding the distribution of the survival times, while the accelerated failure time model can be considered parametric. In the subsequent section, the Cox model is discussed.

2.4.1 Cox-regression/ Proportion hazard model

The model was proposed by Cox (1972) and it is also called proportion hazard model because it is based on an assumption that , for two groups of survival data, the hazard of failure for a subject in one group at time *t* is proportional to the hazard of failure for the similar individual in the other group at the same time *t*.

In its basic form, the hazard function for a subject with predictors

$X_i^T = \left(x_{i1}, x_{i2,} \ldots, x_{ip}\right)$ is

$\lambda_i(t, \boldsymbol{X}) = \lambda_0(t) \exp(\beta^T \boldsymbol{X_i})$,

where β^T is the vector of regression coefficients and $\lambda_0(t)$ is the baseline hazard function, that corresponds to hazard function of a subject with $\beta^T X_i = 0$. X_i is a vector of covariates for the *i*-th subject. Taking the log of the above function gives

$\log\left(\frac{\lambda_i(t,\boldsymbol{X})}{\lambda_0(t)}\right) = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} +, \ldots, \beta_p x_{ip}.$

The above equation means that log hazard ratio is equal to the linear component of the explanatory variables. It can also be shown that e^{β_i} is the hazard ratio.

2.4.2 Cumulative Incidence Curves (CIC)

This is an altenative to Kaplan-Meier in competing risks setting , and involves the use of marginal probabilities as introduced by (Kalbfleisch & Prentince, 1980). The CIC is derived from a cause-specific hazard function, and provides estimates of the "marginal probability" of an event in the presence of competing events, and does not require the assumption that competing risks are independent. In order to come up with CIC, first estimate the hazard at ordered t_j when the event of interest occurs $\hat{h}_k(t_j)$: *estimated proportion of subjects that fail from risk k.* In order to be able to fail at time t_j, a subject must have survived to the previous time t_{j-1}, thus the overall survival to time t_{j-1} is denoted $S(t_{j-1})$. Overall survival is considered than cause-specific survival $S_k(t)$ because the subject must have survived all other competing events. The estimated incidence of failing from event type *k* at time t_j is denoted by:

$$\hat{I}_k(t_j) = \hat{S}_k(t_{j-1})\hat{h}_k(t_j).$$

Thus cumulative incidence at time t_j is then the cumulative sum up to time t_j (*j=1 to j'=j*) of these incidence values over all event type *k* failure times.

or $I_k(t_j) = P(T \leq t, k) = \int_0^t h_k(u)\, S(u) du$ for continuous *t* and

$I_k(t_j) = \sum_{j=1}^{j'} \hat{S}_k(t_{j-1})\hat{h}_k(t_j)$ when *t* is discrete.

This is not a proper probability distribution since the cumulative incidence of failure from event k is below one. Having looked at methods for survival analysis, next are longitudinal data analysis methods.

2.5 Parameter estimation in Cox regression model

The relationship between the hazard function and the explanatory variables is best explored through the estimation of the regression coefficients $\beta_{i,s}$. One way of estimation is to assume the parametric form of the baseline hazard such as exponential and then estimate the coefficients by maximization of the corresponding log-likelihood function. Kalbleish & Pentice (2002), derived a likelihhod involving only $\beta \; and \; X$ based on marginal distribution of the ranks of the observed event times in the absence of censoring. To the contrary, Cox(1972) showed that parameter estimation can be done without specifying the baseline hazard and generalized in case of censoring.

Suppose that all event times are distinct, and let $t_1 < t_2 < t_3, \ldots, < t_k$ be the ordered event times. Let $R(t_i)$ be the risk set at time t_i, and there will be r_i individiuals in $R(t_i)$. The parameter β^T can be estimated using partial likelihood maximization which is the product over the set of observed event times of the conditional probability of seeing the observed events, given the set of individuals at risk at those times, and the partial likelihood is given as:

$$partial \; L(\beta) = \prod_{i=1}^{n} \left[\frac{\exp(\beta^T x_i)}{\sum_{k \varepsilon R_i} \exp(\beta^T x_k)} \right]^{\sigma_i}$$

where σ_i is the failure or censoring indicator with $\sigma = 1 \; fails, 0 \; otherwise$. Inference is done by treating the partial likelihood as if it fulfills all the properties of full likelihood. The log-partial likelihood is given as:

$$l(\beta) = \log \left[\prod_{i=1}^{n} \left[\frac{\exp(\beta^T x_i)}{\sum_{k \varepsilon R_i} \exp(\beta^T x_k)} \right]^{\sigma_i} \right]$$

$$= \sum_{i=1}^{k} \left[\exp(\beta^T x_i) - \log(\sum_{k \varepsilon R_i} \exp(\beta^T x_k) \right]$$

Using partial-likelihood score equations: $\frac{\partial l(\beta)}{\partial \beta} = 0$**,** the maximum partial likelihood estimators $\widehat{\boldsymbol{\beta}}$ are obtained**,** where in the process iterative optimization procedures such as the Newton-Raphson algorithm are applied. The estimated $\widehat{\boldsymbol{\beta}}$ can be used to estimate baseline hazard and cumulative hazard using Breslow's estimates,

$\hat{\lambda}_0(t) = \frac{1}{\sum_{k \varepsilon R_i} \exp(\hat{\beta}^T x_k)}$ and $\hat{\Lambda}_0(t) = \sum_{i:t_i} \frac{1}{\sum_{k \varepsilon R_i} \exp(\hat{\beta}^T x_k)}$.

After fitting a Cox model, it is important to conduct model diagnostics, to check whether the fitted model comforms the data or not. In this next section, model diagnostic procedures in survival data analysis are discussed.

2.6 Model checking in Cox regression model

Once the Cox regression model has been fitted to the observed data, it is very important to check the adequacy of the model. The use of diagnostic procedures is an essential part in model process. For example, the model must include an appropriate set of explanatory variables from those measured in the study, and may need to check that the correct functional form of these variables have been used. It may also be necessary to check the proportional hazards assumptions in the modelling process. In Cox regression model, the procedures for model checking are based on quantities called *residuals.* In survival data analysis, various residuals namely, Cox-Snell, Modified Cox-Snell, Martingale, Deviance, Schoenfeld and Score residuals. In this section, Cox-Snell and Schoenfeld residuals are discussed.

Cox-Snell Residuals

Introduced by Cox and Snell (1968). They are the mostly used residuals in survival data analysis. The Cox-Snell residual for the *i-th* subject at time t_i , *i=1, 2,. . . ,n* is given by

$$r_{C_i} = \exp(\hat{\beta} x_i)\hat{H}_0(t_i)$$

where $\hat{H}_0(t_i)$ is an estimated baseline cumulative hazard at time t_i. Note that the r_{C_i} is just equal to $\hat{H}_i(t_i) = -\log \hat{S}_i(t_i)$,where $\hat{H}_i(t_i)$ and $\hat{S}_i(t_i)$ are estimated cumulative hazard and survival functions for the *i-th individual* from the observed data. In order to check for model adequacy, one expects that if the fitted model is satisfactory, then a model based on estimate of survival function for the *i-th* individual at time t_i will be closer to the corresponding value of $S_i(t_i)$. Here $-\log \hat{S}_i(t_i)$ behaves as *n* observations from exponential distribution with unit mean.

Schoenfeld residuals

The disadvantage of the rest of the residuals is that they depend on estimated survival function and require an estimate of cumulative hazard function. These challenges are addressed by residuals proposed by Schoenfeld (1982). The important feature of this residual is that there is no single value of residual, for each individual, rather a set of values corresponding to each and every predictor included in the fitted model. The *i-th Schoenfeld residual for* $\boldsymbol{x_j}$, the *j-th* explanatory variable in the model is :

$$r_{P_{ji}} = \delta_i(x_{ji} - \hat{a}_{ji}) , \quad \hat{a}_{ji} = \frac{\sum_{l \varepsilon R(t_i)} x_{jl} \exp(\hat{\beta} x_l)}{\sum_{l \varepsilon R(t_i)} \exp(\hat{\beta} x_l)}$$

where x_{ji} is value for the *j-th* explanatory variable *j=1,2,3...*, *p* for the *i-th* individual in the study, $\delta_i = 0$ if censored and $\delta_i = 1$ if uncensored, $R(t_i)$ is the risk set at time t_i. Schoenfeld (1982) showed that the $r_{P_{ji}}$ are asymptotically uncorrrelated and have expected value of zero, thus a plot of $r_{P_{ji}}$ and X_j should be centered around zero.

2.7 The Extended time dependent Cox model

The previous Cox model in section 2.4.1, assumes that the hazards depend on covariates whose values do not change with time. For a given subject, some covariates may be changing with time. These covariates are called *time-dependent*. These can further be classified as *exogenous* or *endogenous* covariates. Exogenous variables have values that change because of causes not related to the subject of the study, 'external' characteristics that affect several individuals simultaneously. Presented in terms of probability, the future path of the exogenous covariate up to any time t > s is not affected by the occurrence of an event at time points, i.e.,

$\Pr(\gamma_i(t)|\gamma_i(s), T_i \geq s) = \Pr(\gamma_i(t)|\gamma_i(s),\ T_i = s)$

where $0 < s \leq t$ and $\gamma_i(t) = \{y_i(s), 0 < s \leq t\ \}$.

The standard examples include period of the year (winter or summer), and environmental factors such as temperature, humidity, and polution levels (Andrinopoulou, 2014). Other factors that can be predetermined from the beggining of study such as treatment dose, are also examples of exogeneous covariates.

The endogenous covariates are time-dependent measurements taken on the subjects under study, such as biomarkers and clinical parameters. An exogenous covariate is a predictable process, while the endogenous covariates are not. Another feature of endogenous covariates is that they are usually measured with error and their complete path up to any time is not fully observed and their complete history is not known (Andrinopoulou, 2014). For example, in malaria studies, blood biomarkers such as haemoglobin levels, number of parasites and others are endogenous covariates. In this

case, there is a need to postulate a model that relates the time varying covariates and the time to event of interest.

This model was proposed by Cox (1972). The extended Cox model is given as:

$\lambda_i(t, y_i(t), x_i) = \lambda_0(\mathrm{t})\exp[(\beta^T x_i) + \alpha y_i(t)]$,

where y_i is a vector of time dependent covariates and x_i denotes the baseline covariates. Interpretation of regression coefficient vector $\boldsymbol{\alpha}$ is that $\exp(\alpha)$ denotes relative increase in the risk for an event at time *t* that results from unit increase in $y_i(t)$. The interpretation of $\boldsymbol{\beta}$ is the same as in the previous Cox model.

Parameter estimation of $\boldsymbol{\alpha}$ is based on partial likelihood estimation function,

$$P\big(l(\beta,\alpha)\big) = \sum_{i=1}^{n} \int_0^{\infty} \{\exp[(\beta^T x_i) + \alpha y_i(t)]\} - log\left[\sum_j \exp\big[\big(\beta^T x_j\big) + \alpha y_j(t)\big]\right] dN_i(t)$$

where $N_i(t)$ is the counting process which counts the number of events for subject *i*, by time *t.*

The above Extended Cox model assumes that the time dependent covariates are a predictable process, measured without errors and their complete path can be specified, and assume that covariates change value at the follow-up visits and remain constant in the time interval in between these visits which is delusive with time dependent endogenous covariates such as biomarkers. Ignoring these features and fit extended Cox regression model will lead to biased estimates of the biomarkers, thus need for dedicated methods of survival analysis.

2.8 Parametric Models for Analysis of Survival data

The previous Cox model seen before, does not make any distribution assumption regarding the event times. In parametric model, the survival times are assumed to have a certain distribution, so the baseline hazard is also assumed to have some distribution. Models in which a specific probability distribution is assumed for the survival times are called *parametric models* (Collet, 2003). Parametric forms that can be assumed for the survival time include, the *exponential, Weibull, Exreme value, log-normal* and *log-logistic* distributions (Tableman & Kim, 2004). However, the mostly used parametric models are Weibull distribution introduced by Weibull (1951) and exponential distribution. This section only presents the Weibull distribution.

Weibull distribution.

Suppose that survival time T follows a Weibull distribution, then

$f(t) = \lambda\gamma t^{\gamma-1}\exp(-\lambda t^{\gamma})$ for $0 \leq t < \infty$ as the Weibull probability density function.

The corresponding hazard function is given as: $h(t) = \lambda\gamma t^{\gamma-1}$

so that the survival function is:

$S(t) = \exp(-\int_0^t \lambda\gamma\mu^{\gamma-1}\, d\mu) = \exp(-\lambda t^{\gamma})$,

where λ is a scale parameter and γ as a shape parameter. This distribution, has the $E(T) = \lambda^{-1}\Gamma(1+\frac{1}{\gamma})$. With the change in shape parameter, the following is the behaviour of $h(t)$, the hazard function.

For $\gamma < 1$, $h(t)$ is a monotone decreasing function, $\gamma \geq 1$, $h(t)$ is an increasing function. In order to relate the hazard function at any time with respect to covariates, then the hazard model at any time *t*, with respect to the explanatory variables will be:

$h_i(t) = \exp(\boldsymbol{\beta x_i})\, h_0(t)$, where baseline hazard is given as: $h_0(t) = \lambda\gamma t^{\gamma-1}$. The corresponding survival function, based on the above hazard is :

$S_i(t) = \exp(-\exp(\boldsymbol{\beta x_i})\, \lambda\gamma t^{\gamma})$.

Fitting these models require parameters to be estimated. The following section, discusses the parameter estimation in the above hazard model.

2.8.1 Parameter Estimation in Weibull distribution

In order to assess the plausibility of the distribution for survival times , one compares the survival function with that of a chosen model. One way to do this, is by transforming the survival function to produce a plot that should give a straight line if the assumed model is appropriate. However, to achieve this, parameters in the model need to be estimated using, the methods of maximum likelihood.

Suppose there are *r* failures among the n individuals with *n-r* censored times. For Weibull distribution function, the likelihood function is given as:

$L(\gamma, \lambda) = \prod_{i=1}^{n} \big(h_i(t)\big)^{\delta_i} S_i(t)$,

where δ_i is zero if censored and unity otherwise and $h_i(t)$ and $S_i(t)$ are hazard and survival functions respectively, as defined in previous subsection. The corresponding log likelihood function after substituting the hazard and survival functions accordingly is:

$\log\big(L(\gamma, \lambda, \boldsymbol{x})\big) = \sum_{i=1}^{n}[\delta_i\{\boldsymbol{\beta' x_i} + \log(\lambda t) + \gamma log t_i\} - \lambda\exp(\boldsymbol{\beta' x_i})t^{\gamma}]$.

Differentiating this function and equating to zero gives the maximum likelihood estimates of $\boldsymbol{\beta'}, \gamma \; and \; \lambda$.

In the next section, methods for analysis of survival data, but in competing risks settings are discussed.

2.9 Competing risks Survival Data Analysis

In medical studies, though researcher may be interested in one particular outcome of interest, there are situations where there are several reasons why an event can occur, and this is known as “competing risks”. Competing risks are said to be present when a patient is at risk of more than one mutually exclusive events, such as death from different causes, and the occurence of one of these will prevent any other event from ever happening (Gichangi & Vach, 2005). Estimation of the marginal or net survival function of the time to an event, in a competing risks framework is a common problem encountered in medical applications of survival analysis (Zheng & Klein, 1995). Often it is impossible to measure the time to occurence of an event of interest due to the occurrence of some other event, a competing risk, at some time before the event of interest. This competing event can be the withdrawal of the subject from the study, death, or failure from some cause other than the one of interest (Williamson *et al.*,2007).

In clinical trials of new treatments, patients may withdraw because they are doing poorly with new treatment. For example, in epilepsy studies, patients diagnosed with epilepsy are given antiepileptic drug. In such studies, if the treatment groups have no seisure and there are few side effects, it is considered successful. Treatment that causes unacceptable side effects are changed to an altenative, whilst a treatment that fails to control seizures will either be changed to an altenative, or a second drug will be added (Williamson *et al.*, 2007).

In such studies, of interest is retention time, defined as the time from randomisation to the withdrawal of the randomised Anti-epileptic drug or addition, a primary outcome recommended by the International League Against Epilepsy (Sander *et al.*, 2007).

The possible competing risks include withdrawal due to adverse effects and inadequate seizure control. Some authors have examined separately the time to withdrawal due to side effects (Lhatoo *et al*., 2000) and have censored patients whose allocated treatment is changed due to inadequate seizure control, which may give misleading results as analyses assume that the competing risks of withdrawal are independent (Kalbfleisch & Prentince, 1980). Ignoring this aspect of an outcome by analysing events overall can result in misleading conclusions (Kalbfleisch & Prentince, 1980). If a patient experiences a competing event, standard survival analysis methods treat that patient as censored for the outcome of interest (Scherzer, 2017). Censoring due to loss to follow-up may be negatively or positively correlated with the event time (Hevia, 2014). This is very important, since Kaplan-Meier curves overestimate the incidence of the outcome over time and use of Cox models inflates the relative differences between groups, resulting in biased hazard ratios (Scherzer, 2017).

Models for Competing risks

A notion of competing risks setting can be graphically presented with an event-free state and various end points corresponding to distinct failure types.

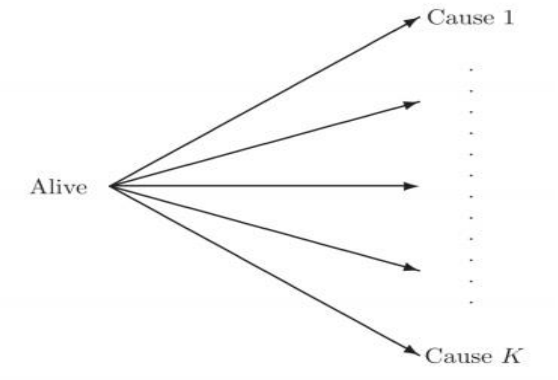

Figure 2: *Multiple events scenario with k distinct events.*

One approach to handle such situations is to consider those that experience the other event types as censored, and then estimate the corresponding probabilities of failure an approach called " *naive Kaplan- Meier"*. The problem with this approach is that, it violates the independence assumption of censoring process. In case where the competing risks survival times distributions were independent of the survival times for the event of interest, one would expect the hazard of event of interest at any time point to be the same for the subjects that are still under follow-up. Howbeit, the naive K-M will overestimate the probability of the failure and underestimate the survival probabilities. The other approach is the use of cause-specific models.

2.9.1 Cause-specific hazard function

Let $C_i(T_i, D_i)$ denote the competing risks survival data on subject i, where T_i is the failure or censoring time , and D_i takes the values $\{0, 1, 2, \dots, g\}$, with $D_i = 0$ indicating a censored event and $D_i = k$ showing that subject i fails from the k^{th} type of failure, where $k = 1, \dots, g$. Throughout, the censoring mechanism is assumed to be independent of the survival time. The cause –specific hazard model is defined as:

$$\lambda(t)_k = \lim_{h \to 0} \frac{P(t \leq T_i \leq t+h, D_i | T_i > t)}{h} = h_{0k}(t) exp\left(\boldsymbol{\gamma_k}^T W_i^T\right),\ t > 0$$

Where $\lambda(t)_k$ is the instantaneous rate for failure of type k, W^T is a vector of covariates that are associated with the hazard function, and $\boldsymbol{\gamma_k}^T$, a corresponding vector of regression coefficients, h_{0k} is a completely unspecified baseline hazard function for risk k. In its basic form, it is assumed that the hazard ratio $\left(\frac{\lambda(t)_k}{\lambda_{0k}(t)}\right)$ depends only on covariates, whose value is fixed during follow-up (baseline covariates).

The exp($\boldsymbol{\gamma}$) is called the cause-specific hazard ratio for the k event, and it represents the relative risk of failing from that event when the correspondent variable increases one unit in its value. In competing risks settings, the other approach is to use the cumulative incidence curves for the event of interest and the competing events.

2.10 Longitudinal Data Analysis

Longitudinal studies are defined as studies in which the outcome variable is repeatedly measured; i.e. the outcome variable is measured in the same individual on several different occasions (Twisk, 2003) . Longitudinal data are frequently encountered in health studies related to humans, animals or laboratory samples. Longitudinal data can be obtained retrospectively or prospectively, but much of longitudinal data is collected in prospective studies. In sociology and economics, longitudinal studies are refered to as panel studies. For example, in Family Medicine, studies on cardiovascular complications among patients with diabetes mellitus, longitudinal outcomes such as glucose levels, blood lipids levels, systollic and diastollic pressure are measured repeatedly over time (Weel, 2005). In cancer studies, the longitudinal data such as circulating tumor cells, immune response to a vaccine, a genetic biomarker, or a health outcome are recorded (Ibrahim *et al.,*2010). Another example is in AIDS studies where repeated measurements of CD4+ and viral load are recorded during disease progression. In clinical trial, comparing different drug regimen therapy for schizophrenia, the Positive and Negative Symptom Rating Scale (PANSS) a measure of psychiatric disorder are repeatedly recorded (Diggle *et al.,* 2004). In reproductive epidemiology, studies on progestrone collect the urinary metabolite progesterone over the course of the women's menstrual cycles (days) (Wu & Zhang, 2006).

Longitudinal data is got via longitudinal research design at a number of seperate occassions in time called "*phases*" or "*waves of the study*". In a randomized clinical trial, investigators often collect prospective longitudinal data on one or more endpoints in response to a particular intervention relative to a control condition. The focus may be either to determine if there is a significant difference between control and treated individuals at the end of the study, often termed an "endpoint" analysis, or to examine differential rates of change over the course of the study in treated and control conditions.

In the case in which subjects are initially randomized to the control and treatment conditions, differences in either the final response or the rate of response over time (e.g., differential linear trends over time) are taken as evidence that the treatment produces an effect on the outcome measure of interest above and beyond chance expectations based on responses in the control condition (Hedeker & Gibbons, 2006). Measuring subjects repeatedly through the duration of the study, one expects positive correlation, which means that standard statistical tools (like the t-test and simple regression) that assume independent observations, are not appropriate for this kind of data analysis (Hevia, 2014). Since longitudinal data are prospective studies, missing data issues are inevitable. In an attempt to treat the longitudinal data, with more realistic assumptions and missing processes, a variety of more rigorous statistical methods have been developed.

The mostly used models in longitudinal data analysis include ANOVA models, MANOVA models, covariance pattern models, mixed-effects regression models (Laird & Ware, 1982), and generalised estimating equations (GEE) (Zeger *et al*., 1988).

Other models for analysis of missing processes such as selection models, pattern mixture models, and shared parameter models for discrete time are existent and applied in longitudinal data analysis (Molenberghs & Kenward, 2007). In the following subsections, the focus is on mixed-effects model, covariance pattern models and GEE models.

2.10.1 Mixed-effects Regression Model (MRM)

A basic characteristic of mixed effects models is the inclusion of random-subjects effects in the model in order to account for the influence of subjects on their repeated measurements. Random effects describe each person's trend across time, and explain correlational structure of the longitudinal data. This intuitive idea about the MRM is depicted by Figure 3.

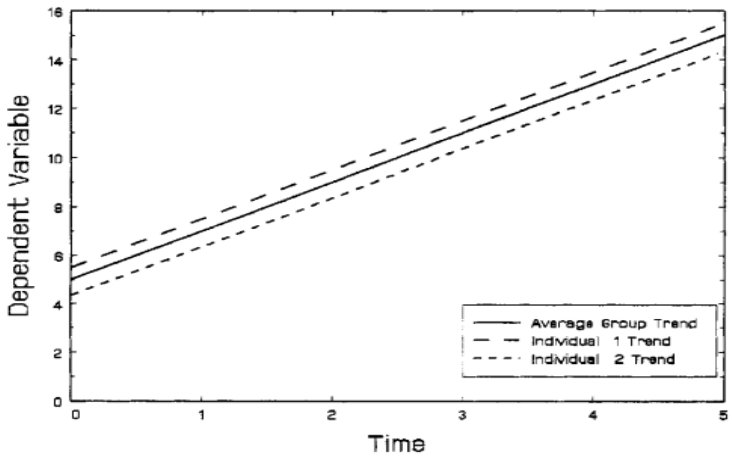

Figure 3: *Random intercept Mixed-effects Model*

Figure 3 depicts the population average trend represented by the solid line, and two individual trends which are subject-specific mean profiles over time, one above and other below the population average trend line.

The basic model is given as follows:

$$y_i(t_{ij}) = \beta_0 + \beta_1 t_{ij} + \varepsilon_{ij}.$$

In this model, both *y* and *t* are allowed to change by individual and occassions, and $\varepsilon_{ij} \sim N(0, \delta^2)$ and independently distributed.

The independence assumption of random errors makes the above model unrealistic for longitudinal data, since *y* are observed repeatedly from same individual and it is more reasonable to assume that errors within an individual are correlated to some degree. The model also posits that the change across time is the same for all indviduals since $\beta_0 \; and \; \beta_1$,(initial level and linear change across time) do not vary by individual. As such it more important to add individual-specific effects that will account for data dependency and describe differential time trends for different individuals.

An extension to the above model is given as:

$y_i(t_{ij}) = \beta_0 + \beta_1 t_{ij} + v_{0i} + \varepsilon_{ij}$,

where v_{0i} describes the influence of individual *i* on their repeated observations, v_{0i} may deviate from zero, since subjects may have negative or positive influence on their longitudinal data. Goldsten(1995) suggests that the model should be presented in a hierarchical or multilevel form to best reflect how the model characterises an individual's influence on their observations. Thus

Level 1 : $y_i(t_{ij}) = b_{0i} + b_{1i} t_{ij} + \varepsilon_{ij}$, as within subject model

Level 2: $b_{0i} = \beta_0 + v_{0i}$, $b_{1i} = \beta_1$ between–subjects model /slope as outcome model (Burstein, 1980).

Individuals in a sample are representative of the general population, then v_{0i} are considered random effects, with $v_{0i} \sim N(0, \delta_v^2)$, and $\varepsilon_{ij} \sim N(0, \delta^2)$, which are conditionally distributed (conditional on individual-specific effects, v_{0i}) and the above model is a random effects model. The generalization of this model, allowing additional

predictors and regression coefficients to vary randomly, is known as the *linear mixed-effects model,* and the standard model for a continuous longitudinal outcome is given as:

$$y_i(t_{ij}) = m_i(t_{ij}) + \varepsilon_{ij} \quad (1)$$

where $m_i(t_{ij})$ is a linear predictor term (mean response):

$$m_i(t_{ij}) = X_i(t_{ij})^T B^{(1)} + Z_i(t_{ij})^T b_i \quad (2)$$

and $\varepsilon_{ij} \sim N(0, \boldsymbol{\delta^2 I_{n_i}})$ are measurement errors , which are assumed to be independently and identically distributed. Let $X_i(t_{ij})$ be a design vector of covariates for subject i associated with fixed effects $B^{(1)}$. Also $Z_i(t_{ij})$ denotes row vector of the design matrix associated with a latent random variable (vector) that can be interpreted as subject-specific random effects b_i.

The standard model assumes that the random effects are distributed as multivariate normal with mean 0 and variance-covariance matrix D, i.e $b_i \sim N(0, D)$ and the model also posits that b_i is independent of ε_{ij} . $y_i(t_{ij})$ is the longitudinal outcome measured at time *t* for subject *i*, where *i* = 1, 2, . . . , n_i, and $\sum n_i$ is the total number of longitudinal observations . The use of n_i, suggests the fact that due to different event times , some patients may miss one or more visits. It is further assumed that the missing values in the longitudinal measurents caused by reasons other than occurence of events are missing at random. This model accounts for correlation for measurements within a subject and handles unequally spaced visit times.

The interpretation of the fixed effects $\boldsymbol{B^{(1)}}$ is the same as in a simple linear regression model: assuming *p* covariates in the design matrix X, the coefficient B_j , for *j=1, 2,...,p*

denotes the change in the average $y_i(t_{ij})$ when the corresponding covariate X_j is increased by one unit, while all other predictors are held constant. In the same way, b_i show how a subset of the regression parameters for the *i-th* subject deviates from those in the population.

With mixed-effects models one is able to include subjects with incomplete data across time, hence increasing the statistical power, and avoid bias presented by complete-case analysis where complete cases may not be representative of the whole population. The other advantage of mixed effects model is that MRM estimates changes for each subjects, and mean response changes in the population. It is important to estimate the parameters in the model, and the next section discusses the methods applied in parameter estimation.

2.11 Parameter Estimation in Mixed-effects regression model

In this section are the methods used to estimate the fixed effects and random effects in the analysis of longitudinal data.

- **Fixed effects estimation**

Parameter estimation can be done using generalised least-squares estimation for $\boldsymbol{B}$ which is the same as maximum likelihood estimation of fixed effects. This can be done, using marginal models,

$y_i(t_{ij}) = X_i B + \varepsilon_{ij}^*$, with $\varepsilon_{ij}^* = Z_i b_i + \varepsilon_{ij}$, with correlated errors , and variance-covariance matrix,

$cov(\varepsilon_{ij}^*) = Vi = ZiDZ_i^T + \delta^2 I_{n_i}$, with I_{n_i} identity matrix. Assuming that V_i is known, minimizing the function,

$\cup = (y - XB)'V^{-1}(y - XB)$, one obtains the generalized least square estimators for $\boldsymbol{B}$, given as:

$$\hat{B} = \left(\sum_{i=1}^{n} {X_i}^T V_i^{-1} X_i\right)^{-1} \sum_{i=1}^{n} {X_i}^T V_i^{-1} y_i$$

This method is based on Liang and Zeger (1986).

- **Random effects estimation**

There are various methods to prediction of random effects, one of which is using Henderson's mixed model equation, that considers joint distribution of *y and b* and the log-likelihhood function of the linear model (Henderson *et al.*, 2000). The Henderson's equation yields the best linear unbiased estimators for both the random effects and the fixed effects. The equation is given as follows:

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + D^{-1} \end{bmatrix} \begin{bmatrix} B \\ b \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix},$$

and this has the solutions

$\hat{B} = (X'V^{-1}X)^{-1}X'V^{-1}y$ and

$\hat{b} = DZ'V^{-1}(y - XB)$, where $R = \delta^2 I_{nxn_i}$.

V is estimated using maximum likelihood estimation or restricted maximum estimation. However, MLE is biased for small samples and RMLE gives better estimates with small sample. To obtained a closed form of the estimates, numerical optimization methods such as Expectatiom–Maximization (EM) methods are used.

2.12 Covariance Pattern Models (CPMs)

These models were introduced by Jennrich and Schluchter(1986), and assumes that timing of the measurements are fixed (subjects intended to be measured at the same finite number of occassions).

They are an extension to MANOVA models, however they allow incomplete data across fixed number of time points and allow a variety of possible variance-covariance structures. The variance-covariance matrix of the repeated measurements are assumed to be of particular form, and not resulting from inclusion of random-subject effects. They are considered as an extension to multiple linear regression with flexibility in the structure of variance-covariance matrix. The model is given as below:

$\boldsymbol{y_i = X_i\vartheta + \varepsilon_i}$, $i=1, 2,..., N$, $j=1 ,2, \ldots, n_i$ observations for individual *i*. $\boldsymbol{y_i}$ is the n_i x 1 dependent variable vector for *i-th* individual, $\boldsymbol{X_i}$ is the n_i x p predictor vector for an *i-th* individual. $\boldsymbol{\vartheta}$ is the $p\times1$ vector of fixed regression parameters and $\boldsymbol{\varepsilon_i}$ is the n_i x 1 error vector. It is further assumed that $\boldsymbol{\varepsilon_i \sim N(0, \sum_i)}$ and $\boldsymbol{y_i \sim N(X_i\vartheta, \sum_i)}$. Note that each $\sum_i$ is a submatrix of *n xn* matrix $\sum$ where *n* is the total number of fixed points.

2.12.1 Variance-Covariance Structures

There are several (error) variance-covariance structures and these include compound symmetry, first-order autoregressive, Toeplitz, unstructured form and random effects, and exponential structures. In this section, the compound symmetry and the Unstructured forms are described.

Compound Symmetry

This specifies equal variances and equal covariances. In a matrix form, it can be written as

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 + \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \dots & \sigma_1^2 \\ \sigma_1^2 & \sigma^2 + \sigma_1^2 & \sigma_1^2 & \dots & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma^2 + \sigma_1^2 & \dots & \sigma_1^2 \\ \cdot & \cdot & \cdot & \dots & \cdot \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \dots & \sigma^2 + \sigma_1^2 \end{bmatrix}$$

With variance of the response variable $\sigma^2 + \sigma_1^2$ and covariance for any pairwise association of the response variable is σ_1^2. The number of parameters in this structure is 2, namely σ^2 and σ_1^2.

Unstructured form

The other variance-covariance structures assume that variance is constant across time (Compound symmetry), and that the lagged correlations are either all the same, decrease exponentially or equal within lag. For example the Toeplitz structure, is reasonable when the time intervals for the measurements are the same or nearly the same. In reality, the above assumptions may not hold, hence resort to unstructured form. This allows that the parameters be different and has a symmetric matrix with *n(n+1)/2* parameters. In a matrix form,this can be written as :

$$\boldsymbol{\Sigma} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} & \dots & \theta_{1n} \\ \theta_{21} & \theta_{22} & \theta_{23} & \dots & \theta_{2n} \\ \theta_{31} & \theta_{32} & \theta_{33} & \dots & \theta_{3n} \\ \cdot & \cdot & \cdot & \dots & \cdot \\ \theta_{n1} & \theta_{n2} & \theta_{n3} & \dots & \theta_{nn} \end{bmatrix}$$

Where $\theta_{j'j} = \theta_{jj'}$ due to the symmetry.

2.12.2 Model Selection for the Variance-Covariance structures

To determine which covariance-variance structure that best fits the data, (Jennrish & Schluchter, 1986) use likelihood ratio test to compare the various structures to unstructured form (saturated model). The test has $\boldsymbol{n(n+1)/2 - q^*}$ parameters where $\boldsymbol{q^*}$ is the number of parameters in the reduced model. This model selection requires that the covariates be the same for both the saturated and reduced models. Both maximum likelihood and restricted maximum likelihood methods can be used in model estimation and likelihood calculations.

2.13 Generalized Estimating Equation Models (GEE)

Longitudinal data are correlated. To account for this correlation, GEE models were developed by (Zeger *et al.,* 1988) as an extension to generalized linear models. GEE models are also called *marginal models* indicating that the model for the mean response depends on covariates of interest not any random effects (Fitzmaurice *et al.,* 2004). The model assumes MCAR missing processs for the data, fixed number of time points, and only the $n \times n$ correlation matrix is considered in GEE models. The model makes specifications on marginal distribution and likelihood of y_{ij} for varying y which is different in CPMs where the joint distribution and likelihood function of $\boldsymbol{y_i}$ are specified. GEE models have a wide range of applications to different types of outcome variables, including count, categorical and continuous data. More literature on GEE models include, Davis(1993), Zeger et al (1988), Diggle *et al* (2002) and Hardin and Hilbe (2003). Here, GEE models based on Liang and Zeger (1986) are described. In order to understand the GEE model, it is important to consider the generalized mixed effects models, since they are considered a root for GEE models, in case of correlated longitudinal measurements.

Generalized Linear Models

GEE basically has the focus on regression of parameters $\boldsymbol{\beta}$ not the variance-covariance structures. Since GEE models are an extension to GLMs for the case of corrrelated data, to motivate the understanding of GEE models one reviews the GLMs. GLMs are family of models that are used to fit fixed effects regression models to normal and non-normal data (Nelder & Wedderburn, 1972).

In these models the response variable is considered to belong to a class of distributions called exponential family with common GLMs such as linear regression for normally distributed response variable, logistic regression for binary response, and Poisson regression for count dependent variable. The GLMs have three specifications, namely linear predictor, link function and variance of the response variable $\boldsymbol{y_i}$ conditional on $\boldsymbol{x_i}$. The linear predictor $\boldsymbol{\varphi_i = x_i'\beta}$ with the possible link functions *g(.)* that convert the expected value μ_i of $\boldsymbol{y_i}$ to the linear predictor and the variance of $\boldsymbol{y_i}$ as is seen below:

$g(\mu_i) = \mu_i$ an identity link function in ordinary multiple regression and $var(\boldsymbol{y_i}) = \emptyset v(\mu_i)$ where $v(\mu_i)$ is a known variance function and $\emptyset$ is a scale parameter that may be known or estimated. In ordinary logistic regression, $\emptyset$ is set to 1, and $v(\mu_i)$ is the error-variance. The link functions and variances of $\boldsymbol{y_i}$ can be specified for both logistic and Poisson regressions.

To obtain estimates for $\boldsymbol{\beta}$, the estimation equation

$$U(\boldsymbol{\beta}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \left(\frac{\boldsymbol{\delta \mu_i}}{\boldsymbol{\delta \beta}}\right)' (v(y_i))^{\boldsymbol{-1}}(\boldsymbol{y_i} - \mu_i) = \mathbf{0}$$

which depends on mean and variance of *y* and not the distribution of *y* is used. Thus solutions of this equation provided estimates called "quasi-likelihood estimates". GLMs are fixed effects regression models and assume that the observations are independent, hence not suitable for analysis of longitudinal data which is highly correlated. However, they are extended to account for correlation inherent of longitudinal data, thus GEE models.

- **GEE Models**

Since these models assume fixed time points, they only need to assume distribution of $\boldsymbol{y_{ij}}$ at time *j*. There is no need to assume joint distribution of $\boldsymbol{y_i}$, but only the marginal distributions of the independent variable at the time points. GEE focus on regression of *y* on X. Since GEE models are an extension to GLMs, they also make the following specifications:

Linear predictor: $\boldsymbol{\varphi_{ij} = x'_{ij}\beta}$, with $\boldsymbol{x'_{ij}}$ as covariate vector for subject *i* at time *j*.

Link function: $g(\mu_{ij}) = \boldsymbol{\varphi_{ij}}$

Varaince of y: $v(y_{ij}) = \emptyset v(\mu_{ij})$, with $v(\mu_{ij})$ as known variance function and $\emptyset$ a scale parameter that can be estimated.

Working Correlation Structure

This is an additional component that misses in GLMs. The working correlation structure of the repeated measures is **R** of size *n x n* since subjects are measured at fixed time points: This specification, now accounts for correlation inherent of longitudinal data. The subject *i* correlation matrix is $\boldsymbol{R_i}$ of size n_i *x* n_i if $n_i < n$, since subjects need not to measured at each time point. The **R** thus $\boldsymbol{R_i}$, is assumed to be dependent on a vector of

correlation parameters denoted $\boldsymbol{\alpha}$ which present the average dependence among the repeated observations across subjects.

There are various working correlation structures, namely, *independence, exchangeable, AR(1), m-dependent and unspecified.* The next section discusses the independence, exchangeable and AR(1) structures.

Independence Structure

This structure assumes that $\boldsymbol{R_i(\alpha) = I}$ an nxn idenity matrix, an equaivalence to an assumption that data is not correlated. Howbeit, the assumption seems illogical with longitudinal data where correlation can not be ruled out. It is proposed that this structure leads to loss of efficiency with binary outcomes, but has an advantage to models that include time varying covariates.

Exchangable Structure

It assumes that all the correlations in **R** are the same. It specifies that $\boldsymbol{R_i(\alpha) = \rho,}$ an equivalence to compound symmetry in CPMs.

AR(1) structure

In this structure, the within-subject correlation over time is an exponential function of the lag. It is denoted as $\boldsymbol{R_i(\alpha) = \rho^{|j-j'|}.}$ With the order of the time lag, and dependence on one term , the correlations tend to decline.

Now parameters estimation in generalised estimation equations is presented.

2.14 Parameter Estimation in GEE models

In order to estimate parameters in GEEs, let $\boldsymbol{B_i}$ be an $n \times n$ diagonal matrix with $V(\mu_{ij})$ as the jth diagonal element. Also let $\boldsymbol{R_i(\alpha)}$ $\boldsymbol{n \times n}$, be the the working correlation matrix for subject *i*. The associated working variance-covariance matrix for y_i is defined as, proposed by Liang and Zeger (1986).

$$\boldsymbol{V(\alpha) = \emptyset B_i^{1/2} R_i(\alpha) B_i^{1/2}}$$

To find the estimates for $\boldsymbol{\beta}$, one uses solutions to the equation,

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \mathbf{Z}_i' \, [\boldsymbol{V(\hat{\alpha})}]^{-1} (\boldsymbol{y_i - \mu_i}) = \mathbf{0}$$

where $\boldsymbol{\hat{\alpha}}$ is a consistent estimator of α and $\boldsymbol{Z_i = \frac{\partial u_i}{\partial \beta}}$. The above formula is just an extension to estimation equation in GLM but now for correlated longitudinal data. The equation depends only on mean of $\boldsymbol{y_i}$ and its variance, thus the associated solutions are called *quasi-likelihood estimates.*

Having reviwed the basics for survival data analysis and longitudinal methods, it is important to extend to joint modelling process.

2.15 Joint Modelling for Longitudinal and Survival data

Here, the interest is on building a model that takes into account all three aspects namely, longitudinal data, survival data and competing risks. In the first place an approach that jointly takes into account the survival and longitudinal data without the competing risks data is presented. In the course, reasons for adopting joint models than the seperate methods of longitudinal and survival data are also discussed. Later, an extension will be made, to introduce joint models in the competing risks settings.

2.15.1 Basic Nature of Joint Models

The extended Cox regression model may take account of longitudinal data and survival data, but with limitations as previously observed. The longitudinal markers and the time-to-event outcomes in survival data, may be interrelated, suggesting the dependence. Under extended Cox regression model, only the exogenous covariates are taken care of, but not the endogenous. Also under time-dependent Cox model it is postulated that the value of the longitudinal outcome remains constant between the observed times. Assumption which are not valid for time endogenous covariates. In order to illustrate the whole idea behind joint modelling, an example is given by (Rizopoulos, 2012), in Figure 4.

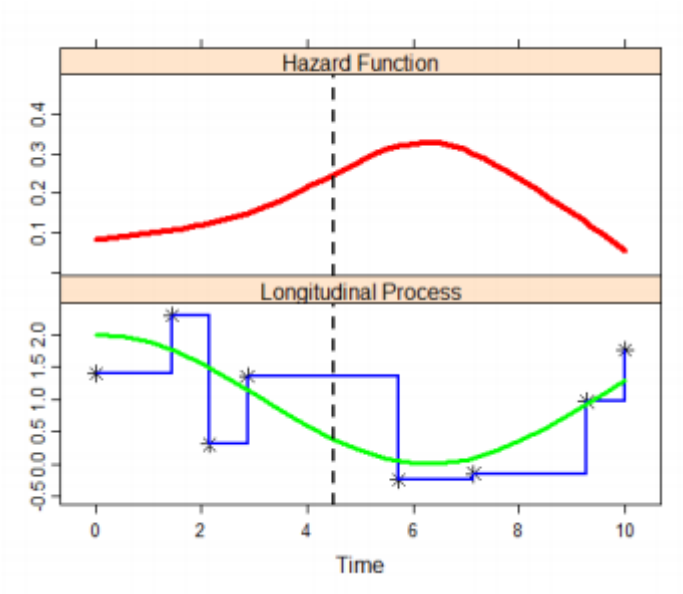

Figure 4: *An intuitive idea of joint models.*

In Figure 4, the bold line in the upper part of the figure shows the instantaneous rate of failure (hazard). The lower part shows a faint line that represents the longitudinal process. The starred line in longitudinal process represents the extended-Cox approximation of the longitudinal trajectory. From Figure 4, one can tell how the hazard function is associated with the longitudinal process, suggesting the dependence

between the longitudinal process and survival process in time. The blue line indicates an assumption under time-dependent Cox model which posits that the value of time dependent covariates remain constant between the visits, yet the line is staggered in the figure.

Modelling the longitudinal outcome with Cox models seperately, will lead to an introduction of errors in the estimation of the longitudinal process if the two processes are associated . This is where, the joint modelling of the longitudinal and survival data comes in for it takes into account both the exogenous and endogenous time-dependent covariates.

However, in the absence of correlation between longitudinal and survival outcomes, each outcome can be analysed separately (Marchenko, 2016) using the separate methods. Joint modelling of longitudinal and time-to-event data is an area of increasing research, which allows the simultaneous modelling of a longitudinal (repeatedly measured over time) outcome such as weekly biomarker measurements, and a time-to-event (survival) outcome such as time to death (Sudell *et al.*, 2016). The model is a combination of longitudinal and survival submodels that are linked using an association structure that quantifies the relationship between the outcomes of interest.

There are various forms to joint modelling approach of longitudinal and survival data. The basic joint model which consists of one longitudinal and one survival outcome was introduced by Self and Pawitan (1992), further work was done by DeGruttola and Tu (1994), Tsiatis et al. (1995), Faucett and Thomas (1996) and Wulfsohn and Tsiatis (1997). Also , Taylor et al. (2005), Garre et al. (2008), Yu et al. (2008), Proust-Lima

and Taylor (2009) and Rizopoulos (2011) considered the joint modelling framework to derive individualized predictions for a longitudinal and a survival outcome that are updated at each new visit. However, in this project the focus is on joint models by Rizopoulus (2012) whose focus is on individual's survival. In the process, submodel specifics to joint modelling, likelihood functions, parameter estimation, and underlining statistical inference are introduced. In order to be in line with project objectives, an extension will be made to competing risks settings. Before introducing the submodels, here are reasons for joint modelling.

2.16 Why Joint Modelling of Longitudinal and Survival Data?

Longitudinal measures are commonly incomplete or may be prone to measurement errors. As the longitudinal covariates are measured with errors, there is a requirement for more complex analysis than one that treats covariates as fixed markers in survival models. The inclusion of raw longitudinal measurements in the survival analysis leads to bias (Prentice, 1982). Studies on joint models have been proposed to solve difficulties in Cox proportional hazard model with time-dependent covariates, which are possibly missing at some event times or subject to substantial measurement error (Brown & Ibrahim, 2003). In other cases, undertaking a joint model that evaluates both longitudinal and survival data simulteneously, reduces biases and improve precision over simpler approaches (Henderson *et al.*,2000). For example, if a particular drug reduces the hazard of a particular disease by 30%, then a joint model may lead to an estimated hazard ratio of 0.75, whereas a conventional model (eg, a Cox model) that does not incorporate the longitudinal data into the analysis may yield a hazard ratio of 0.80.

In this case, one says that the estimate based on the joint model is less biased than the Cox model estimate because 0.75 is closer to the true hazard ratio of 0.70 (Ibrahim, Chu, & Chen, 2010). Thus joint models provide efficient estimates of the treatment effects on both markers and on time to event of interest. When the longitudinal marker is correlated with a survival outcome, joint modeling framework shows superiority over modeling the two processes separately (Liu, 2016).

Often, longitudinal and survival data are collected together, hence it may be important to investigate the relationship between the serial biomarker and event of interest. A paper by Andrinopoulou (2014) suggests that joint models are an appropriate statistical tool for assessing the progression of serial biomarkers accounting for patients drop-out due to reasons associated with the study. Next are models involved in joint modelling of longitudinal and survival data.

2.17 Submodels in Joint Modelling

As highlighted in section 2.1, joint model is a combination of longitudinal and survival submodels that are linked using an association structure that quantifies the relationship between the outcomes of interest. The model in this project is based on proposition by Rizopolous (2012). Let T_i^* be the true event time for the *i-th* subject , and let T_i be the observed event time, where $T_i = \min(T_i^*, C_i)$, where C_i is the censoring time. Let $\delta_i = I(T_i^* \leq C_i)$, i.e. δ_i is unity for the true event. Let also assume that $m_i(t)$ is true or unobserved value of the longitudinal marker at time t. Then $y_i(t)$ is the observed value of the time dependent covariate at time t , and $y_{ij}(t) = \{y_i(t_{ij}), j = 1, 2, \ldots, n_i\}$. The aim is to associate the true unobserved longitudinal outcome $m_i(t)$ with the hazard of an event.

Survival Submodel

As stated under extended time dependent Cox regression model, the standard relative risk model is defined as :

$$h_i\big(t\big|\boldsymbol{\mathcal{M}}_i(t)\big) = h_0(t)\exp\big(\gamma^T w_i + \boldsymbol{\alpha}(m_i(t)\big). \qquad (1^*)$$

Where $\boldsymbol{\mathcal{M}}_i(t) = \{m_i(s),\ 0 < s \leq t\}$ is longitudinal history of the unobserved, $h_0(t)$ is the baseline hazard , and w_i is a vector of baseline covariates. $\boldsymbol{\alpha}$ quantifies the strength of the association between the marker and the risk of an event / the effect of underlying longitudinal outcome to the risk for an event. Model regression coefficients are interpreted as seen in previous sections. *exp(*γ_j*)* denotes the hazards ratio from one unit increase in *j-th* covariate, *exp(*$\boldsymbol{\alpha}$*)* denotes relative increase in the risk of an event at time *t* resulting from one unit increase in $m_i(t)$.

In survival analysis, it is important to consider all history for the covariate, and not just a value $m_i(t)$. The survival function depends on full history of the marker. The survival function is:

$$S_i\big(t\big|\boldsymbol{\mathcal{M}}_i(t)\big) = \exp\{-\int_0^t h_0(t)\exp\big(\gamma^T w_i + \boldsymbol{\alpha}(m_i(t)\big)\ ds\}.$$

In literature , Li *et al*. (2008), suggests that the baseline hazard is unspecified, however, Hsieh *et al*.(2006) suggested that to avoid misspecification of the underlying parametric distribution of the survival times which in turn leads to under-estimation of standard errors of parameter estimates in the joint model settings, it is imperative to have it specified, using parametric forms as seen in section **2.8**. However, it is advisable to use a more flexible form of model for the baseline hazard function. Here, are two possible forms for baseline hazard function by Rizopoulos (2012).

- ***Regression spline model***

In this model, the log baseline hazard function is given as:

$$\log h_0(t) = k_0 + \sum_{d}^{m} k_d B_d\,(t, q)$$

where $k^T = (k_0, k_1, \ldots, k_m)$ are the spline coefficients, q denotes the degree of the B-splines basis function B(.) proposed by de Boor (1978) and the $m = \ddot{m} + q - 1$ where $\ddot{m}$ is the number of interior knots.

- ***Piece-wise Constant model***

In this model, the baseline function, takes the form,

$h_0(t) = \sum_{q=1}^{Q} \mathcal{E}_q I\big(V_{q-1} < t < V_q\big)$,

where $0 = v_0 < v_1 < v_2 < \ldots < v_Q$, a partition of time scale with v_Q the largest than the largest observed time, and $\mathcal{E}_q$ the value of the hazard in the interval $(v_{q-1}, v_q]$.

In both models , as the number of knots increases, the specification of the hazard becomes more flexible. In both models, it is important to avoid overfitting and keep balance between bias and variance. However, there is no ideal strategy to achieve this, still Harrel (2001) gives a standard rule of thumb based on keeping the total number of parameters between 1/10 and 1/20 of the number of events in the sample.

- ***ch-Laplace***

Under this method of specifying the baseline hazard, fully exponential Laplace approximation is used for integration over the random effects. The method is suitable where the subject specific longitudinal profiles are nonlinear and are modelled using higher dimensional random effects structures.

Longitudinal Submodel

The previous relative risk submodel, uses the unobserved longitudinal value $m_i(t)$. In order to determine the effect of the longitudinal outcome on the risk of an event, it is important to have $m_i(t)$ estimated and that the complete true longitudinal history $\boldsymbol{\mathcal{M}}_i(t)$ is reconstructed. Our focus is on continuous longitudinal markers, and the mixed-effect regression model is given as:

$$y_i(t) = m_i(t) + \varepsilon_i(t)$$

$$= \beta X_i^T(t) + Z_i^T(t)b_i + \varepsilon_i(t), \text{ where } m_i(t) = \beta X_i^T(t) + Z_i^T(t)b_i \qquad (2^*)$$

where X_i^T is the design vector of fixed coefficients β , and Z_i^T is a design vetor for random effects b_i. This vector b_i is a latent random variable that can be interpreted as subject specific effects of $Z_i^T(t)$, and $\varepsilon_i(t)$ are random errors that are assumed to be independent and normally distributed with mean zero and variance $\delta^2 I_{n_i}$, i.e $\varepsilon_i(t) \sim N(0, \delta^2 I_{n_i})$, for all $t \geq 0$.

It is also assumed that $\varepsilon_i(t)$ and b_i are independent. This vector b_i, follows the same distribution and assumptions as seen in section 3.1, $b_i \sim N(0, D)$ and accounts for association between the longitudinal and the event processes, and the correlation between the repeated measurements in the longitudinal outcome.

The joint model allows to settle that the longitudinal markers are a function of true unbserved longitudinal value $m_i(t)$ and some error, an attribute absent in time-dependent Cox model. In general, the joint model is given as a piece-wise function of equations (1^*) and (2^*) as shown below:

$$\begin{cases} y_i(t) = m_i(t) + \varepsilon_i(t) = X_i(t)^T\beta + Z_i(t)^T b_i + \varepsilon_i(t) \; longitudinal \;\; submodel \\ h_i\left(t \middle| \boldsymbol{\mathcal{M}}_i(t)\right) = h_0(t)\exp\left(\gamma^T w_i + \boldsymbol{\alpha}(m_i(t)\right) \quad survival \; submodel \end{cases}$$

2.18 Parameter Estimation in the Joint Model

In separate methods of analysis for both longitudinal and survival data, maximum likelihood estimation procedures are used to estimate model parameters. In joint modelling settings, the same approach of maximum likelihood estimation is applied by Rizopoulos (2012) to approaximate model parameters. Prior to estimation, the likelihood formulation in the joint modelling process, and estimation process are discussed.

2.18.1 Likelihood Function in the Joint Model

As previously specified in section 2.17 under the longitudinal submodel of the joint model, b_i account for association between the failure process and the repeated longitudinal observations for the outcome variable. The vector of random effects also account for correlation between the repeated longitudinal measurements.

In the joint likelihood formulation of longitudinal and survival data (T_i, δ_i, y_i), the survival and longitudinal processes are assumed to be conditionally independent given the vector of random effects b_i. The model also posits that the repeated measurements of the longitudinal outcome are independent of each other. Under above assumptions, thus

$p(T_i, \delta_i, y_i \ |b_i, \theta_i) = p((T_i, \delta_i \ |b_i, \theta_i)p(y_i \ |b_i, \theta_i)$ and

$$p(\ y_i \ |b_i, \theta_i) = \prod_j p(\ y_{ij} \ |b_i)$$

where $\boldsymbol{\theta} = (\theta_t^T, \theta_y^T, \theta_b^T)^T$ denotes the parameter vector for the event time outcome, the longitudinal outcomes and the random effects variance-covariance matrix respectively.

Thus, the joint likelihood contribution for the i^{th} subject as proposed by Tsiatis and Davidian (2004) is :

$$p(T_i, \delta_i, y_i; \theta_i) = \int p(T_i, \delta_i, y_i, b_i\,; \theta_i) db_i$$

$$p(T_i, \delta_i, y_i; \theta_i) = \int p(y_i|b_i) p(T_i, \delta_i|b_i; \theta_t, \beta)\; p(b_i)\; db_i$$

$$= \int \{h(T_i|\mathcal{M}_i(T_i);\; \theta)^{\delta_i} S(T_i|\mathcal{M}_i(T_i);\; \theta)\} \left[\prod_j p(\; y_i(t_{ij})\;\; |b_i;\; \theta_y)\right] p(b_i)\; db_i$$

where b_i explains the interdependencies, and $p(.)$ is the probability density function and $S_i(t|b_i) = \exp(-\int_0^t h_0(s) \exp\big(\gamma^T w_i + \boldsymbol{\alpha}(m_i(s)\big)\, ds$ is the survival function which depends on the whole longitudinal history. Also the product

$$\prod_j p(\; y_i(t_{ij})\;\; |b_i;\; \theta_y)\, p(b_i)$$

$$= (2\pi\delta^2)^{\frac{-n_i}{2}} \exp\{-\frac{\|\; y_i X_i\, \beta - Z_i b_i\;\|^2}{2\delta^2}\}$$

$$\times (2\pi)^{-\frac{q_b}{2}} \det(D)^{-\frac{1}{2}} \exp(-\frac{{b_i}^T D^{-1} b_i}{2})$$

where q_b denotes the dimensionality of the random effects vector, and $\|.\|$ denotes Euclidean vector norm.

The joint log-likelihood function with respect to θ is given as:

$$l(\theta)$$

$$= \sum_{i=1}^{n} log \int \{h(T_i|\mathcal{M}_i(T_i);\; \theta)^{\delta_i} S(T_i|\mathcal{M}_i(T_i);\; \theta)\} \left[\prod_j p(\; y_i(t_{ij})\;\; |b_i;\; \theta_y)\right] p(b_i)\; db_i$$

The maximization of this log likelihhod function, demands the optimization algorithms and numerical integration techniques to be applied. The existence of integrals in the

random effects, and the survival function that formulate the joint likelihood , result in no closed form solution, since it may be of high dimension hence approximated numerically. Standard numerical integration techniques such as Monte Carlo, Gausian quadrature, Laplace approximation and Expectation-Maximization techniques are applied. The latter receives more preference in literature than the others. As decribed by Wulfsohn and Tsiatis (1997), E-M algorithm intuitively involves treating the random effects as missing data, where in Expectation step, the missing data are filled, and the log-likelihood function of the observed data is replaced with a surrogate function, and maximization step where the surrogate function is maximized. Recently, Rizopoulos et al (2009) introduced a hybrid algorithm for maximazation of log likelihood which start with EM and continue with quasi-Newton (direct maximization).

2.18.2 Random Effects Estimation

In joint modelling, it is also important to consider estimation of subject-specific effects on their outcomes. The random effects b_i are estimated using Bayes Theory. As introduced by Rizopoulos (2012), assuming $p(b_i, \theta)$ as the posterior distribution, and $p(T_i, \delta_i | \, b_i; \, \theta) p\big(\, y_i\big(t_{ij}\big) \, \big| b_i; \, \theta \big)$ as conditional likelihood part, the condition posterior distribution of b_i is :

$$p(b_i | T_i, \delta_i, y_i \, ; \theta) = \frac{p(T_i, \delta_i | b_i; \, \theta) p\big(\, y_i\big(t_{ij}\big) \, \big| b_i; \, \theta \big) p(b_i; \theta)}{p(T_i, \delta_i, y_i, \theta)}$$

$$\propto \boldsymbol{p}((T_i, \delta_i | b_i; \, \theta) p\big(\, y_i\big(t_{ij}\big) \, \big| b_i; \, \theta \big) p(b_i; \, \theta)$$

This has no closed form solution, and numerical methods are applied to approximate the random effects. Standard summary measures for the posterior distribution are given as:

$$mean: \overline{b_i} = \int b_i \, p(b_i|T_i, \delta_i, y_i \, ; \theta) db_i$$

$$mode: \widehat{b_i} = argmax_b(log \;\; p(b_i|T_i, \delta_i, y_i \, ; \theta)$$

The impressive part about this distribution is that as the number of repeated longitudinal measurements increases , the distribution converges to normal distribution.

2.19 Competing Risks Joint Models

As seen in section 2.9, there are situtions when apart from an event of interest, there may be competing risks. In this section, an extension is made to standard joint model for longitudinal and survival data, to accomodate the competing risks settings. Though, joint modelling apparoach has been an increasing area of research, much of the research on joint modelling of longitudinal and survival data have been focused on data with a single event time and a single mode of failure, combined with an assumption of independent censoring of event times (Tsiatis & Davidan, 2004). However, in some situations interest lies with more than one possible cause of event or where the censoring is informative. Some considerable works on joint modelling of longitudinal and survival data in competing risks settings has been done by Gueorguieva et al.,(2012) ; Andrinopoulou et al., (2014) and Proust-Lima et al., (2016), Hevia, (2014) and others. In all work by above authors, cause- specific hazard regression and mixed effect models are used, with an extension to the relative risk model for basic joint model to account for competing risks.

The cause-specific model, postulates relative risks models for each of the competing event type. The idea behind these models is to couple a cause- specific hazard model for the continuous time-to-event process with a mixed-effects regression model for the longitudinal outcome.

Assuming one has K different event types, let T_{i1}^{*} , $T_{i2}^{*}, \ldots, T_{iK}^{*}$ be the true failure times for K event types. Let T_i be the observed failure time such that $T_i = \min(T_{i1}^{*}, \ldots, T_{iK}^{*}, C_i)$, where C_i is the censoring time. Let D_i takes the values $\{0, 1, 2, \ldots, g\}$, with $\delta_i = 0$ indicating a censored event and $\delta_i = k$ showing that subject i fails from the $k - th$ type of failure, where $k = 1, \ldots, g$. The relative risk model for competing risks is now the cause-specific hazard model given as:

$$h_{ik}\left(t\middle|\boldsymbol{\mathcal{M}}_i(t)\right) = h_{0k}(t)\exp\left({\gamma_k}^T w_i + \boldsymbol{\alpha_k}(m_i(t)\right)$$

where w_i vector of baseline covariates, $m_i(t)$ true value of the longitudinal marker with $\boldsymbol{\mathcal{M}}_i(t) = \{m_i(t), 0 \leq s < t\}$. The corresponding mixed effects model is given by:

$$y_i(t) = m_i(t) + \varepsilon_i(t)$$

$$= \beta X_i^T(t) + Z_i^T(t)b_i + \varepsilon_i(t), \quad b_i \sim N(0, D) \text{ and } \varepsilon_i(t) \sim N(0, \delta^2 I_{n_i})$$

Parameter estimation is just the same as in the basic joint model, with some changes to the likelihood formulation. For competing risks and longitudinal outcomes joint model, the likelihood is given as:

$$p(T_i, \delta_i | b_i; \theta_t, \beta)$$

$$= \prod_{k=1}^{K}\left[h_{0k}(T_i)\exp\left({\gamma_k}^T w_i + \boldsymbol{\alpha_k}(m_i(t)\right)\right]^{I(\delta_i = k)}$$

$$\times \exp\left(-\sum_{k=1}^{K}\int_0^{T_i} h_{0k}(s)\exp\left({\gamma_k}^T w_i + \boldsymbol{\alpha_k}(m_i(s)\right) ds\right)$$

In the likelihood function, the baseline hazard is estimated using regression spline method, as in section 4.3.

2.19.1 Assessing Model Assumptions in competing risks joint model

In mixed-effects and cause-specific hazard models, there are methods that are used to test model assumptions . In joint modelling alone, Rizopoulos (2012), propose the use of multiple imputation residuals with the fixed visit times to validate the model assumptions. However, when the longitudinal data, survival data, and competing risks components are amalgamated, the assessment of model assumptions becomes complicated.

CHAPTER 3

METHODOLOGY

In the previous Chapter, methods for analysis of longitudinal and survival data that are collected in biomedical studies are discussed. Next, is to apply the methods for joint modelling of survival data with competing risks and repeated measures of longitudinal data to real data. This section presents an overview of the data that are used for analysis of this project. It also explains the statistical package that is used for the analysis of the data.

3.1 Methods

The data used for this project is part of the primary dataset that was collected from a randomized controlled trial that aimed at evaluating strategies to delay the emergency of resistance to anti-malarial drugs in children by Malawi Liverpool Wellcome Trust and College of Medicine in 2001 to 2003 (Bell *et al.,* 2008). For the primary study, written informed consent was required from the parent of each child recruited and the study was explained in parent's preferred language . The study protocol was approved by ethics committees of the College of Medicine, University of Malawi, and Liverpool School of Tropical Medicine.

The study primarily targeted children aged 12 to 60 months of age, weight $\geq 6\ kg$, no feature of severe malaria (event of interest) on enrollment, hemoglobin $\geq 5.0 g/dl$ was measured using hamocue. The children were randomized to four treatment armsnamely,

Sulfadoxine-Pyrimethamine(SP),Chloroquine(CQ) + Sulfadoxine-Pyrimethamine(SP), Amodiaquine(AQ) + Sulfadoxine-Pyrimethamine(SP) and Artesunate (ART) + Sulfadoxine-Pyrimethamine (SP) and followed up for a period of six weeks.

The children were recruited at, and followed up from, Chileka Health Centre which is about 19 kilometers from Queen Elizabeth Central Hospital (QECH), and serves all the immediate needs of the local population, referring major problems to QECH.

All children were also required to provide venous and capillary blood samples on assessment days for parasite microscopy. From the blood samples taken from the children, biomarkers such as hemoglobin, white cell count, red blood cell count, platelets , creatinine and Bilirubin were examined in full blood count. Children were considered pure, with P. *falciparum* parasitaemia parasite density between 2000 to 200 000 parasites per μl. Children were removed from the primary study after enrolment if their full blood count showed severe anaemia, that is hemoglobin level less than 5 g/dl. On the other hand, during follow-up withdrawal was based on adverse reactions to the randomised drug, protocol violation and consent withdrawal.

The study was a double blinded trial as all members of study team and patients were uninformed of study treatments allocation. The patients were assessed on days 0, 7, 14, 28 and 42 and any other day if unwell. In order to obtain this data, one of the supervisors who was part of the study team had rights to share the data for academic purposes and other use.

3.2 Outcomes of Interest

For a period of two years, the study recruited 500 children in all treatment arms. However in this project, data that was collected in 2001 were used, which consist of 101 children. The study managed to collect baseline characteristics of children such as age in months, sex, and body weight in kilograms(kg). During follow up, the repeated longitudinal measurements for hemoglobin and parasite counts were recorded. This study uses longitudinal outcomes of hemoglobin level, collected on days, 0 ,7 , 14 , 28, and 42 and parasites counts recorded on same visit times.

Time to severe malaria/ treatment failure was an outcome of interest, in the presence of a competing risk of withdrawal from the study . In the study, withdrawal was due to adverse reaction to the drugs used, protocol violation and consent withdrawal as a competing risk for severe malaria. All the children that were withdrawn on above grounds, were considered withdrawn on reasons associated with the study in this project. Status indicator for subjects that dropped out of the study from other unknown reasons, loss to follow up (censored), severe malaria, and withdrawal was defined.

3.3 Statistical Analysis

The data for this project were analyzed using a statistical package R version 3.5.1. Firstly, descriptive statistics for baseline covariates and other biomarkers were presented. Then seperately, longitudinal mixed-effects models for longitudinal outcomes, hemoglobin and parasite counts were applied to this data set, leaving aside the survival models.

The next models are competing risks models where two different failures of severe malaria and withdrawal were taken into account, by including the baseline covariates. Finally, in order to meet the objectives of the study, joint models were applied to competing risks data. Since there were two longitudinal outcomes, they could not be included in the joint model at the same time. As a result, models were fitted for each of the repeated longitudinal outcomes.

3.3.1 Mixed-effects Model for Real Data

In this case, two separate mixed effects models of hemoglobin and parasites counts with fixed and random effects were considered. Let $y_{h,i}$ and $y_{P,i}$ denote the hemoglobin level and parasite counts for the i-th individual, $i = 1, \dots, n$. Then

$$y_{h,i} = \beta_{h,0} + \beta_{h,1}\, sex_i + \beta_{h,2}\, age_i + \beta_{h,3}\, weight_i + \beta_{h,4}\, time_i + \beta_{h,5}\, treat_i + b_{h,0i} + e_{h,i} \quad (3)$$

$$y_{P,i} = \beta_{P,0} + \beta_{P,1}\, sex_i + \beta_{P,2}\, age_i + \beta_{P,3}\, weight_i + \beta_{P,4}\, time_i + \beta_{p,5}\, treat_i + b_{P,0i} + e_{P,i}, \quad (4)$$

where sex, age, weight, time, and treatment are baseline covariates, $\beta_{h,0}$, $\beta_{h,1}$, $\beta_{h,2}$, $\beta_{h,3}$, $\beta_{h,4}$, $\beta_{h,5}$ are coefficients for baseline covariates respectively, and $b_{h,0i}$ and $e_{h,i}$ are random effect and errror terms.

The models above are for longitudinal biomarker hemoglobin and parasite counts respectively. The model for hemoglobin assume that sex, age, and weight, are fixed effects, with random slope of subjects (patients). The model assumes that there may be different longitudinal profiles from subject to subject. It also assumes that the random effects and the random errors come from normal distribution, just as discussed

previously. The model for estimated parasites counts includes the random-slope of subjects, with the same underlying assumptions as the hemoglobin model.

3.3.2 Competing Risks Survival Model for Real Data

The competing risks survival models for the severe malaria and withdrawal with baseline covariates, age, sex, treatment, weight, baseline hemoglobin level, and parasite counts are the extended time-dependent cause-specific Cox models with no interaction and given as:

$$h_{i\,sm}(t) = h_{0sm}(t)\exp\{\gamma_{1sm,}treat_i + \gamma_{2sm}age_i + \gamma_{3sm}sex_i + \gamma_{4sm}weight_i + \gamma_{5sm}hb0 + \gamma_{6sm}parasite0\}$$

In this model, since severe malaria is an event of interest, withdrawal is treated as censored in addition to usual censored observations, resulting from lost to follow-up.
In the similar manner, the extended time-dependent cause-specific model for withdrawal, treating severe malaria and lost to follow-up as censored is given as:

$h_{i\,wd,k}(t) = h_{0wd}(t)\exp\{\gamma_{1wd}treat_i + \gamma_{2wd}age_i + \gamma_{3wd}sex_i + \gamma_{4wd}weight_i + \gamma_{5wd}hb0 + \gamma_{6wd}parasite0\}$.

3.3.3 Joint Models for Real Data with Competing Risks and Longitudinal Markers

In order to evaluate the relationship between the longitudinal marker and severe malaria in the presence of competing risk of withdrawal, two separate joint models were analyzed, each including a different longitudinal outcome. The longitudinal markers of hemoglobin and parasite counts separated by the event type of severe malaria or withdrawal were considered.

This approach is recommended when focus is on survival outcome and allows the evaluation of the impact of serial longitudinal markers. As detailed in section chapter 2, section 2.7 and subsection 2.10.1, the true submodels in joint modeling approach are as follows:

- The longitudinal mixed effect submodel for hemoglobin.

$$y_{h,i} = \beta_{h,0} + \beta_{h,1}\, sex_i + \beta_{h,2}\, age_i + \beta_{h,3}\, weight_i + \beta_{h,4}\, time_i + \beta_{h,5}\, treat_i + b_{h,0i} + e_{h,i}$$

- Survival submodels for hemoglobin

$$\begin{cases} h_{h,i\,sm}(t) = h_{h,0sm}(t)\exp[\gamma_{h,1sm} sex_i + \gamma_{h,2sm} age_i + \gamma_{h,3sm} weight_i + \beta_{h,sm4}\, treat_i + \partial_{h,sm} m_{h,i}(t)] \\ h_{h,i\,wd}(t) = h_{h,0wd}(t)\exp[\gamma_{h,1wd} sex_i + \gamma_{h,2wd} age_i + \gamma_{h,3wd} weigh_i + \beta_{h,wd4}\, treat_i + \partial_{h,wd} m_{h,i}(t)] \end{cases}$$

Here $m_{H,i}(t)$ is the true value of the longitudinal marker hemoglobin. Similary, the longitudinal submodel for parasites is given as follows:

$$y_{P,i} = \beta_{P,0} + \beta_{P,1}\, sex_i + \beta_{P,2}\, age_i + \beta_{P,3}\, weight_i + \beta_{P,4}\, time_i + \beta_{p,5}\, treat_i + b_{P,0i} + e_{P,i}$$

And the corresponding survival submodel are:

$$\begin{cases} h_{p,i\,sm}(t) = h_{p,0sm}(t)\exp[\gamma_{p,1sm} sex_i + \gamma_{p,2sm} age_i + \gamma_{p,3sm} weight_i + \beta_{p,sm4}\, treat_i + \partial_{p,sm} m_{p,i}(t)] \\ h_{p,i\,wd}(t) = h_{p,0wd}(t)\exp[\gamma_{p,1wd} sex_i + \gamma_{p,2wd} age_i + \gamma_{p,3wd} weigh_i + \beta_{p,wd4}\, treat_i + \partial_{p,wd} m_{p,i}(t)] \end{cases}$$

where $m_{p,i}(t)$ is the true value of the longitudinal marker parasite counts.

CHAPTER 4

RESULTS

This chapter presents the analysis results obtained in R version 3.5.1, after analyzing the malaria data. Firstly, the chapter presents the basic descriptive statistics and then the separate models for longitudinal and survival malaria data. In the last part of this chapter, the joint models applied to real malaria data in the presence of competing risk of withdrawal are presented followed by model comparision for the separate and competing risks joint models.

4.1 Basic Descriptive Analysis

This study used data for 101 children. In this study, male children (57.4%) and female children (42.6%) data were analyzed. The children were on average aged 2.22 years (std=1.12) . The average body weight of the chidren was 11.03 kilograms and the median follow-up time of 28 days. The different outcomes that were observed are severe malaria (5.0%) and withdrawal (10.9%). In the competing risks analysis, severe malaria was an event of interest and withdrawal as a competing risk. The remaining children (84.2%) were censored in the analysis. The clinical biomarkers considered in the analysis were parasites and hemoglobin level. The analysis used measurements obtained on visit days 0, 7, 14, 28 and 42, for parasites and hemoglobin longitudinal measurements. The results, report an average of six parasite count and an average of 9.38 g/dl hemoglobin level for the baseline data on day 0.

In Figure 5 are boxplot graphics for the two longitudinal biomarkers, parasites and hemoglobin levels seperated by the event types, and barplots for treatment and sex against the event type.

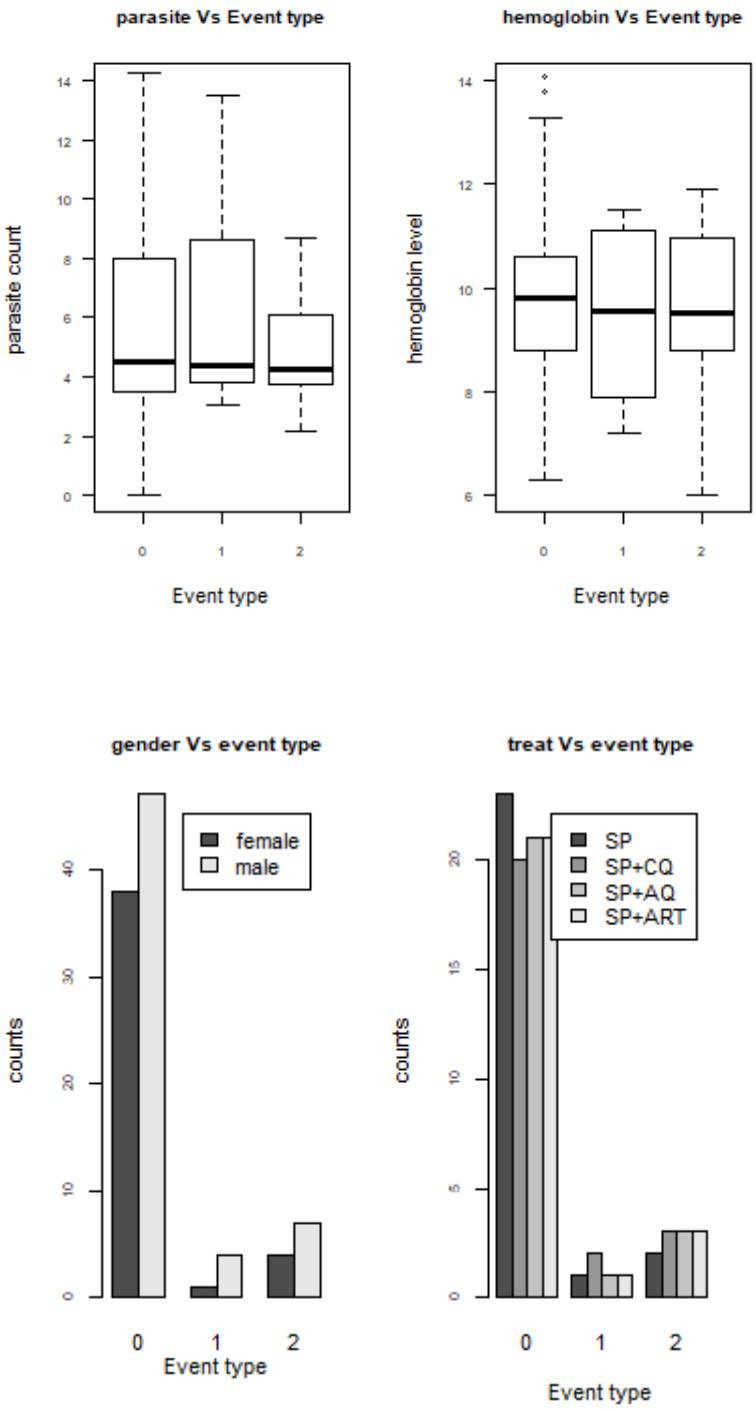

Figure 5: *Baseline explanatory variables classified by event type a patient experienced: 0 for censored event, 1 for severe malaria and 2 for withdrawal.*

For children that failed (experienced severe malaria), there were more males than females and the same was the case for children who experienced withdrawal in the course of the study. As can be observed in Figure 5, the median hemoglobin level for children that experienced severe malaria is slightly the same as the medians for hemoglobin levels of children that were censored and withdrawn from the study. It was also observed that the medians number of parasites were the same in all event types.

On randomized treatment and severe malaria experinece, children that were randomized to treatment arm "Chloroquine and Sulfadoxine-pyrimethamine (SP+ AQ)" had slightly more cases of severe malaria than those in the other treatment arms that experienced severe malaria. The numbers of patients that were withdrawn seemed to be slightly the same in all treatment arms except in the Sulfadoxine-pyrimethamine group.

4.2 Linear Mixed-Effects Models

As highlighted previously, this section presents the linear mixed-effects model to describe the evolution in time of longitudinal biomarkers, hemoglobin level and parasites.

In Figure 6, subject-specific evolutions in time of the longitudinal biomarkers hemoglobin level and parasites measurements are presented. From Figure 6, it was observed that subjects showed similar variability in their longitudinal profiles for hemoglobin and parasite in all treatment groups.

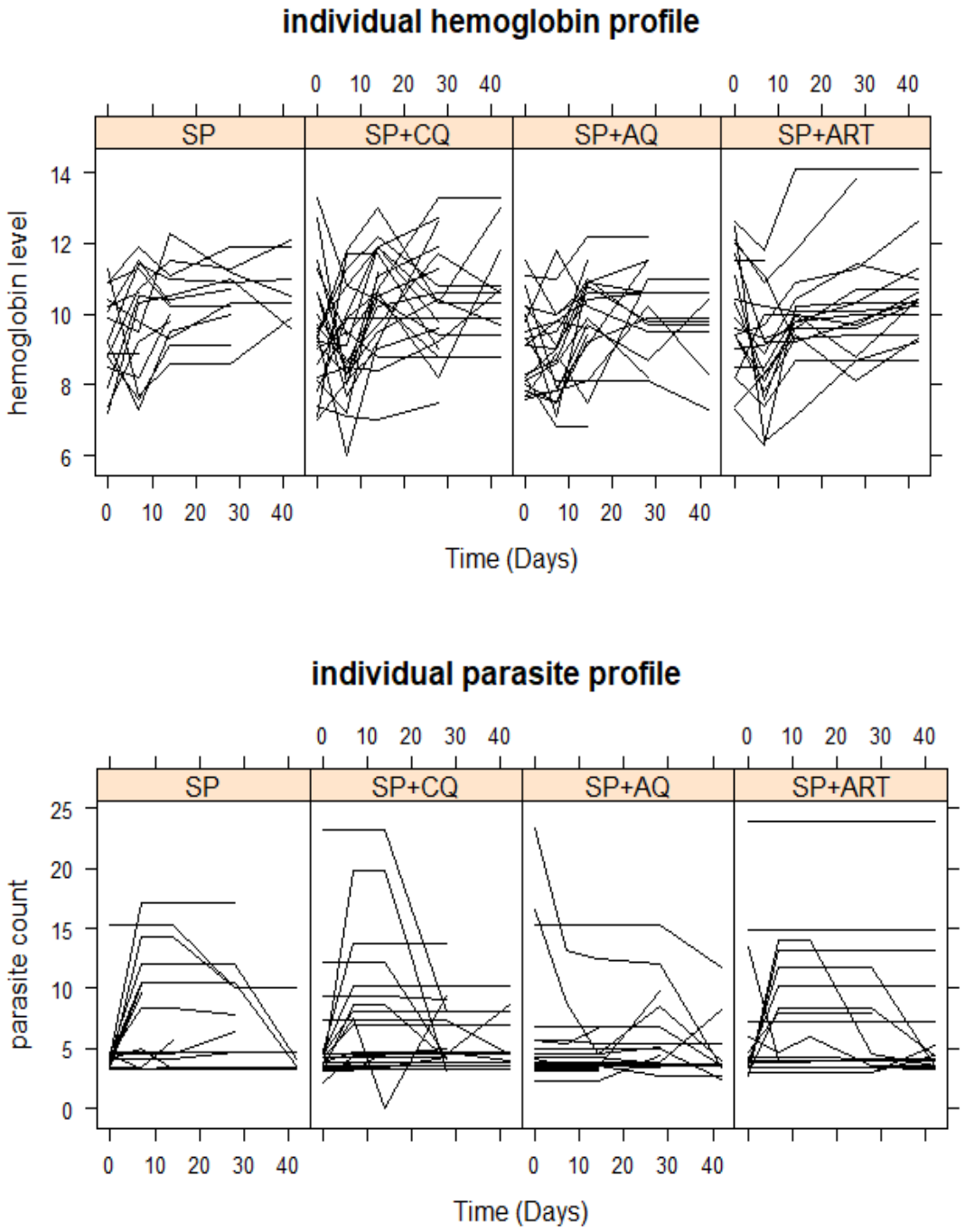

Figure 6: *The individual hemoglobin and parasites profiles over time in days seperated by treatment that a patient received.*

Next the mixed effects regression models for the malaria data are presented as stated in section 3.3.1.

Table 2 summarizes the fitted mixed-effects regression models for hemoglobin and parasites counts.

From the table, the results suggest that time was statistically significant predictor of the longitudinal scores of hemoglobin levels, with hemoglobin level increasing by 0.03 g/dl units for any passing day (s.e = 0.004). The intercept was also statistically significant, implying that the value for hemoglobin level is 8.38 g/dl, when all parameters are zero (s.e=0.65).

For the parasite count scores, the results show that the intercept was statistically significant in predicting the parasite counts, with parasite counts of six parasites when all other parameters are zero. The parasite counts for male patients increase by two parasites than the parasite counts for female counterparts, however this result is not statistically significant. The other variables, time , age and weight and treatment were not statistically significant in the prediction of parasite counts.

Table 2: Fitted values for the linear mixed-effects models for the longitudinal variables hemoglobin level, and parasite counts with standard deviations (sde), and the p-values

	Slope	SE	p-value
Hemoglobin			
$\beta_{H,0}$ (Intercept)	8.38	0.65	< 0.0001
$\beta_{H,1}$ (sexmale)	-0.05	0.22	0.835
$\beta_{H,2}$ (age)	0.12	0.14	0.376
$\beta_{H,3}$ (weight)	0.06	0.07	0.409
$\beta_{H,4}$ (time)	0.03	0.004	< 0.0001
$\beta_{H,5}$ (SP+CQ)	0.19	0.31	0.549
$\beta_{H,5}$ (SP+AQ)	-0.13	0.32	0.695
$\beta_{H,5}$ (SP+ART)	0.20	0.32	0.528
Random effects			
$b_{H,0i}$ (Intercept)		0.89	
Residual		1.10	
Loglik:	-623.9	**BIC:** 1306.7	
Parasites counts			
$\beta_{p,0}$ (Intercept)	5.61	2.30	0.016
$\beta_{p,1}$ (sexmale)	1.55	0.79	0.054
$\beta_{p,2}$ (age)	-0.46	0.49	0.345
$\beta_{p,3}$ (weight)	0.08	0.24	0.744
$\beta_{p,4}$ (time)	-0.01	0.01	0.493
$\beta_{H,5}$ (SP+CQ)	0.30	1.10	0.786
$\beta_{H,5}$ (SP+AQ)	-0.59	1.14	0.608
$\beta_{H,5}$ (SP+ART)	0.65	1.13	0.568
Random effects			
$b_{p,0i}$ (Intercept)		3.52	
Residual		2.76	
Loglik:	-988.2	**BIC**: 2035.4	

4.3. Survival Competing Risks Models

In competing risks settings, the cumulative incidence curves for the two competing events , severe malaria and withdrawal, accounting for failure times and the cause of the failure were estimated. Graphically, the cumulative incidence functions provide an insight of how the events of severe malaria and withdrawal evolve over time.

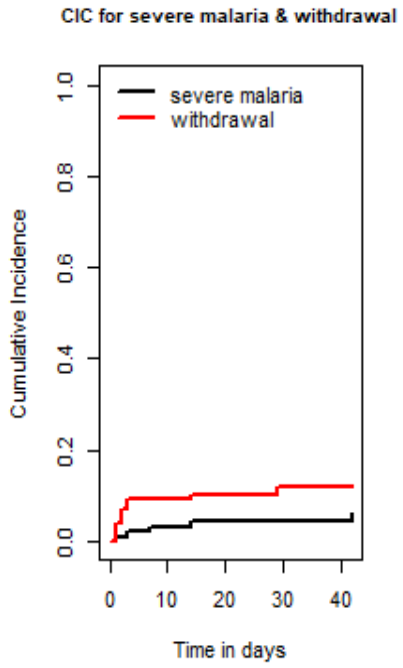

Figure 7: *Cumulative incidence curves for the two competing events, severe malaria and withdrawal.*

The cumulative incidence rates are higher for an event of withdrawal (study related conditions) than cumulative incidence rate for severe malaria, with more withdrawal events between day 0 and day 10.

In Table 3**,** parameter estimates and standard errors after fitting the cause-specific hazard regression models for the two events, severe malaria and withdrawal are presented.

Table 3: Fitted values for the competing risk models.

Parameter	Log(RR)	SE	p-value
CQ + SP	0.25	0.74	0.737
CQ + SP : CR	1.6e-15	1.05	1.000
AQ + SP	0.060	0.80	0.940
AQ + SP: CR	1.0e-15	1.13	1.000
ART+ SP	-0.03	0.81	0.969
ART+ SP:CR	2.0e-15	1.15	1.000
Age	-0.03	0.34	0.369
Age: CR	8.1e-16	0.49	1.000
Weight	0.212	0.16	0.189
Weight: CR	-5.5e-16	0.23	1.000
Sexmale	0.49	0.56	0.377
Sexmale: CR	-9.6e-16	0.79	1.000
Hb0	0.09	0.17	0.596
Hb0 : CR	-5.5e-16	0.24	1.000
Parasite0	0.04	0.05	0.416
Parasite0: CR	-5.9e-17	0.07	1.000

Considering the results in Table 3, it was observed that the relative risk for severe malaria increased by exp(0.21) = 1.24 (HR) for unit increase in body weight of the child. Also, the risk of severe malaria is reduced by exp (- 0.03) = 0.73 (73.4%) in older children than in younger children. However, these results are not statistically significant in predicting the risk of severe malaria. The covariates: sex, and treatment were not statistically significant in predicting the risk of severe malaria.

All the covariates were not statistically significant in predicting the hazard for the competing risk event ‘ withdrawal’. It also observed that in the extended time-dependent Cox model, the baseline longitudinal markers of hemoglobin and parasite

counts showed no association with the hazards of severe malaria and the competing event withdrawal

4.4 Joint Modelling and Competing risks Models

In order to evaluate the relationship between the longitudinal scores of parasite counts and hemoglobin and the risk of severe malaria in the presence of competing risk withdrawal, one of the recommended approaches is to plot the longitudinal scores of parasites and hemoglobin seperated by event type that occured. This approach works when an interest is on survival outcome, and allows evaluation of impact of longitudinal outcome on survival (Hevia, 2014).

In Figure 8, the longitudinal progression of parasites count, and hemoglobin separated by the event type that a patient experienced with the fitted lines are shown. The results in Figure 8, showed that for the patients that experienced severe malaria, as the parasites were clearing probably due to patients taking the randomised medication, the hemoglobin levels for the patients were decreasing between day 0 and day 14. The behaviour was different for patients that were withdrawn, who had their hemoglobin level and parasite counts increasing between day 0 and day 10.

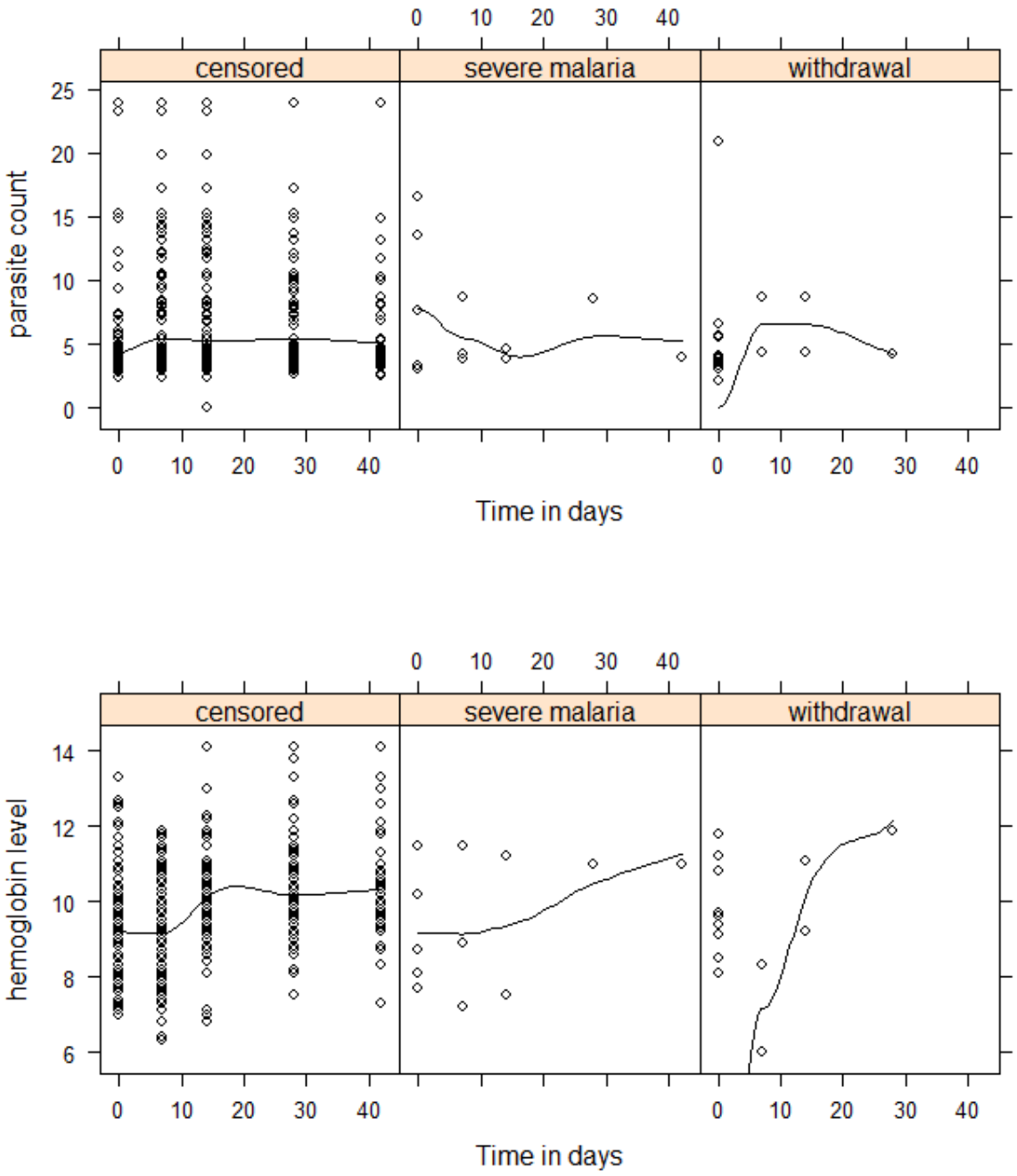

Figure 8: *Longitudinal scores showing the progression of hemoglobin , and parasite variables separated by the event type.*

Table 4: Estimates for competing risks survival and longutudinal parasite count processes in joint model settings

EVENT PROCESS			
Parameter	**RR**	**SE**	**p-value**
CQ + SP	2.33	1.23	0.491
CQ + SP : CR	1.00	1.14	0.999
AQ + SP	1.08	1.26	0.952
AQ + SP: CR	0.96	1.23	0.975
ART+ SP	1.08	1.28	0.953
ART+ SP:CR	0.99	1.20	0.998
Age	0.32	0.58	0.058
Age: CR	1.02	0.62	0.977
Weight	1.82	0.27	0.025
Weight: CR	0.99	0.26	0.988
Sexmale	6.75	0.93	0.040
Sexmale: CR	1.04	0.97	0.965
Assoct:	0.27	0.25	<0.0001
Assoct :CR	1.01	0.32	0.988
LONGITUDINAL PROCESS (parasite count)			
Slope			
Intercept	6.88	1.57	<0.0001
Time	-0.03	0.01	0.028
Sexmale	2.04	0.61	0.001
Age	-0.17	0.37	0.642
Weight	-0.18	0.16	0.261
CQ + SP	0.94	0.85	0.268
AQ + SP	-0.02	0.85	0.985
ART+ SP	1.38	0.88	0.116
Random effects			
Intercept		3.68	
Residual		2.97	
Log-Lik	-1134.2		
BIC	2344.3		

Results in Table 4, indicated that in joint modelling setting of longitudinal biomarker parasite count and the cause-specific hazard model processes, body weight of the child was significantly associated with the risk of severe malaria. With any unit increase in body weight, the relative hazard of severe malaria was increasing by 1.82 (HR). The true parasite count was also strongly associated with the risk of severe malaria, such that for a unit decrease in true parasite count, the relative hazard of severe malaria was decreasing by 0.27 (27.3%). The sex of an individual was also statistically associated with the risk of severe malaria, with relative risk of 6.76 (HR) higher in male patients than in female counterparts. The covariates, including the true parasite counts were not statistically significant in predicting the risk of withdrawal. On the longitudinal process, sex was strongly associated with the longitudinal scores of parasite counts, with two parasite counts more in male patients than in female patients. Also, with each passing day, the parasite counts were decreasing by -0.03 counts, as time was strongly associated with parasite counts. The intercept was also significant, with parasite counts of seven when other parameters in the model equal to zero.

Table 5: Estimates for a fitted joint model for longitudinal marker hemoglobin and competing risks survival processes

EVENT PROCESS			
Parameter	**RR**	**SE**	**p-value**
CQ + SP CQ + SP : CR	3.13 1.02	1.15 1.20	0.322 0.984
AQ + SP AQ + SP: CR	1.88 1.01	1.22 1.31	0.607 0.996
ART+ SP ART+ SP:CR	2.72 0.97	1.17 1.26	0.392 0.982
Age Age: CR	0.91 0.96	0.53 0.60	0.859 0.948
Weight Weight: CR	1.63 1.02	0.32 0.22	0.124 0.940
Sexmale Sexmale: CR	2.20 1.04	0.80 0.88	0.328 0.961
Assoct: Assoct :CR	0.06 0.98	1.26 0.78	0.029 0.982
LONGITUDINAL PROCESS (hemoglobin level)			
Slope			
Intercept	7.51	0.76	<0.0001
Time	0.02	0.01	<0.0001
Sexmale	0.08	0.24	0.752
Age	0.11	0.13	0.383
Weight	0.12	0.07	0.067
CQ + SP	0.25	0.30	0.406
AQ + SP	-0.12	0.31	0.704
ART+ SP	0.30	0.32	0.352
Random effects Intercept Residual		 0.86 1.20	
Log-Lik **BIC**	-767.13 1709.6		

Results in Table 5, indicated that the covariates, age, sex, and weight were not statistically significant in predicting the longitudinal hemoglobin scores. However, time and the intercept were statistically significant. With any passing day, the hemoglobin scores increase by 0.02 g/ul adjusting for other variables in the model. Also, without all other covariates, the hemoglobin level had a value of 7.51 g/ul. On competing risks survival process, all covariates were not statistically significant in predicting the risks of both severe malaria and withdrawal. In this setting, only the true hemoglobin level was significant in predicting the hazard of severe malaria, where the relative risk of severe malaria was reduced by 0.06 (6.0%) for a unit change in true hemoglobin level. The results also suggest that much of the variation was not resulting from subjects, as is indicated by the residuals.

4.5 Model Comparison

As shown in Table 2, Table 3, Table 4, and Table 5 of previous sections, it was observed that the separate linear mixed-effect model for parasite, only the intercept was associated with parasite counts. However, in the joint model process, the longitudinal process showed that the intercept, time, and sex were significantly associated with the longitudinal scores of parasite counts. Moreover, the estimates for all covariates had smaller standard errors in the joint model longitudinal process of parasite counts than in the separate mixed-effect parasite model. On the separate cause-specific Cox model, no variable was found to be significantly associated with the risks of severe malaria and withdrawal. To the contrary, in the joint model setting, the true parasite count was significantly associated with the risk of severe malaria in the presence of withdrawal. However, the coefficient estimates in the separate model had reduced standard errors than in the event process of the joint model.

The separate mixed-effect model for the hemoglobin scores, had time and the intercept strongly associated with the longitudinal marker hemoglobin. In the joint model setting, the same was the case. However, the estimates had small standard errors in the seperate models than the joint model setting. For the competing risks models, the true hemoglobin value was significant in the joint modelling setting. There was no covariate that was significant in separate cause-specific hazard model. The results also suggested elevated standard errors for the estimates in the joint model estimates for the survival process.

In order to compare the models, the log-likelihood estimates were used. The separate hemoglobin longitudinal model, had the log- likelihood -623.91, and the joint model had the log-likelihood estimate of -767.13, indicating that the separate model had the better fit to the data than the joint model. Also for the parasite count model, the separate model had log- likelihood of -988.24 where as the joint model had the log likelihood estimate -1134.2, thus the separate model preferable than the competing risks joint model. The same conclusion was reached, when estimates for Bayesian Information Criteria were used for the models. The choice of seperate model is also attributed to small standard errors in the separate models than the joint models, as models with small standard errors are preferable than models with large standard errors (McCrink *et al*., 2011).

CHAPTER 5

DISCUSSION

In the analysis of this data, for the mixed-effects model for hemoglobin level, time was significant in determining the hemoglobin level for each passing day. The increase in hemoglobin level as time passes could be due to clearence of parasites, hence reducing their attack of the red blood cells. Normally the Plasmodium *falciparum* survives up to four days in the host cells (White, 2017). In the longitudinal model for parasite count, none of the covariates considered showed significant results. This is contrary to observations in biomedical studies where in parasite clearence curve, time shows to be significant (White,2011).

For the survival processes alone in time-dependent Cox model, none of the covariates considered were associated with the risk processes. This suggested that when analyzing this data these covariates could not be considered. However in a different study, for malaria, age was found to be significantly associated with the risk of severe malaria, with higher odds of malaria in younger children (Nyirakanani *et al*.,2018), which suggested need to categorise age variable. In all risk models, no covariate was associated with the risk of withdrawal. This could be the case as withdrawal conditions might not be clinically associated the covariates used in this study.

The analysis revealed that when joint model of parasite count was fitted, body weight, sex and true parasite count were associated with the risk of severe malaria. The study found that the risk of severe malaria increased with an increase in body weight. This result was different from some work in literature where the higher parasitological response is expected to be higher in the underweight children than the overweight children (Djimde *et al*., 2019). This suggests need to categorize body weight when analysing malaria data. The higher risk in male children than female children could be due to more cases of male participants in the study resulting in slightly increased cases of severe malaria in males than in children. The significant result between true parasite count and risk of severe malaria is an effect traditionally presented in medicine, that as the parasite counts decrease, the risk of severe malaria also decreases as patients take the randomised drugs and have the hemoglobin levels increasing. There was also an improvement in the longitudinal process of parasite counts as time was statistically significant with a reduction in parasite count as time passed. This is what is clinically expected in biomedical studies of malaria. The parasite counts were higher in male children than in female children. This could be the case possibly due to more males in the study than females participants, as there is no biological association between the parasite count and gender of children (Nyirakanani *et al*.,2018).

For the joint model of hemoglobin level, in the risk process, only the true hemoglobin level was associated with the risk of severe malaria. For any unit change in hemoglobin level, the relative risk of severe malaria decreased. This could be the case due to reduction in the parasite counts as days passed. This could result into more hemoglobin levels hence reduced risk of severe malaria as work by Lombardo et al., (2017) yielded the same results. In the longitudinal process, time was also statistically significant with

increased hemoglobin level as time passed. The possibility in clinical studies is that time has detrimental effect on parasite count hence reduced attack of Parasitaemia *falciparum* on red blood cells.

For this data in general, the separate models seemed to be better to perform the analysis than the joint models with competing risks. This is attributed to lack of association between the risk processes and the baseline longitudinal markers in the time-dependent Cox models. Smaller standard errors for the seperate models also contributed to the choice of seperate models as models with small standard errors are preffered to models with large standard errors (Nguti et al., 2005). However, in literature, joint models considering competing risks by Hevia (2014), Hickey et al., (2017), and Andrinopoulou et al., (2014, 2017) were preffered for analysis of data.The choice was clear as there were associations between the longitudinal markers of interest in each study and the event or survival procesess in seperate time-dependent Cox models. In the work by Hevia (2014), the joint model had smaller standard errors than the seperate model, hence giving the joint models preference.

CHAPTER 6

CONCLUSION, RECOMMENDATIONS, AND LIMITATIONS

This chapter gives the summary of the results obtained in the analysis of malaria data using joint models with competing risks data. In it recommendations and possible limitations of the study are also included.

6.1 Conclusion

In clinical studies, it is common to collect survival data and longitudinal data. When an interest lies on association between the survival process and longitudinal markers, joint models are applied to the analysis of such data. From the results, it was observed that when analyzing the longitudinal markers of hemoglobin and parasite count in mixed effects models, time should be considered as it is statistically significant in predicting the scores of hemoglobin and parasite count. In light of these results, covariates treatment, sex, weight, age, parasite count and hemoglobin level may not be considered in seperate time-dependent Cox model. However, grouping weight and age may give some insight in the risk process as was seen in some literature. As higlighted in the joint models, it is important to consider both true parasite counts and true hemoglobin levels when assessing the risk of severe malaria in the presence of competing risk withdrawal.

When analysing the longitudinal malaria outcomes together with competing risks survival malaria outcomes in randomized controlled trials for malaria studies, seperate methods for longitudinal data and survival data can be used when there is no association

between baseline parasite count, hemoglobin level and the risks of severe malaria and withdrawal. However, joint models should only be considered when there is an association between the parasite count, hemoglobin and events' processes. Since there was no association between the longitudinal and survival processes, then separate models proved to be the better model fits to analyze these malaria outcomes data than the joint models with competing risks as shown by estimates of Bayesian information criteria.

6.2 Recommendations

In clinical studies for malaria, when longitudinal data and survival data are available, and there is no association between the survival process and the longitudinal process, then, the separate analysis of these data can be done. However, it is recommended that where the association does exist, use of joint models should be considered. For this malaria data, the use of separate models for longitudinal and competing risks survival malaria outccomes as there is no association between the events' process and longitudinal process. It is also recommended that doing the same study with different correlation levels between the survival and longitudinal outcomes, may improve the results possibly using simulation. Exploring newly developed methods with possible diagnostic methods may help to improve model selection procedures, and improve the results.

6.3 Limitations

In biomedical studies, where statistical tools are used, further progress is needed in this area of joint modelling of longitudinal data with competing risks survival data to advance tools for better analysis, as the field is in its early developmental stages, and

restricted in its application to biomedical studies. For instance, there is a need for development of diagnostic methods for model validation, selection and comparison, and also models that can include more than one longitudinal biomarkers of interest into a single competing risks joint model.

REFERENCES

Andrinopoulou, E.-R. (2014). Joint Modelling of Longitudinal and Survival Data with Applications to heart valve Data. *ISBN*, 978-994.

Bell, D. J., Nyirongo, S. K., Mukaka, M., Zijlstra, E. E., Plowe, C. V., Molyneux, M. E., Winstanley, P. A. (2008). Sulfadoxine-Pyrimethamine–Based Combinations for Malaria: A randomized Blinded trial to compare efficacy, Safety and Selection of Resisstance in Malawi. *PLos One*, 3(2), 1578.

Boor, C. D. (1978). *A practical Guide to splines.* Berlin: Springer.

Brown, E. R., & Ibrahim, J. G. (2003). A Bayesian Semi-parametric Joint Hierarchical Model for Longitudinal and Survival Data. *Biometrics, 59*, 221-228.

Collet, D. (2003). *Modelling Survival data in Medical Research.* London: Chapman & Hall/CRC.

Diggle, P. J., Heagerty, P., Liang, K.-Y., & Zeger, S. L. (2004). *Analysis of Longitudinal Data.* Great Britain: Oxford University Press.

Djimde, M., Samouda, H., & Djimde, A. A. (2019). Relationship between weight status and anti-malaria drug efficacy and safety in Children in Mali. *Malaria Journal*, 40.

Elashoff, R. M., 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.

Erango, M. A. (2018). Comparision analysis of seperate and Joint Models in case of time-to-death event of HIV/AIDS patients under ART follow-up. *Open Access Medical Statistics*, 8, 25-33.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied Longitudinal Data Analysis.* New Jersey: John Wiley & Sons Inc.

Gichangi, A., & Vach, W. (2005). The Analysis of Competing risks Data . *American Statistical Association,94 (446),* 496-509.

Gueorguieva, R., Rosenheck, R., & Lin, H. (2012). Joint modelling of longitudinal outcome and interval-censored competing risk dropout in a schizophrenia clinical trial. *J. R. Statist. Soc*, 175, 417–433.

Harrel, F. (2001). *Regression Modelling strategies: With Application to Linear Model, Logistic regression and Survival Analysis.* New York: Springer.

Harville, D. (1977). Maximum likelihood approaches to variance Component estimation to related problem. *American Statistical Association*, 72,320-340.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal Data Analysis.* Chicago: A John Wiley & Sons Inc.

Henderson, R., Diggle, P., & Dobson, A. (2000). Joint Modelling of Longitudinal Measurements and Event-time Data. *Biostatistics 1,4*, 465-480.

Hevia,A.F.(2014,July 8). *eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto.* Retrieved from eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto: http://www.eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto

Hsieh, F., Tseng, Y.-K., & Wang, J.-L. (2006). Joint Modelling of Survival and Longitudinal Data: Likelihood approach revisited. *Journal of the International Biometic Society, 62*, 1037-1043.

Ibrahim, J. G., Chu, H., & Chen, L. M. (2010). Basic Concepts and Methods for Joint Models of Longitudinal and Survival Data. *Clinical Oncology 28*, 2796-2801.

J.White, N. (2017). Malaria parasite clearence. *Malaria Journal*, 16:194.

Jennrish, R., & Schluchter, M. (1986). Unbalanced repeated measures models with structured covariance matrices. *Biometrics, 42(4)*, 805-820.

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data.* New Jersey: A John Wiley & Son Inc.

Kalbfleisch, J. D., & Prentince, R. L. (1980). *Statistical Analysis of failure time data.* New Jersey: A John Wiley and Sons Inc.

Kalbfleisch, J. D., & Prentince, R. L. (1980). *Statistical Analysis of failure Time Data.* New Jersey: A John Wiley & Sons.

Kalbfleisch, L. D., & Prentice, J. R. (2002). *Statistical Analysis of Failure Time data.*

Kaplan, E. L., & Meier, P. (1958). Non-parametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282): 457-481.

Kleinbaum, D. G., & Klein, M. (2005). *Survival Analysis: A self-learning text.* Antlanta GA 30306: Springer.

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

Lee, E. T., & Go, O. T. (1997, May 1). Survival Analysis in Public Health Reseach. *Annuals of Public Health*, pp. 105-134.

Lhatoo, S., Wong, I., & Sander, J. (2000). Prognostic factors affecting long-term retention of Topiramate in patients with chlonic epilepsy. *Epilepsia*, 338-341.

Liu, P. (2016). Retrieved from https://scholarcommons.sc.edu/etd/3829.

Lombardo, P., Vaucher, P., Rarau, P., Mueller, I., Faurat, B., & Seun, N. (2017). Risk of Malaria in Papua New Guinea in Infants. A nested cohort study. *The American society of tropical medicine and hygiene*.

Marchenko, Y. (2016, August 29). *Stata Corp LLC*. Retrieved from Stata Corp LLC[US]: https://www.stata.com.

McCrink, L., Marshall, A., & Cairnsk, K. (2011). Joint Modelling of Longitudinal and Survival Data: A comparision of Joint and Independent models, Int. *Session CPS044* (pp. 4971-4976). Dublin: World Statistical Congress.

Molenberghs, G., & Kenward, M. G. (2007). *Missing Data in Clinical Studies.* England: John Wiley & Sons .

Nelder, J., & Wedderburn. (1972). Generalized Linear Models. *A journal of Royal Statistical Society*, 370-384.

Nguti, R., Burzykowski, T., Rowlands, J., Renard, D., & Janssen, P. (2005). Joint Modelling of repeated measurements and event-time: An application to performance traits and survival of lambs bred in sub-humid tropic. *Genetic Selection Evolution: Biomedical Central*, 37(2), 175-197.

Prentice, R. L. (1982). Covariate Measurement errors and parameter estimation in a failure time regression model . *Biometrika*, 331-342.

Prousta-Lima, C., Dartigues, J.-F., & Jacqmin-Gadda, H. (2016). Joint latent class models for longitudinal and time-to-event data: a review. *Statistics in Medicine,* 23, 74-90.

Rizopoulos, D. (2010). JM: An R package for the Joint Modelling of longitudinal and time-to-event data. *Journal of Statistical Software,* 35(9),1-33. Retrieved from htpp://www.jstatsoft.org/v35/i09.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data with applications in R.* New York: CRC Press.

Scherzer, R. (2017, January 31). *khrc.ucsf.edu.* Retrieved from khrc.ucsf.edu: https://khrc.ucsf.ed.

Singh, R., & Mukhopadhyay, K. (2011). Survival Analysis in Clinical Trials : Basics and Must know areas. *STATISTICS*, 145-148.

Smith, P. J. (2002). *Analysis of Failure and Survival Data.* London: Chapman & Hall/CRC.

Stepniewska, K., & White, N. (2006). Some considerations in the designs and interpretation of antimalaria drug trials in uncomplicated falciparum malaria. *Malaria Journal*, 5,127.

Sudell, M., Kolamunnage-Dona, R., & Tudu-Smith, C. (2016). Joint Models for longitudinal and time-to-event data: a review of reporting qiality witha view to meta-analysis. *BMC Medical Research Methodology*, 16-168.

Tableman, M., & Kim, J. S. (2004). *Survival Analysis Using S.* Londom: Chaoman & Hall/CRC.

Tsiatis, A. A., & Davidan, M. (2004). Joint Modelling of Longitudinal and Time-to-event Data: An Overview. *Statistica Sinica*, 14, 809-834.

Twisk, J. W. (2003). *Appied Longitudinal Data Analysis for Epidemiology.* New York: Cambridge University Press.

Verbeke, G., & Lesaffre, E. (2009). Fully exponential Laplace approximations of the joint modelling of survival and longitudinal data. *Royal Statistical Society*, Series B 71, 637-654.

Weel, C. v. (2005). Longitudinal Research and Data Collection in Primary Care. *Annals of family Medicine*, s46-s51.

White, N. J. (2011). The parasite clearence curve. *Malaria Journal*, 10,278.

Williamson, P. R., Smith, C. T., Sander, J. W., & Marson, A. G. (2007). Importance of Competing Risks in the Analysis of Anti-epileptic drug failure. *Biomed Central, 8,12,* 1-10.

Wu, H., & Zhang, J.-T. (2006). *Non-parametric regression methods for longitudinal Data Analysis.* Hoboken, New Jersey: A John WIley & Sons, Inc.

Wulfsohn, M., & Tsiatis, A. (1997). A Joint Model for Survival and Longitudinal Data Measured with errors. *Biometrics*, 53, 330-339.

Zeger, S. L., Liang, K.-Y., & Albert, P. S. (1988). Models for Longitudinal Data. A Generalized Estimating Equation Approach. *Biometrics, 44*, 1049-1060.

Zheng, M., & Klein, J. P. (1995, March 1). *Biometrika.* Retrieved from Biometrika Web site: https://doi.org/10.1093/biomet/82.1.127.

Zwiener, I., Blettner, M., & Hommel, G. (2011). Survival Analysis. *PubMed*, 163-169.

APPENDICES

CODES USED FOR ANALYSIS IN R

PACKAGES USED

```
library(survival)
library(lattice)
library(splines)
library(foreign)
library(nlme4)
library(reshape2)
library(nlme)
library(JM)

library(cmprsk)
```

DESCRIPTIVES

```
summary(CMPDATA$age)
summary(CMPDATA$sex)/2
summary(CMPDATA$status)
summary(CMPDATA$parasite0)
summary(CMPDATA$hb0,CMPDATA$parasite0)
```

BOXPLOTS

```
boxplot(parasite~status2, ylab = "parasite count", xlab = "Event type", main="parasite Vs Event type", cex.main=0.9, cex.axis=0.7, las=1, ylim=c(0,14), data=RMW)


boxplot(hb~status2, ylab = "hemoglobin level", xlab = "Event type",
main="hemoglobin Vs Event type", cex.main=0.9, cex.axis=0.7, las=1, ylim=c(6,14),
data=RMW)
```

BAR PLOTS

```
attach(CMPDATA)
par(mfrow=c(1,3))


Table1<- table(sex, status3)
barplot(Table1, beside=T, legend.text = c("female", "male"), xlab="Event type", main
= "gender Vs event type", cex.main=0.9, cex.axis = 0.7)
```

```
Table2<- table(treat, status3)

barplot(Table2, beside=T, legend.text = c("SP", "SP+CQ", "SP+AQ", "SP+ART"),
xlab="Event type", main = "treat Vs event type", cex.main=0.9, cex.axis = 0.7)
```

LONGITUDINAL PROFILES: USING PACKAGE: lattice

```
xyplot(parasite~obtime|treat, group= patient, data=RMW, xlab=" Time (Days) ",
ylab="parasite count", col=1, type = "l", main = "individual parasite profile",
cex.main = 0.1)

xyplot(hb~obtime|treat, group= patient, data=RMW, xlab=" Time (Days) ", ylab=
"hemoglobin level", col=1, type = "l", main = "individual hemoglobin profile",
cex.main = 0.1)
```

LINEAR MIXED EFFECTS MODELS FOR PARASITES AND HEMOGLOBIN

```
FITP<-lme(parasite~ treat + sex + age + weight+ obtime, random = ~1|patient,data=
RMW)
FITH<-lme(hb~ obtime + treat + sex + age + weight, random=~1|patient,
data=RMW)
summary(FITH)
intervals(FITH) // Obtain 95% CI for coef
```

SURVIVAL ANALYSIS

CUMULATIVE INCIDENCE CURVES: Using cmprsk package

```
CUM<-cuminc(ftime = time,fstatus = status,rho = 0,cencode = "censored")

par(mfrow=c(1,3))

plot(CUM,xlab="time in days",  col=c(1,2))
```

COMPETING RISKS MODEL [Cause-specific time-dependent Cox Model]

```
NCOMP<-coxph(Surv(time, status2)~(treat + sex + age + sex + weight + hb0 +
parasite0)*CR + strata(CR), data = CMPDATA)
```

JOINT MODELS

```
xyplot(parasite ~ obtime | status, data = RMW, panel = function(x, y) { panel.
xyplot(x, y, grid = FALSE, type = c("p", "smooth"), col.line = "black") },ylab =
"parasite count",xlab = "time in days", pch=16)
```

```
xyplot(hb ~ obtime | status, data = RMW, panel = function(x, y) { panel.xyplot(x, y,
grid = FALSE, type = c("p", "smooth"), col.line = "black") }, ylab = "hemoglobin
level",xlab = "time in days", pch=16)
```

THE JOINT MODELS IN THE PRESENCE OF COMPETING RISKS

```
JOINTH<-jointModel(FITH, COMPCOX, timeVar = "obtime", method = "spline-PH-
aGH", CompRisk = TRUE, interFact = list(value=~CR, data= CMPDATA))
```

```
JOINTP<-jointModel(FITP, COMPCOX, timeVar = "obtime", method = "spline-PH-
aGH", CompRisk = TRUE, interFact = list(value=~CR, data= CMPDATA))
```

DATA OBJECTS IN THE PROJECT R FILE

BCOMP: This contains the fitted Extended Cox Model with baseline longitudinal markers

FITH: Linear Mixed-effect Model for hemoglobin

FITP: Linear Mixed-effect Model for parasite

JCOX: Cox model used in the joint modelling

JOINTH: Hemoglobin Competing Risks Joint Model

JOINTP: Parasite Competing Risks Joint Model

CMPDATA: This is the competing risks Data used in the Analysis of this project

RMW: This is the Longitudinal Data for parasites and hemoglobin used in the analysis