APPLICATION OF THE EXTENDED COX-PROPORTIONAL HAZARDS MODEL FOR THE ANALYSIS OF MASS ORAL AZITHROMYCIN FOR REDUCTION OF CHILDHOOD MORTALITY IN THE PRESENCE OF TIME-DEPENDENT COVARIATES

MASTER OF SCIENCE (BIOSTATISTICS) THESIS

ALVIN BLESSINGS CHISAMBI

UNIVERSITY OF MALAWI

DECEMBER, 2022

APPLICATION OF THE EXTENDED COX-PROPORTIONAL HAZARDS MODEL FOR THE ANALYSIS OF MASS ORAL AZITHROMYCIN FOR REDUCTION OF CHILDHOOD MORTALITY IN THE PRESENCE OF TIME-DEPENDENT COVARIATES

MASTER OF SCIENCE (BIOSTATISTICS) THESIS

By

Alvin Blessings Chisambi

BSc. (Statistics and Computing)- University of Malawi

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

University Of Malawi

December, 2022

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 people's work has been used acknowledgements have been made.

ALVIN BLESSINGS CHISAMBI

Full Legal Name

_____

Signature

_____

Date

CERTIFICATE OF APPROVAL

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

Signature: _____Date:_____

Mavuto Mukaka, PhD (Professor)

Main Supervisor

Signature:_____Date:_____

Patrick Sawerengera,

Programme Coordinator

DEDICATION

To my beloved wife – Ellen Phiri Chisambi, you can do this!

ACKNOWLEDGMENTS

I would like to give it all to my supervisor Prof. Mavuto Mukaka, PhD for his guidance, constructive ideas and for never giving up on me when I allowed work to take more of my time on this thesis. I really appreciate the encouragement and his critiques on my work which has helped me structure the entire thesis with understanding. If it was not for him, I wouldn't have been submitting this today! If we had supervisors like him, then more students would graduate on this program in Malawi. It is all to you dear Prof.

I acknowledge Prof. Khumbo Kalua for his financial support and mentorship in operational research. This can't go without mentioning Prof. Robin Bailey for providing this dataset from MORDOR clinical trial and Dr John Hart for structuring the dataset for this thesis. I appreciate their assistance.

ABSTRACT

The famous Cox proportional hazard model is applied in most medical research studies that involve time to event data like mass oral azithromycin for the reduction of childhood mortality. Literature has shown that this model overestimates estimates in the presence of time varying covariates. This thesis compared the Cox proportional hazard model to Extended Cox model that account for time varying covariates. The models were applied to model the effect of age, weight, and treatment in relation to death as the outcome of interest. The study findings showed that the Cox PH model overestimated the effect of the covariates to the hazards of a participant dying as it did not take into consideration the presence of time dependent covariates in the data as compared to the extended Cox model. Kaplan Meir survival curves were plotted to compare survival in the two study arms (Placebo and Azithromycin drug groups). The hazard of death was associated with covariates age, weight, and treatment-received in both study arms. The results from the Cox model showed that the expected hazard is 13.56 (4.89, 37.64) times higher in a person who is one year older than another. For the variable weight, the expected hazard is 0.763 (0.64, 0.90) times lower risk reduction in the drug group as compared to placebo group. For the variable treatment received, the expected hazard is 0.03 (0.02, 0.08) times lower risk reduction in the drug group as compared to placebo group for those treated than those who did not receive treatment. Whereas in the extended Cox model, the main model showed that the expected hazard is 3.74(0.53, 26.35) times higher in a person who is one year older than another, the expected hazard is 21.01(6.82, 64.72) times higher for those who did not receive drug than those who received drug (***p*=<0.001)**. The time varying covariates model showed that the expected hazard is 1.24(0.84, 1.83) times higher in a person who is one year older than another. The expected hazard is 1.85(1.48, 2.31) times higher in those who did not receive drug than those who received drug (***p*=<0.001)**. The Extended Cox model was a better model when studying data that involves time varying covariates to avoid reporting overestimated estimates.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

This chapter presents a brief background of the study area, the problem statement, study objectives and its significance in the theoretical and empirical knowledge.

1.1 Background

World health Organization (WHO) through one of the fact sheets released in 2022 found that Trachoma, an eye disease that causes blindness of around 1.9 million people, is hyperendemic in the poorest/rural areas of most parts of the world. The statistics indicate that Trachoma is one of the main causes of 1.4% of blindness in the whole world. Programs use the SAFE strategy to eliminate blinding Trachoma in the world including Malawi which was declared free of blinding Trachoma in 2020. The "A" in the SAFE strategy is Antibiotic (Azithromycin). Studies done in Ethiopia had shown that Azithromycin played a big role in reducing childhood mortality.

According to a 2021 report by United Nations Inter-agency group for Childhood Mortality Estimation (UN IGME), it was stated that despite the global efforts, the world appears to be significantly off track to achieve the Sustainable Development Goals (SGDs) on preventing childhood mortality for under-fives. As per the findings of this report, more than 5 million children died before the age of 5 in the year 2020 alone. The report further states that SDGs aims at having a neonatal mortality rate of 12 or fewer deaths per 1,000 live births, and an under-five mortality rate of 25 or fewer deaths per 1,000 live births for all countries including Malawi, by the year 2030. Research has shown that three-quarters of all childhood deaths occur after the neonatal period, with malaria, pneumonia, and diarrhoea accounting for the majority. A study done by (Sazawal S et al., 2003 and Moonen B et al., 2010) showed that Management is often case detection and individualized therapy, but concluded that novel strategies such as mass treatment may play a role. Studies have shown that mass drug administration (MDA) of

azithromycin has proven to reduce childhood mortality as it appears to have collateral benefits against several diseases (Keenan, J.D. et al., 2015)

A Larger cluster randomized controlled trial was done in three countries (Malawi, Niger, and Tanzania) to see the effect of Azithromycin on reduction of childhood mortality. Communities were randomized to placebo and Azithromycin groups. In their study, they found that Azithromycin had collateral benefits against several diseases that affects the under five children (Keenan, J.D. et al.). The nature of the study was done in phases and individuals were followed until the event of interest; death occurs. This kind of data is called Survival data and researchers, or investigators use survival methods of analysis to analyze data of this nature.

Survival studies follow subjects until an event of interest occurs and measure explanatory variables for that event, sometimes repeatedly over the course of follow up. Researchers have mostly applied or used the Cox regression model in the analyses of time to event data. "The associations between the survival outcome and time dependent measures may be overestimated unless they are modeled appropriately to avoid wrong recommendations (Ngwa et al., 2016).'' The Survival methods of analysis, for example, Cox regression model has been used in several Mass Oral Azithromycin for childhood mortality studies. For example, (Porco, Travis C. et al.) used Cox regression model to investigate the effect of Mass Oral Azithromycin on Childhood mortality. Another study by (Keenan, J.D. et al.) also looked at the effect of Mass Azithromycin distribution for reducing childhood mortality in sub-Saharan Africa.

In this thesis we explore the Time Dependent Cox Regression Model, which quantifies the effect of repeated measures of covariates in the analysis of time to event data (Ngwa et al., 2016). Literature has shown that the Time Dependent Cox Regression Model is commonly used in biomedical research but sometimes if not handled properly, it does not explicitly adjust for the times at which time dependent explanatory variables are measured (Barnett et al., 2011). If the this is overlooked in the analysis, it can yield different estimates of association compared to using a model that adjusts for these times. To address the question of how different these estimates are from a statistical perspective, we compare the traditional Cox Proportional Hazards Model to extended Cox model, considering models that adjust and do not adjust for time.

1.2 Problem Statement

Epidemiological or biomedical studies requires investigators or researchers are to evaluate the effect of exposures, such as antibiotics, on clinical outcomes (Munoz-Price et al.). However, many of these fluctuating exposures occur at intervals and are not present throughout the entire time of observation. These "fluctuating" variables that occur at intervals are called time-dependent variables. Researchers or investigators need to be careful when performing analysis on these kind of variables by incorporating time-dependent exposure status in the statistical models that are being fitted. When these time dependencies of antibiotic exposures are not handled properly or ignored when fitting these models, one might end up with incorrect or overestimated estimates of both hazards and hazard ratios (Munoz-Price et al.). Despite the availability of the traditional Cox Proportional Hazards model and its capacity to incorporate time-dependent covariates, investigators do not often utilize this.

There has been growing literature on survival models for analysis of time-to-event data in medical research which handles both time independent covariates and time dependent covariates. Studies in the past have employed different statistical techniques such as Cox regression model, Kaplan Meier estimate, log-rank test, Logistic regression to study childhood mortality. However, few studies tend to account for the presence of time varying covariates in the models used in analyzing the data which may result in overestimation.

Sometimes just using this simple Cox proportional hazard model would not be ideal in situations where the hazards are not proportional like in the cases where time-varying covariates are present. A study done by (Zhang Z et al, 2018) on Time-varying covariates and coefficients in Cox regression models, recommended that when time varying covariates or coefficients are present, an analyst should consider taking them into account in survival modelling in order to improve the estimation. This thesis will use extended Cox proportional hazard model which takes into account time varying covariates present in analyzing child mortality data after mass oral azithromycin to compare the estimates with that of just computing traditional Cox proportional hazard model.

1.3 Study Objectives

This section presents the main and specific objectives of the study.

1.3.1 Main Objective

To fit and compare survival models that do not account for time varying covariates and extended Cox regression model that account for time varying covariates for childhood mortality data in Mangochi, Malawi.

1.3.2 Specific Objectives

1. To estimate and compare survival functions in the treatment and placebo arm between the phases and between those aged less or more than 6 months at treatment
2. Estimate and comparing hazards at each phase with time fixed vs time dependent exposure.
3. To assess if there is a difference in inference from standard survival methods and the methods that adjust for time-varying covariates.
4. To assess the validity of fitted model assumptions.

1.4 Significance of the study

Analysis of time-varying covariates without accounting for them results in overestimation which may make researchers come up with wrong conclusions. Therefore, modelling using standard methods that do not account for time-varying covariates would not give analysts true picture of what is on the ground. The study contributes to available work done in the field of survival analysis when modeling survival time while accounting for time varying covariates. The study will assess whether childhood mortality can be reduced using a novel approach—mass administrations of azithromycin like those used for trachoma. Furthermore, results of the research will help researchers understand appropriate methods to use when modelling survival data with time-varying covariates.

In this chapter, we have presented a brief background of Childhood mortality reduction after Oral azithromycin, the problem statement, objectives, and significance of the study. The

subsequent chapters present the literature review of the study area, the methodology used in the study, the results and discussion of the study and conclusion and recommendation(s).

CHAPTER 2
LITERATURE REVIEW

This section presents an overview of survival analysis, reviews survival methods and approaches in survival data analysis and introduces models that have been developed in modelling survival data. In detail, this chapter reviews survival function, hazard function, hazard ratio, Kaplan Meier (KM) methods, tests for survival analysis, Models in survival analysis, model parameter estimation, model comparisons (Model diagnostics) and handling of time-varying covariates and competing risks approach.

2.1 Overview of Survival Analysis

(Machin, Cheung, & Parmar, 2006) defines Survival analysis as "a set of methods for analyzing data where the result variable is the time until the occurrence of an event of interest. They further state that an event can either be death, occurrence of a disease, marriage, divorce, etc." Cornel Statistical Consulting Unit states that "the time to event or survival time are often measured in days, weeks, years, etc. as an example, if the event of interest is heart failure, then the survival time can be the time in years until someone develops a heart attack (Altman, 1991)".

Failure time or survival time, as well as event time is called the response variable of survival analysis. To analyze data in which the time until the event is the main outcome or of interest, researchers/study Investigators use survival methods of analysis to analyze such data. The failure time is usually continuous and hard to be determined for some subjects, 'that is for some subjects we may know that their survival time was at least equal to sometime *t*; Whereas, for other subjects, we will know their exact time of event (Altman, 1991).'' "Survival time responses that have not been observed completely are said to be censored, and if there is no censoring, standard regression procedures could be used to predict the outcome (Machin,et.al 2006).''

Other lectures and papers in introduction to survival analysis indicate that time to event is restricted to be positive and has a skewed distribution making these to be inadequate. ‘’The probability of surviving past a particular point in time could be of more interest than the expected time of event, in this case, the hazard function used for regression in survival analysis, can lend more insight into the failure mechanism than simple regression (Kleinbaum & Mitchel, 1996).’’
In survival methods of analysis, subjects are followed over a specified period and focus being on the time at which the event of interest occurs.

2.2 Censoring

Survival data is sometimes censored, censoring occurs when information about the survival time of some individuals is incomplete (Kleinbaum et. al,1996). Censoring is an important issue in survival analysis; it represents a specific form of missing data. The foremost ideal data for survival analysis are those yielded by cases within which the time of treatment is clearly established, and all participants are followed up until they experience the event (In, Junyong and Dong Kyu Lee. “Survival analysis: Part 1 – analysis of time-to-event,” page 183). A lecture in Introduction to Survival Analysis indicates that there are generally three reasons why censoring might occur; a subject does not experience the event before the study ends; a person is lost to follow-up during the study period and a person withdraws from the study. (Kleinbaum et. al,1996).

There are three general types of censoring, right and left and interval censoring (Leung. et al., 1997). (In, Junyong and Dong Kyu Lee. “Survival analysis: Part 1 – analysis of time-to-event,” page 183) states that t right censoring is the common form of of censored data that researches mostly encounter, within which the event of interest does not happen to the subject at the end of the study period (end-of-study censoring) or the observation is terminated for reasons other than death (loss-to-follow-up censoring) (Kleinbaum et. al,1996). Right-censored data can increase the estimated overall survival time but may cause bias. Interval-censoring occurs in survival analysis when the time until an event of interest is not known precisely (and instead, only is known to fall into a particular interval). Such censoring commonly is produced when periodic assessments (usually clinical or laboratory examinations) are used to assess if the event has occurred. Left censoring concerns cases with unclear first exposure to the treatment event prior

to inclusion in the study. Left censored data can occur when a person's survival time becomes incomplete on the left side of the follow up period. Censored observations may not only be due to losses to follow-up or administrative cessation of the period of consideration but can also be due to events not of interest. This situation is problematic if these "other events" preclude observation of the primary event under consideration. In survival analysis, censoring should be non-informative (participants who drop out of the study should do so due to reasons unrelated to the study.) Informative censoring occurs when participants are lost to follow-up due to reasons related to the study, for example in a study comparing disease-free survival after two treatments for cancer, the control arm may be ineffective, resulting in more recurrences and patients becoming too sick to follow-up.

2.3 Survival Function, S(t)

According to (In, Junyong and Dong Kyu Lee. "Survuval analysis: Part 1 – analysis of time-to-event," page 183), they defined survival function as the probability of the outcome event not occurring up to a specific point in time, including the time point of observation (t), and is denoted by S(t). That is, if the event is "recurrence of back pain," it is the "probability of not having back pain" up to a specific time. In the survival function, t = 0 corresponds to a probability of 1.0 (i.e., 100% survival at the onset), and the point in time with 50% survival probability is the median survival time. (Kleinbaum et. al,1996).

Let T be a non-negative random variable denoting the time to a failure event. The survivor function $S(t)$ gives the probability that a person survives longer than some specified time ***t:*** that is, $S(t)$ gives the probability that the random variable T exceeds the specified time t (Kleinbaum & Klein, 2005). In other words, the survivor function also known as survivorship function is simply the reverse of the cumulative probability function of T. Where the cumulative distribution is given by

$$F(t) = \Pr(T < t) = \int_0^t f(u)du \qquad (1)$$

and the survival function, S(t), is defined to be the probability that the survival time is greater than or equal to t, and is given by

$$S(t) = P(T > t) = 1 - F(t) \qquad (2)$$

It is simply the probability that there is no failure event prior to time *t*. The survivor function can therefore be used to represent the probability that an individual survives from the time origin to sometime beyond t. The function is equal to 1 at *t=0* and decreases toward zero as *t* goes to infinity (∞). The probability density function is expressed as;

$$f(t) = \frac{dF(t)}{dt} = \frac{d\{1-s(t)\}}{dt} = -S'(t) \qquad (3)$$

The hazard and survival functions are alternative forms of describing the distribution of survival times. The survival time is most useful for comparing the survival progress of two or more patient groups, the hazard function since it is an instantaneous measure gives a more useful graphical description of the risk of failure at time t. Hazards and survival functions can be expressed in terms of each other (Machin, Cheung, & Parmar, 2006).

2.4 Hazard Function, h(t)

According to Cleves, Mario's Introduction to Survival Analysis Using Stata, 2nd edition, page 7, they define "the hazard function or conditional failure rate as the instantaneous rate of failure. It is the limiting probability that the failure event occurs in each interval, conditional upon the subject having survived to the beginning of that interval, divided by the width of the interval (Cleves. et al., 2010)."

The hazard function is widely used in survival analysis by researchers to express the risk or hazard of death at some time *t* and is obtained from the probability that an individual dies at time *t*, conditional on he or she having survived to that time (Lalanne & Mounir, 2016). (Collect, David, in his book called Modelling Survival data in medical Research, 2nd edition states that" the ratio of the number of events occurring during the entire study period to the total number of observations is termed the "incidence rate." For example, if the event is death, mortality is the incidence rate. However, since the incidence may not be constant throughout the study period, it may be necessary to calculate the incidence rate at a specific time (*t*)."

First, the incidence rate for the period between a specific time t and the next measurement time $t + \alpha$ can be calculated by dividing the number of events occurring between *t* and $t + \alpha$ by the total number of observations at time *t*. By α approaches 0, i.e., by taking the limit as the interval between *t* and $t + \alpha$ closest to *0*, the instantaneous incidence rate at t, which constitutes

the hazard, can be calculated. The hazard function is a function for calculating the instantaneous incidence rate at any given point in time and is denoted by h (*t*). "The function h(t) is also referred to as the hazard rate, the instantaneous death rate, the intensity rate, or the force of mortality. (Cleves. et al., 2010)." (In, Junyong and Dong Kyu Lee. "Survival analysis: Part 1 – analysis of time-to-event," page 184) .

In simple terms it is the probability that an individual encounters an event of interest at time t, conditional on having survived to that time. If $\boldsymbol{t}$ is a continuous function with density function $\boldsymbol{f}$, then the hazard function is defined by:

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t < T \leq t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{S(t)} \qquad (4)$$

Thus, the hazard function, *h(t)*, is the instantaneous rate at which events occur, given no previous events.

2.5 Hazard Ratio (HR)

"The hazard ratio has been specifically developed for survival data and is used as a measure of the relative survival experience of two groups" (Machin, D. et al. "Survival Analysis: A Practical Approach," page 12). Survival analysis hazard ratio aims at obtaining the ratio of two hazards, it compares the hazard of one group against the hazard of another. This method allows for the censoring which occurs in nearly all survival data (Machin, Cheung, & Parmar, 2006). For instance, suppose that patients are randomized to receive either a standard treatment or a new treatment, and let $h_s(t)$ and $h_N(t)$ be the hazards of death at time t for patients on the standard treatment and new treatment, respectively. This proportional hazard model can be expressed in the form

$$h_N(t) = \emptyset h_s(t), \qquad (5)$$

For any non-negative value of t, where $\emptyset$ is a constant. An implication of this assumption is that the corresponding true survivor functions for individuals on the new and standard treatments do not cross. The value of $\emptyset$ is the ratio of the hazards of death at any time for an individual on the new treatment relative to an individual on the standard treatment, and so $\emptyset$ is known as the relative hazard or hazard ratio.

The hazard ratio is the ratio of the hazard rates corresponding to the conditions described by two levels of an explanatory variable as defined by (Hazard Ratio- Wikipedia). The hazard ratios represent instantaneous risk over the study time period or some subset thereof. If a $\emptyset < 1$, the hazard of death at time t is smaller for an individual on the new drug, relative to an individual on the standard. The new treatment is then an improvement on the standard. On the other hand, if $\emptyset > 1$, the hazard of death at time t is greater for an individual on the new drug, and the standard treatment is superior. The hazard ratio, $\emptyset$, of 1 corresponds to equal hazards between the two groups. While a hazard ratio, $\emptyset$, of 2 implies that at any time twice as many in the treatment group are having an event proportionately compared with the control group (Deurden., 2009).

2.6 Non-time-varying covariates

Literature shows that the Cox regression model has been used widely in the analyses of time to even data. The Cox proportional hazards (PH) model is the widely used approach in survival analysis of clinical trials because it requires a few assumptions (Ngwa, et al., 2016). The model was proposed by Cox in 1972 for analysis of survival data with and without censoring; for identifying differences in survival due to treatment and prognostic factors in clinical trials (Singh, Ritesh and Keshab Mukhopadhyay. "Survival analysis in clinical trials: Basics and must know areas," page 146). The Cox proportional hazards (PH) model allows one to explain the survival time as a function of multiple prognostic factors (Lalanne & Mounir, 2016). This model relies on a fundamental assumption that the proportionality of the hazards, implying that the factors investigated have a constant impact on the hazard - or risk - over time. However as observed by Ponnuraja and Venkatesan (2010), this is not appropriate all the time as the assumptions do not hold always. If time-dependent variables are included without appropriate modeling, the PH assumption is violated and this leads to deriving misleading effect estimates, and significant effect in the early (or late) follow-up period may be missed (Bellara, MacGrogan, Debled, & Brouste, 2010). Unfortunately, most researchers in practice often do not test models' assumptions for fitted models and this practice does not spare the fitting of Cox PH model (Stanley, Molyneux, & Mukaka, 2016). Checking the proportionality of the hazards should thus be an integral part of a survival analysis by a Cox model. (Bellera, Carine et al. "Variables with time-varying effects and the Cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer").

“Models that can accommodate time-dependent covariates are commonly used in biomedical research but sometimes do not explicitly adjust for time in the model. Not adjusting for time can yield different estimates of association compared to a model that adjusts for time.” (Ngwa, et al., 2016). When analysing survival data and time-varying covariates or coefficients are present, an analyst should consider taking them into account in survival modeling to improve the estimation (Zhang, Zhongheng, 2018).

2.7 Time-varying covariates

A time-dependent variable is defined as any variable whose value for a given subject may differ over time (t). In contrast, a time-independent variable is a variable whose value for a given subject remains constant over time (Kleinbaum & Mitchel, 1996). When explanatory variables do not change over time or when data is only collected for explanatory variables at one time point, it is appropriate to use static variables to explain the outcome. On the other hand, there are many situations where it is more appropriate to use time varying covariates. Using time varying explanatory variables, when appropriate, is more robust because it utilizes all available data (Allison, 2010).

Time-varying covariance occurs when a given covariate changes over time during the follow- up period, which is a common phenomenon in clinical research (Zhang, Zhongheng, 2018). “For example, in a patient with sepsis, the C- reactive protein (CRP) may be measured repeatedly to evaluate inflammatory status until it returns to normal. In clinical oncology, the recurrence status of a patient is usually checked at a predefined time interval. In many cases when studying the relation between a survival outcome and covariate(s), investigators will only consider the baseline value of the covariate, which however, fails to consider the relation of the survival outcome as a function of the change of the covariate. For example, the effect of smoking on status is ever changing during the follow up period. Such a covariate can be considered as a time -varying covariate.” (Zhang, Zhongheng, 2018).

2.8 Models in Survival Analysis

This subsection explores the relationship between survival experience of an individual and explanatory variables using an approach based on statistical modelling. Two broad classes of

regression models are considered: Proportional Hazard (PH) Models and Accelerated Failure Time (AFT) Models. Models used to describe survival time in a comparative sense are often called semi-parametric regression models and are the major focus of this thesis. We will also distinguish between semi-parametric and parametric models but for the purpose of this thesis, Cox PH models and Extended Cox models will be discussed in detail. These models suggested in the literature include the Cox semiparametric proportional hazard model and some parametric models like the exponential model, and Weibull Model, Gomperz (Gamma) and Log-Normal model.

2.8.1 Non-parametric models

The first approach in survival model fitting is a non-parametric strategy that focuses on estimation of the regression coefficients leaving the baseline hazard $\partial_o(t)$ completely unspecified. This approach relies on a partial likelihood function proposed by Cox (1972) in his original paper.

2.8.1.1 Kaplan-Meier survival estimate

The Kaplan–Meier estimator is the nonparametric maximum likelihood estimator of the underlying survival function (Kaplan and Meier, 1958). The Kaplan-Meier (KM) method is a non-parametric method used to estimate the survival probability, S(t), from observed survival times (Kaplan and Meier, 1958). It is a popular method because it requires very weak assumptions (assumes no form of distribution) but utilizes information content of both fully observed and right censored data. Suppose that n individuals have experienced an event of interest, such as death in a group of individuals. If we let $0 \leq t_1 < \cdots < t_n$be the observed ordered death times. Let n_i be the number of individuals who are at risk at $t_{(n)}$. Let z_i be the number of observed deaths at t_i, i= 1…n. Then the Kaplan Meier estimate at any time t is given by

$$\hat{S}(t) = \prod_{i:\tau_{i\leq t}} \frac{r_i - z_i}{r_i} = \prod_{i:\tau_{i\leq t}} 1 - \frac{z_i}{r_i} \ , \qquad (6)$$

where r_i is the number of individuals at risk at time t_i, and the product is overall observed failure times less than or equal to t (Kaplan. & Meier., 1958). The estimator is a step function that

changes values only at the time of each. It is also possible to compute confidence intervals for the survival probability. The KM survival curve, a plot of the KM survival probability against time, provides a useful summary of the data that can be used to estimate measures such as median survival time. (Etikan, Abubakar, & Alkassim, 2017)

2.8.1.2 Log-rank test or Mantele-Haenzel test

Another possible objective of the analysis of survival data may be to compare the survival times of two or more groups. A simple test of statistical significance is the log-rank or Mantele-Haenzel test. Comparison of two survival curves can be done using a statistical hypothesis test called the log rank test. It is used to test the null hypothesis that there is no difference between the population survival curves (i.e. the probability of an event occurring at any time point is the same for each population). The test statistic is calculated as follows:

$$x^2(\log rank) = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \qquad (7)$$

Where the O_1 and O_2 are the total numbers of observed events in groups 1 and 2, respectively, and E_1 and E_2 the total numbers of expected events.

The log rank test is used to test whether there is a difference between the survival times of different groups, but it does not allow other explanatory variables to be taken into account or considered.

2.8.1.3 Partial likelihood estimator

Cox (1972, 1975) introduced the ingenious partial likelihood principle to eliminate the infinite dimensional base-line hazard function from the estimation of regression parameters with censored data. In a seminal paper, Andersen and Gill (1982) extended the Cox regression model to general counting processes and established the asymptotic properties of the maximum partial likelihood estimator and the associated Breslow (1972) estimator of the cumulative base-line hazard function via the elegant counting process martingale theory. The maximum partial likelihood estimator and the Breslow estimator can be viewed as non-parametric maximum likelihood estimators (NPMLEs) in that they maximize the non-parametric likelihood in which the cumulative base-line hazard function is regarded as an infinite dimensional parameter (Andersen *et al.* (1993), pages 221–229 and 481–483, and Kalbfleisch and Prentice (2002), pages 114–128).

Intuitively, it is a product over the set of observed death times of the conditional probabilities of seeing the observed deaths, given the set of individuals at risk at those times. It is given by

$$l_p(\theta) = \prod_{i=1}^{n} \frac{e^{x_{(i)}\theta}}{\sum_{j \epsilon R(t_{(1)})} e^{x_j \theta}}, \qquad (8)$$

Where the product is over n death (or failure) times. The contributions occur only at death times. The partial likelihood is not a product of independent terms, but conditional probabilities.
In 1975 Cox provided a more general justification of L as part of the full likelihood—in fact, a part that happens to contain most of the information about _—and therefore proposed calling L a partial likelihood. This justification is valid even with time-varying covariates. A more rigorous justification of the partial likelihood in terms of the theory of counting processes can be found in Andersen et al. (1993).

2.8.1.4 The Nelson-Aalen Estimator

Consider estimating the cumulative hazard $\hat{\theta}(t)$. A simple approach is to start from an estimator of S(t) and take minus the log. To estimate and plot the cumulative hazard function, the Nelson-Aalen estimator can be used. The Nelson-Aalen estimator is a non-parametric estimator of the cumulative hazard function;

$$\hat{\theta}(t_{(i)}) = \sum_{j=1}^{i} \frac{d_j}{r_j}, \qquad (9)$$

Intuitively, this expression is estimating the hazard at each distinct time of death $t_{(j)}$ as the ratio of the number of deaths to the number exposed. The cumulative hazard up to time t is simply the sum of the hazards at all death times up to t, and has a nice interpretation as the expected number of deaths in (0, t] per unit at risk. This estimator has a strong justification in terms of the theory of counting processes.

Breslow (1972) suggested estimating the survival function as

$$\hat{S}(t) = exp\ \{-\hat{\theta}(t)\}, \qquad (10)$$

where $\hat{\theta}(t)$ is the Nelson-Aalen estimator of the integrated hazard. The Breslow estimator and the K-M estimator are asymptotically equivalent, and usually are quite close to each other, particularly when the number of deaths is small relative to the number exposed.

2.9 Semi-parametric models

The second approach to survival model fitting is regarded to be a flexible or semi parametric strategy, where mild assumptions about the baseline hazard $\theta_o(t)$ are applied. Using this approach, time is subdivided into reasonably small intervals and the assumption that the baseline hazard is constant in each interval is applied leading to a piecewise exponential model (Altman, 1991).

2.9.1 Cox Proportional Hazard Model

The basic model of consideration in this thesis is the proportional hazard model proposed by cox(1972). The model has come to be known as the cox regression model; although assumptions of proportional hazards are the base for this model, there is no form of a probability distribution that is assumed for the survival times, and this model is therefore referred to as semi-parametric model. Apart from the mentioned assumptions, the Cox assumes that the hazards are proportional and from model estimation it uses partial likelihood which is more generalized than the maximum likelihood (Hosmer, Lemeshow, & May, 2006).

This is the most common approach used in research to model the effects of covariates on survival when analyzing survival data, and since the only assumption made is on the proportionality of the baseline hazard; it therefore means that the hazard ratio is constant over time or that the hazard for an individual is proportional to the hazard for any other individual. (Therneau & Grambsch, 2000). The model can be used for comparison of the hazard functions for individuals in two groups as it also takes into account the effects of censored observation. (Cox., 1972).

Defined as: Let $x_1 \dots \dots \dots x_k$be the values of n covariates $X_1 \dots \dots \dots . X_k$, the hazard function is then given as follows as this model (Cox regression model).

$$h(t) = h_0(t) \exp\left(\sum_{i=1}^{k} \delta_i x_i\right), \qquad (11)$$

Where $\delta_i = (\delta_1, \delta_2, \delta_n)$ is a (1 × k vector of regression coefficients and $h_0(t)$ is the baseline hazard function at time t.

The study goal is addressed by the regression model for the hazard function below:

$$h(t, x, \delta) = h_0(t) r(x, \delta), \qquad (12)$$

Equation (12) above is the product of hazard function $h_0(t)$and the function of subject covariates $r(x, \delta)$. The way how the hazard function changes as a function of survival time is given by the hazard function $h_0(t)$. While the function $r(x, \delta)$, characterizes how the hazard function changes as a function of subject covariates (Hosmer & Lemeshow, 1999). The functions should be chosen such $h(t, x, \delta) > 0$. Note that $h_0(t)$ is the hazard function when $r(x, \delta) = 1$. When the function $r(x, \delta)$ is such that $r(x = 0, \delta) = 1, h_0(t)$ is frequently referred to as the *baseline hazard function.* ***T***he ratio of the hazard functions for two subjects with covariate values denoted x_1 and x_0 in equation (12) is given by;

$$HR(t, x_1, x_0) = \frac{h(t, x_1, \delta)}{h(t, x_0, \delta)}, \qquad (13)$$

So

$$HR(t, x_1, x_0) = \frac{h_0(t) r(x_1, \delta)}{h_0(t) r(x_0, \delta)}$$

$$= \frac{r(x_1, \delta)}{r(x_0, \delta)} \qquad (14)$$

The function $r(x, \delta)$ is the only function the hazard ratio (HR) depends on. Cox (1972) suggested that using $r(x, \delta) = exp(x\delta)$, with this parameterization as a founder of the model in (12), the hazard function is

$$h(t, x, \delta) = h_0(t) e^{x\delta} \qquad (15)$$

and the hazard ratio is

$$HR(t, x_1, x_0) = e^{\delta(x_1 - x_0)} \qquad (16)$$

Many researchers in the literature will refer to this model by different names. Some literature will term it the Cox model whereas the others will term it the cox proportional hazards model or simply the proportional hazards model.

For example, in trial that compared the rate of deaths among the two gender groups where gender is a covariate and is dichotomous, with a value of $x_1 = 1\ for\ males\ and\ x_0 = 0\ for\ females, the\ hazard\ ratio\ in\ (10) becomes$

$$HR(t, x_1, x_0) = e^{\delta}, \qquad (17)$$

If the value of the coefficient is $\delta = In(3)$, then the interpretation is that males are dying at three times the rate of females. The survivorship function in Cox model is given by the following equation

$$S(t, x, \delta) = [S_0(t)]^{exp(x\delta)} \qquad (18)$$

The hazard function described in this section are called semi-parametric functions since they do not explicitly describe the baseline hazard function, $h_0(t)$.

2.9.2 ***Extended Cox Model***

The Cox proportional hazard model whose general form is illustrated in the equation (19) below, presents a function for the hazard at the time *t* for an individual with a given specification of a set explanatory variables denoted by X. The **X** represents a vector of explanatory variables that are modelled to predict an individual's hazard.

$$h(t|\boldsymbol{X_i}) = h_0(t)\exp[\textstyle\sum_{i=1}^{k} \beta_i \boldsymbol{X}_i] \qquad (19)$$

or as

$$h(t|\boldsymbol{X_i}) = h_0(t)\exp(\boldsymbol{X}_i\beta) \qquad (19)$$

where, $X = (X_1, \ldots., X_n)'$ is a $(k \text{ x } 1)$ vector of predictor or explanatory variables and $\beta = (\beta_1, \ldots, \beta_k)'$ is the a $(k \text{ x } 1)$ vector of parameters.

The Cox model formula above states that the hazard at time *t* is a product of two functions, the baseline hazard function $h_0(t)$ and the exponential function e to the linear sum of $\beta_i X_i$, where the sum is over the n explanatory **X** variables. The most important feature of the formula above entails the proportional hazards (PH) assumptions underlined in the model application. This feature is that the baseline hazard function is a function of t, it does not involve **X**'s, whereas the Exponential function involves **X'**s but does not involve t. In this particular case the **X'**s are said to be time-independent.

In some cases, however, it is possible to have **X**'s that do involve t, and such **X**'s are referred to as time-dependent **X**'s. If these variables are considered, the Cox Ph model form can still apply even though such a model does not satisfy the Proportional hazards assumption. When this happens, it now called the extended Cox model, and its general form is as follows:

$$h\big(t, X(t)\big) = h_0(t)\exp[\textstyle\sum_{i=1}^{n_1}\beta_i X_i + \sum_{j=1}^{n_2}\delta_j X_j(t)] \qquad (20)$$

In simple terms, the extended Cox model is a Cox regression model which has the addition of a dependent time variable on variables that do not meet the proportional hazard assumption. From the equation above, the extended Cox PH model also includes the baseline hazard function $h_0(t)$ which is multiplied by an exponential function just like the Cox proportional hazard model but unlike in the Cox proportional hazard model, the extended Cox model includes both time-independent predictors denoted by X_i variables and the time-dependent predictors denoted by$X_j(t)$ variables in the exponential part. (Therneau & Grambsch, 2000). In this case $X(t)$ denotes the entire vector of predictors at time t.

The regression coefficients in the extended Cox model are estimated using a maximum likelihood procedure just as it is the case with the Cox proportional hazard model. However, the calculations for the latter are more complicated than those of the Cox proportional hazard model since the risk set used to form the likelihood function become more complicated as they include time-dependent variables in them (Hosmer & Lemeshow, 1999). Whereas the statistical inferences methods are essentially the same as for both models where Wald test or likelihood ratio (*LR*) test can be used and also the large sample confidence interval methods.

The effect which a time-dependent variable $X_j(t)$ has on survival probability at time *t* depends on the value of the said variable at the same time *t* and not on the value at an earlier or later time. This is an important assumption of the extended Cox proportional hazard model and that is, the hazard at time *t* depends on the value of $X_j(t)$ at the same time (Hosmer & Lemeshow, 1999). The values of the variable $X_j(t)$ may change over time but the hazard model provides only one coefficient for each time dependent variable in the model. This statement simply says that, at a

given time *t*, there is only one value of the variable $X_j(t)$ that influences the hazard, that value being measured at time *t*.

The formula for the hazard ratio as derived from the extended Cox model is as shown below:

$$\widehat{HR}(t) = \frac{\hat{h}(t,X^*(t))}{\hat{h}(t,X(t))}, \qquad (20)$$

$$= \exp[\sum_{i=1}^{n_1} \widehat{\beta}_i(X_i^* - X_i) + \sum_{j=1}^{n_2} \delta_j(X_i^*(t) - X_i(t))\,] \qquad (21)$$

The assumption of the proportional hazards is not satisfied when we apply the extended Cox model. The general hazard ratio formula for the extended Cox model is as shown in equation 21 and describes the ratio of hazards at a particular time *t*, and requires the specification of two sets of predictors at time *t*. These two sets are denoted as $X^*(t)$and $\boldsymbol{X}(t)$. "The two vectors of predictors, $X^*(t)$and $X(t)$, identify two categories at time t for the combined vector of predictors containing both time-independent and time-dependent variables" (Ingabire, Mwalili, & Orwa, 2015).

2.10 Parametric Survival Model (AFT)

Literature shows that different types of parametric models have been used, suggested, or proposed for use with survival data. James (1988) cite that the hazard functions shapes associated with most of these models can be classified as; constant, monotonically increasing, monotonically decreasing, U-shaped, and bell-shaped. Unknown parameters values are the ones that makes a hazard function to take on a variety of shapes. Gross and Clark (1975) cite individuals whose only risks of death are accidents or rare illness as an example of a population where a constant hazard is applicable.

To run away from specifying the hazard function completely, we employed nonparametric or semiparametric models for the analysis of censored survival time data in so doing reducing the set of assumptions required to supply the hazard ratios formed from the coefficients whose

clinical meaning can easily be explained. Aside from the nonparametric or semiparametric models, we have parametric methods available to use in situations where the distribution of survival time is known from prior research. (Hosmer & Lemeshow, 1999) states that these models have advantages in terms of; using a full Maximum likelihood to provide estimates to the parameters; the Coefficients can be clinically meaningful and for some models are related to those from a proportional hazards model; survival time estimates can be determined by the fitted values; and Residuals can be computed that are different between observed and predicted values of time. Analysis results of using a fully parametric model can have bring a normal error as that of a linear analysis (Hosmer & Lemeshow, 1999).

Wei (1992) suggests that the parameters within the AFT models will be easily understood by clinical investigators as they are interpreted as effects on the duration than the hazard ratios. Fisher (1992) on the other hand commented that most research-oriented clinicians have little or no trouble understanding the proportional hazards model or the hazard ratio. Investigators in research have employed the proportional hazards model as seen from the literature. This is so as it has become the standard method for multivariate analysis for survival times in many applied settings. These models also can provide concise and simply interpreted analysis as discussed on the benefits earlier as some even have proportional hazards and may be ready to provide an alternate way of explaining covariate effects albeit the PH models interpretation would be first choice.

When survival time is assumed to follow a known distribution, it is referred to as a parametric survival model. Some of distributions that are used for survival time in research are: the lognormal, the log-logistic, the Weibull, the exponential (a special case of the Weibull), and the generalized gamma. The Cox proportional hazards model is a semiparametric model as listed in the sections above because even when we have known regression parameters, the distribution of the result/outcome are unknown. That is why the baseline hazard function is not laid out in a Cox model.

2.10.1 The Weibull Model

The Weibull distribution is one of the continuous distributions in probability theory and statistics. From literature, the Weibull model is found to be the most widely used parametric survival model. Weibull distribution can be employed in data if survival hazard function increases or decreases monotonically with an increase in survival time. Cox proportional hazard model uses Weibull distribution whose hazard function, for t>0, follows:

$$h(t) = \theta z t^{z-1}, \qquad (22)$$

where z and $\theta > 0$.

The survival function is given by

$$S(t) = exp(-\theta t^{z}), \qquad (23)$$

The density function,

$$f(t) = \theta z t^{z-1} exp(-\theta t^{z}), \qquad (24)$$

The Weibull distribution is a generalization of the exponential distribution and is characterized by θ which is the scale parameter and will be reparametrized with regression coefficients as it is with the exponential model. The additional parameter z is called a shape parameter. The shape of the hazard function shape is determined by z, the shape parameter. For example, If the shape parameter z > 1 then the hazard increases as time increases. The hazard is constant if the shape parameter z = 1 and the Weibull model is reduced to the exponential model (exponential distribution is generated) . If the shape parameter z < 1 then the hazard function decreases over time. The Weibull model is given greater flexibility by p than the exponential model yet the hazard function remains relatively simple.

The shape parameter z and scale parameter θ makes it impossible to have two-dimensional sufficient statistics hence there has been much work done with the Weibull distribution (Lawless, 1982). The Weibull model possess the unique property in that if the AFT assumption does not hold then the PH assumption does not hold also (Cox and Oakes, 1984).

In absence of a close form solution, numerical optimization are used in order to get the maximum likelihood estimates. A lot of work with the Weibull distribution has been done since there are no two-dimensional sufficient statistics for θ and z (Lawless, 1982).

2.10.2 Log-Logistic Model

Unlike the Weibull distribution described above, the hazard function for the log-logistic distribution allows for a couple of nonmonotonic behavior within the hazard function. The log-logistic hazard has the following form;

$$h(t) = \frac{\theta z t^{z-1}}{1+ \theta t^{z}} \qquad (25)$$

, (where $z > 0$ and $\theta > 0$)

The log-logistic distribution can work with an AFT model but not a PH model. The shape parameter is $(z > 0)$. If $z \leq 1$ then the hazard decreases over time. If $z > 1$, however, the hazard increases to a maximum point then decreases over time. During this case $(z > 1)$, the hazard function is claimed to be unimodal**.**

The survival odds are the odds of surviving beyond time t $(i.e., S(t)/(1 - S(t))$.

$$\frac{S(t)}{(1 - S(t))} = \frac{Z(T > t)}{Z(T \leq t)} \qquad (26)$$

This is the probability of the event not happening in time t divided by the probability of getting the event by time t. The failure odds are the odds of getting the event by time t $(i.e., (1 - S(t)/ S(t))$, which is the reciprocal of the survival odds

$$\frac{(1 - S(t))}{S(t)} = \frac{Z(T \leq t)}{Z(T > t)} \qquad (27)$$

The log-logistic survival function $(S(t))$

$$S(t) = \frac{1}{1 + \theta t^{z}} \qquad (28)$$

and failure function $(1 - S(t))$

$$1 - S(t) = \frac{\lambda t^z}{1 + \theta t^z} \qquad (29)$$

in a log-logistic model, the failure odds are simplified to θt^z

$$\frac{1 - S(t)}{S(t)} = \frac{\frac{\theta t^z}{1 + \theta t^z}}{\frac{1}{1 + \theta t^z}} = \theta t^z \qquad (30)$$

2.10.3 The Exponential Regression Model

The exponential distribution is different with the other distributions like the Weibull and log logistic distributions which have two parameters θ and *z*, the exponential distribution is a one-parameter distribution with a constant hazard θ. During the overview with the Weibull distribution in section (2.10.1) above, we saw that the exponential distribution was generated if the shape parameter z = 1. The product of h(t) and S(t) will give use the probability density function for this distribution. The exponential model, which is the simplest parametric survival model in that the hazard is constant over time (i.e., h(t) = θ).

For the Exponential, S(t)

$$exp(-\theta t)$$

reparameterization of S(t) as an AFT model is shown below.

$$S(t) = exp(-\theta t)$$

$$t = [- ln \left(S(t)\right) x \frac{1}{\theta}$$

$$let \frac{1}{\theta} = exp\left(\delta_0 + \delta_1 x\right), equivalent\ to\ \theta = exp[-(\delta_0 + \delta_1 x)]$$

By reparameterization, we get

$$t = [- ln \left(S(t)\right)]\ exp(\delta_0 + \delta_1 x) \qquad (31)$$

median survival time is found by substituting $S(t) = 0.5$ in above equation

and the h(t)

θ

, the exponential hazard is constant and can be parametrized as a PH model,

$$h(t) = \theta = exp-(\delta_0 + \delta_1 x)$$

The hazard ratio for a dichotomous covariate is

$$HR(x = 1, x = 0) = ex\,p-(\delta_1) \qquad (32)$$

James (1988) cites that the exponential distribution manifest clearly the unique memoryless property since the hazard function is independent of t. James (1988) further cites that it is for this property and ease of estimation that provides justification for its use in survival studies and biomedical application. In summary, defined for t>0, the hazard function is $\theta(t) = \theta$, survival function, S(t) = $\exp(-\theta t)$, and the density function, f(f) = θ $\exp(-\theta t)$.

The exponential model is a proportional hazards model making it an accelerated failure time (AFT) model. Cox and Oakes (1984) show that the only AFT models that have proportional hazards are exponential and Weibull regression models. If an exponential regression model fits the data, one may express the effect of covariates as a time ratio or a hazard ratio. The assumption that the hazard is constant for each pattern of covariates outweighs the PH assumption. If the hazards are constant, then of course the ratio of the hazards is constant. However, the hazard ratio being constant does not necessarily mean that each hazard is constant. In a Cox PH model, the baseline hazard is assumed to be a variable. In fact, no studies have specified, the form of the baseline hazard is not even specified. To estimate regression coefficients, maximum likelihood estimators (MLE) are used and they tend to be normally distributed. (Kleinbaum & Mitchel, 1996)

The key assumption for survival models has been the proportional hazard assumption. However, parametric survival models should not always be PH models. Many parametric models are acceleration failure time models and not PH models. The exponential and Weibull distributions can use both the PH and AFT assumptions (Kleinbaum & Mitchel, 1996). The interpretation of parameters is different for AFT and PH models. The AFT assumption is valid for a comparison

of survival times whereas the PH assumption is applicable for a comparison of hazards. The base assumption for AFT models is that the effect of covariates is multiplicative (proportional) with respect to survival time, whereas for PH models the underlying assumption is that the effect of covariates is multiplicative with reference to the hazard (Kleinbaum & Mitchel, 1996).

2.10.4 Generalized gamma model

The generalized gamma model is one of the parametric survival models which is also a generalization of the exponential distribution. Both the hazard and the survival function for this model are complicated and can only be expressed in terms of integrals. (Kleinbaum & Mitchel, 1996). The generalized gamma distribution has only three parameters allowing for flexibility in its shape. The Weibull and lognormal distributions are special cases of the generalized gamma distribution (Kleinbaum & Mitchel, 1996).

2.10.5 Lognormal model

Kalbfleisch and Prentice (1980) cite that the lognormal model is easy to use if there is no censoring occurring since the likelihood function does not involve the incomplete normal integral. When censoring occurs, it is almost impossible to do the computations. When this happens, the log-logistic distribution provides a good approximation to the log-normal model.

The hazard function for this distribution is given by

$$h(t) = \frac{\frac{1}{t\delta\sqrt{2\alpha}} exp[-\left(\frac{\log t - \mu}{2\delta^2}\right)^2]}{1 - \dot{\phi}\left(\frac{log\ t - \mu}{\delta}\right)}\ , \qquad (33)$$

where $\dot{\phi}$ is the cumulative distribution function of standard normal variable and $\delta > 0$.

The survival function is given by

$$S(t) = 1 - \dot{\phi}\left(\frac{log\ t - \mu}{\delta}\right), \qquad (34)$$

and the density function by

$$f(t) = \frac{1}{t\delta\sqrt{2\alpha}} exp[-\left(\frac{\log t - \mu}{2\delta^2}\right)^2], \qquad (35)$$

The lognormal model has a relatively complicated hazard and survival function that can only be expressed as integrals (Kleinbaum & Mitchel, 1996). The shape of the lognormal distribution is very similar to the log-logistic distribution and yields similar model results. The difference is that although the lognormal model accommodates an accelerated failure time model, it is not a proportional odds model.

2.10.6 The Gompertz

The Gompertz distributions are widely used in actuarial work in literature. The following hazard function describes the distribution well:

$\theta(t) = exp\,(z + yt)$, (36)

The survival function is given by

$$S(t) = \exp\left[-\frac{e^z}{y}\,(exp\,(yt) - 1)\right], \quad (37)$$

And the density function,

$$f(t) = \exp\left[(z + yt) - \frac{1}{y}\,(exp\,(z + yt) - e^z)\right], \quad (38)$$

when z=0, the Gompertz distribution reduces to the exponential distribution. Parametric models need not be AFT models. The Gompertz model is a parametric proportional hazards model but not an AFT model. The model can be expressed in a form like that of a Cox PH model except that the baseline hazard is specified as the hazard of a Gompertz distribution containing a shape parameter γ.

2.11 Model parameter estimation

Maximum likelihood estimation is used to estimate the unknown parameters of the parametric distributions. Kalbfleisch & Prentice,(1973) derived a likelihood involving on $\beta\ and\ Z\ (\ not\ \lambda_0(t))$ based on the marginal distribution of the ranks of the observed failure times (in absence of censoring). Cox (1972) derived the same likelihood and generalized it for censoring using the idea of a partial likelihood. If Y_n is uncensored, the *n*th subject contributes $f\ (Y_n)$to the likelihood. If Y_n is censored, the *n*th subject contributes $Pr(y > Y_n)$to the likelihood. The joint likelihood for all p subjects is

$$L = \prod_{i:\beta_i=1}^{p} f(Y_i) \prod_{i:\beta_i=0}^{p} S(Y_i) , \qquad (39)$$

The log-likelihood can be written as

$$\log \mathrm{L} = \sum_{\mathrm{i}:\,\beta_{\mathrm{i}=1}}^{p} \log\big(\mathrm{h}(\mathrm{Y_i})\big) - \sum_{\mathrm{i}=1}^{p} \mathrm{H}\,(\mathrm{Y_i}) \qquad (40)$$

2.12 Model Comparison

2.12.1 Evaluation Criteria

Several survival models with semi-parametric or parametric approaches are used in different fields including medical, natural, and social sciences. The choice of the foremost appropriate model for data at hand is as important as the analysis itself. In literature available, it has been noted that the likelihood-based model selection for example Akaike information criteria (AIC) or Bayesian information criteria (BIC) are used to select among nested models. The AIC has been used as a measure of goodness of fit that balances model fit against simplicity. (Akaike, 1981) while BIC has been used as a similar measure (Simonoff, 2003) .

2.12.2 Akaike's Information Criterion

Akaike's information criterion named after Hirotugu Akaike, (1927–2009) provides an estimator for predicting error and the relative quality of Statistical models. The AIC compares related models and helps in selecting a model that fits the available data with less margin of error (Glen, 2015). The formula for AIC is given below:

$$AIC = -2l\big(\hat{\vartheta}_{ML}\big) + 2p\,, \qquad (41)$$

penalizes the maximized log-likelihood with the number of parameters p. The criterion is negatively oriented, i.e., the model with minimal AIC is selected. Therefore, a difference of $2q$ is sufficient for a model with q additional parameters to be preferred (Akaike, 1981).

2.12.3 Bayesian Information Criterion

As an alternative to the AIC, we sometimes use the *Bayesian information criterion with* the formula below:

$$BIC = -2l\big(\hat{\vartheta}_{ML}\big) + p\ log\ (n), \qquad (42)$$

where n denotes the size of the sample. Half of the negative BIC is additionally referred to as the *Schwarz criterion*. It has an equivalent orientation as AIC, such models with smaller BIC are preferred. It penalizes model complexity in general (i.e., if $\log(n) \geq 2 \Leftrightarrow n \geq 8$) more distinctly than AIC.

The AIC is a way of putting off the complexity of an estimated model against how well the model fits the data. For this study models discussed, the AIC was given by $AIC = -2^{*} log\ (likelihood) + 2(p + k),$ (43)

Where p is the number of parameters, k=1 for the exponential model, k=2 for the Weibull, log-logistic, and log normal models (Klein et al., 1997). Lower AIC indicates better likelihood.
Other studies in literature use just the Akaike information criterion (AIC) and residues review to compare the performance of the parametric models.

2.13 Model Diagnosis

This section presents different approaches to assess the assumptions under different models. These include the use of time varying covariates, Cox Snell for goodness of fit test and graphical approach using Schoenfeld residuals. Three approaches are also used for evaluating the proportional hazards (PH) assumption of the Cox model—a graphical procedure, a goodness-of-fit testing procedure, and a procedure that involves the use of time-dependent variables.

2.13.1 Cox Snell Residuals

The use of the Cox-Snell residuals is goodness of fit of the Cox PH model. As defined by Collet (2003), Cox-Snell residuals are given as

$rC_i = exp\ \left(\hat{\theta}' x_i\right) \hat{H}_0(t_i)$ (44)

When assessing the model, the plot of the integrated hazard based on the residuals against the hazard rate estimates backed out of the Cox model should have a 45-degree slope. Therefore, if the Cox model fits, then the residuals should be distributed as unit exponential.

2.13.2 Schoenfeld Residuals

The idea behind the statistical test is that if the PH assumption holds for a specific covariate, then the Schoenfeld residuals for that covariate won't be associated with survival time. Schoenfeld (1982) proposed the primary set of residuals to be used with a fitted proportional hazards model and packages providing them as the "Schoenfeld residuals." These are based on the individual contributions to the derivative of the log partial likelihood. Collet (2003) denotes the ith Schoenfeld residual for Xj, jth explanatory variable in the model as given by;

$$r_{pji} = \beta_i\{x_{ji} - \hat{\theta}_{ji}\}, \qquad (45)$$

Where xji is the value of the jth explanatory variable, j=1,2,3, …, p, for each individual in the study. The schoenfeld residuals are particularly useful in evaluating the PH assumption after fitting a Cox regression model.

2.13.3 Martingale's Residuals

Hosmer and Lemeshow (1999) define the martingale residuals as; $\widehat{M}_n = C_n - \widehat{H}_n$

Where the components of the residual for the *nth* subject are the values of the censoring variable C_n and the estimated cumulative hazard $\widehat{H}_n = \widehat{H}(t_n, x_n, \hat{\delta})$.

2.13.4 Time-dependent covariates

The proportional Hazard assumption can also be assessed using the time-dependent explanatory variables. When this is done the Cox model is extended to contain a product that shows the interaction involving the time independent variable being assessed and some function of time. (Hosmer & Lemeshow, 1999). This as explained earlier in this context leads to the extended Cox model;

$$h(t, X) = h_0(t)\, exp[\delta X + \sigma X * g(t)] \qquad (46)$$

When assessing predictors one-at-a-time, the extended Cox model takes the general form shown above for the predictor *X*. (Hosmer & Lemeshow, 1999)The test can be carried out using either a likelihood ratio statistic or a Wald statistic. In either case, the test statistic has a chi-square distribution with one degree of freedom under the null hypothesis. The extended Cox model in some cases can also be used to assess the PH assumption for several predictors simultaneously as well as for a given predictor adjusted for other predictors in the model (Kleinbaum & Mitchel, 1996).

$$h(t, X) = h_0(t)\, exp\left(\sum_{n=1}^{p} \delta_n X_n + \sigma_n X_n * g_n(t)\right) \quad (47)$$

The primary limitation to using the extended cox model for assessing the PH assumption involves selection of functions since different functions will result in different conclusions on whether PH Assumption is Satisfied (Kleinbaum & Mitchel, 1996).

2.13.5 The log(−log) of S(t)

The proportional Hazard assumption can also be assessed using the time-dependent explanatory variables. When this is done the Cox model is extended to contain a product that shows the interaction involving the time independent variable being assessed and some function of time. (Hosmer & Lemeshow, 1999). This as explained earlier in this context leads to the extended Cox model;

$$h(t, X) = h_0(t)\, exp[\delta X + \sigma X * g(t)] \quad (48)$$

When assessing predictors one-at-a-time, the extended Cox model takes the general form shown above for the predictor *X*. (Hosmer & Lemeshow, 1999). The test can be carried out using either a likelihood ratio statistic or a Wald statistic. In either case, the test statistic has a chi-square distribution with one degree of freedom under the null hypothesis. The extended Cox model in some cases can also be used to assess the PH assumption for several predictors simultaneously as well as for a given predictor adjusted for other predictors in the model (Kleinbaum & Mitchel, 1996).

$$h(t, X) = h_0(t)\, exp\left(\sum_{n=1}^{p} \delta_n X_n + \sigma_n X_n * g_n(t)\right) \quad (49)$$

The primary limitation to using the extended cox model for assessing the PH assumption involves selection of functions since different functions will result in different conclusions on whether PH Assumption is Satisfied. (Kleinbaum & Mitchel, 1996).

2.14 Review of Previous Research

Different survival studies, where the hazards are not proportional, for example in cases where time-varying covariates are present; researchers have opted not to use a traditional cox model to model the survival as it would not be ideal. To deal with this situation, a study by Zhang Z et al. (2018) on Time-varying covariates and coefficients in Cox regression models, recommend that

when time varying covariates or coefficients are present, an analyst should consider taking them into account in survival modelling in order to improve the estimation.

On the same, another study by Dekker, et al. (2016) looked at time-dependent effects and time-varying risk factors. The first approach studied the effect of a fixed baseline risk factor on Mortality in different time windows (time stratified effects). It resulted in separate HRs for distinct time windows. In the second approach, a risk factor that changed over time was studied in relation to subsequent mortality. This approach resulted in one HR that could be considered as a weighted average of short-term effects on mortality. They noted that the dependence of the hazard function for an individual on the values of certain explanatory values can be modelled. When explanatory variables are incorporated in a model for survival data, the values taken by such variables are those recorded at the origin of the study. The impact of these variables on the hazard of death is then evaluated (Dekker, Mutsert, van Dijk, Zoccali, & Jager, 2016).

In studies that generate survival data, subjects are monitored for the entire duration of the study and values for different variables are captured and recorded on a regular basis during this period as frequent enough as defined by the study. (Machin, Cheung, & Parmar, 2006). Where an account is taken of the values of the explanatory variables as they evolve through the study, the resulting model for the hazard of event at any given time would be more satisfactory, this is because more recent values of variables provide a better reflection of the situation than the values at time origin of the study.

It is evident that Cox Proportional hazard model has been widely used in survival analysis; however, all the studies reviewed above identified several weaknesses like overestimation which make them less suitable than extended Cox regression adopted in this thesis. Cox proportional hazard model ignores the effect of time dependent variables which makes it weak compared to extended Cox as it did not take into account the effect of survival times; thus, the main aim of this thesis is to model childhood mortality reduction after mass azithromycin by comparing Cox proportional hazard model and extended Cox model and determine which model best explains the estimates.

CHAPTER 3
METHODOLOGY

This chapter presents the methodology used for this project. This chapter will describe the research design and the procedures. In particular, study design; data collection and data analysis; analysis approach and lastly ethical consideration.

3.1 Study design

The study used secondary data from the MORDOR (Macrolides Oraux pour Réduire les Décès avec un Oeil sur la Résistance), a cluster-randomized trial. MORDOR was a community-randomized trial conducted in the Malawian district of Mangochi, the Nigerien districts of Boboye and Loga, and the Tanzanian district of Kilosa [10]. The 1533 randomization units were the health surveillance assistant area in Malawi, the *grappe* in Niger, and the hamlet in Tanzania. Communities with a population between 200 and 2000 inhabitants on the most recent census were eligible for enrollment. Enrollment was based on census information available prior to the study. Communities remained in the study even if the population size drifted out of this numerical range. Children aged 1–59 months who weighed at least 3800 g were eligible for azithromycin or placebo. Biannual distributions were performed over each district in a rolling fashion, over a 6-month prespecified time period (8 months for the initial census and distribution) [10]. Thus, treatments could be given any time during the year. The estimated time of death was collected during the subsequent census.

3.2 The MORDOR Data

The MORDOR (Macrolides Oraux pour Réduire les Décès avec un Oeil sur la Résistance) is a cluster-randomized multi-country trial which was double blinded and collected real time data using the customized android based WUHA-MORDOR mobile application using salesforce server database. The application automatically assigned unique IDs to households and individuals once enrolled and recorded in the mobile application. All household census data and

individual details were recorded. Upon data synchronization, all information could be retrieved from the system in the future follow-up visits to that household by searching the household name.

The MORDOR data was availed for this thesis in stata format. It covered client's information from enrollment until death or end of study period (2014 to 2017). The dataset used in this study contained information on infants only. An infant here was defined as any individual below the age of 5.

3.3 Data collection and data management

The study data were extracted from the MORDOR Malawi database. The data was extracted in stata format for the whole study period (2014-2017). The study utilized information collected for only infants aged 1 – 12 months old from both the intervention and control arm of the study from Mangochi and Namwera zones of Mangochi district based on the researchers interests and convenience. Social – demographic characteristics and clinical information were captured from all subjects enrolled in the study. The socio-demographic characteristics included were age and sex. The clinical data included date of death, dosage taken and treatment drug letter. All residents from consented households who were under 5 and were eligible were enrolled in the study and were treated with oral azithromycin or placebo during rounds of MDA. All this information (births, deaths, weight and location-GPS) were being entered in the MORDOR database using MORDOR application on Nexus 7 tablets during census after every 6 months.

3.4 Sample size and sampling procedure

The data for this study was made available through London School of Hygiene and Tropical Medicine which was one of the implementing partners in Malawi. The data had information on all subjects who participated in the trial form both arms of the study. A representative sample was made available for this study which included all children between 1 month to 12 months as it was the target population for this study. The study analyzed information from two zones in Mangochi which included Namwera and Mangochi zones. A total of 36,349 infants aged 1 to 12 months were included in the analysis for this study.

3.4.1 Inclusion and exclusion criterion

The study looked at participants who were aged 1 to 12 months who were enrolled in the study. The entry point was at least 1 month old and weigh 3800g and should be from Mangochi and Namwera zones of Mangochi district. All those from other zones, aged more than a month and weighed less than 3800g were excluded in this study.

3.5 Study Outcome

The outcome variable is the occurrence of death in an individual (a binary indicator variable for each individual aged 1 to 12 months who were present at baseline).

3.6 Data handling and description

This study data was extracted from the MORDOR Malawi database which was collected upon approval from the College of Medicine Research Ethics Committee (COMREC) where the primary study protocol was reviewed and granted ethical approval before any patient-related research activities began in Malawi. The London School of Hygiene and Tropical Medicine Interventions Ethics Committee also reviewed the study protocol and granted ethical approval (reference # 6500). The MORDOR (Mortality Reduction after Oral Azithromycin) framework covering trials in three countries had been approved by the Committee on Human Research (Parnassus Panel) of the University of California, San Francisco IRB # 10-01036, reference 001294.

Data was recorded electronically using handheld Google Nexus 7 mobile devices with custom-made software applications and uploaded onto a secure, password protected, central server. Rapid transfer of electronically captured data allowed nearly real time monitoring of activity at the study site. Data generated in Malawi was reviewed in real time by a data coordinating centre based in Blantyre. All handheld devices and data entry coordinating centres were password protected, and all changes in data were noted, including the date of the change, and the person who made the change. Training sessions were conducted before each biannual census. The central database application used hard disk encryption and physical protection of the server (which was maintained in a locked room accessible only to authorized personnel). The database was based on mySQL (which supports standard SQL queries). Data was backed up off site

(providing integrity in case of the physical loss of the server). Data was never deleted from mobile capture devices until at least one offsite backup had been completed. Data security during electronic transfer was achieved through use of the Advanced Encryption Standard (AES). Data was validated by the Malawi data coordinating centre data manager/study coordinator and de-identified before uploading to the central database.

The data was explored to obtain important variables that would be used for analysis. The data was first cleaned, and then sorted for easy navigation when doing analysis. For the purpose of this thesis, the randomized subject who consented to be enrolled in the study was the unit of analysis and the outcome variable was time to death. Categorical variables were coded using numbers e.g. female =1, male=2; azithromycin = 1, Placebo=0; died=1, alive=0. Age was a continuous variable. Survival time was measured in months.

3.7 Data analysis

3.7.1 The Estimates, Statistical Tests and the Level of Significance

The analysis first looked at some descriptive statistics (frequencies, Inter-quartile range, mean and median) for the baseline characteristics. The baseline characteristics of the participants in the study such as age and weight were presented as median and, naturally, the measure of dispersion was the interquartile range. Rank-based measure of central tendency and its subsequent measure of dispersion are ideal in survival data since survival data are typically right skewed. The Hazard ratios, their corresponding coefficients and 95% confidence intervals were presented for Cox Proportional hazards and extended Cox proportional hazards models. Also included were the calculated p-values for all statistics. All statistical tests were made at 5% level of significance.

Kaplan Meier curves for the 2 groups were compared. The cumulative incidence curve and Kaplan Meier curves were compared. CI curves for categorical variables; gender, HIV status and ART status were obtained and comparison between the different groups for the patients in terms of survival was performed.

Statistical analysis were done using stata version 14, a statistical software package created in 1985 by StataCorp used in data management, statistical analysis, graphics and simulations.

The Schoenfeld residuals and plots were used to test the PH assumption. The Martingale residuals were used to check the Linearity of variable age. Time-varying covariates were used to test for PH assumption for the extended Cox model.

3. 8 Model Specification

3.8.1 Cox Proportional Hazards model

We will fit Cox Proportion hazard model as it is the most common approach to model covariate effects on survival. It takes into account the effect of censored observations (Cox., 1972). The only assumption made is on the proportionality of the baseline hazard. The proportional hazard assumption means that the hazard ratio is constant over time or that the hazard for an individual is proportional to the hazard for any other individual (Therneau. & Grambsch., 2000). Let x1, … ,xn be the values of p covariates X1, …. , Xn, according to the Cox regression model, the hazard function is given as follows;

$$h(t) = h_0(t)\exp(\sum_{i=1}^{n} \delta_i x_i), \qquad (50)$$

Where $\delta_i = (\delta_1, \delta_2, \delta_p)$ is a 1 × n vector of regression coefficients and h0(t) is the baseline hazard function at time t.

The Cox proportional Hazard model is used to estimate the effects of covariates on child survival in this thesis. The model has the capacity to identify covariates with little effect on survival and risk of death of an individual based on prognostic covariate.

3.8.2 Extended Cox model

The extended Cox model was used to take into account time varying covariates to control for survival overestimation. The time-dependent covariate and a covariate that has no time-dependency will be modelled together using this model. The time-dependent covariate will be interacted with time function because it does not meet the proportional hazard assumption.

The Cox proportional hazard model whose general form is illustrated in the equation below, presents a function for the hazard at the time *t* for an individual with a given specification of a set explanatory variables denoted by X. The **X** represents a vector of explanatory variables that are modelled to predict an individual's hazard.

$$h(t, \boldsymbol{X}) = h_0(t) \exp(\sum_{i=1}^{n} \beta_i \boldsymbol{X}_i), \qquad (51)$$

where, $X = X_1, X_2 \ldots \ldots \ldots X_n$ Explanatory/predictor variables.

The Cox model formula above states that the hazard at time *t* is a product of two functions, the baseline hazard function $h_0(t)$ and the exponential function e to the linear sum of $\beta_i X_i$, where the sum is over the n explanatory X variables. The most important feature of the formula above entails the proportional hazards (PH) assumptions underlined in the model application. This feature is that the baseline hazard function is a function of t, it does not involve X's, whereas the Exponential function involves X's but does not involve t. In this particular case the X's are said to be time-independent.

In some cases, however, it is possible to have X's that do involve t, and such X's are referred to as time-dependent X's. If these variables are considered, the Cox Ph model form can still apply even though such a model does not satisfy the Proportional hazards assumption. When this happens, it now called the extended Cox model and its general form is as follows:

$$h\big(t, X(t)\big) = h_0(t) \exp\big[\sum_{i=1}^{n_1} \beta_i X_i + \sum_{b=1}^{n_2} \delta_b X_b(t)\big], \qquad (52)$$

In simple terms, the extended Cox model is a Cox regression model which has the addition of a dependent time variable on variables that do not meet the proportional hazard assumption. From the equation above, the extended Cox PH model also includes the baseline hazard function $h_0(t)$ which is multiplied by an exponential function just like the Cox proportional hazard model but unlike in the Cox proportional hazard model, the extended Cox model includes both time-independent predictors denoted by X_i variables and the time-dependent predictors denoted by$X_b(t)$ variables in the exponential part. (Therneau & Grambsch, 2000). In this case $X(t)$ denotes the entire vector of predictors at time t.

The regression coefficients in the extended Cox model are estimated using a maximum likelihood procedure just as it is the case with the Cox proportional hazard model. However, the calculations for the latter are more complicated than those of the Cox proportional hazard model

since the risk set used to form the likelihood function become more complicated as they include time-dependent variables in them. (Hosmer & Lemeshow, 1999). Whereas the statistical inferences methods are essentially the same as for both models where Wald test or likelihood ratio (*LR*) test can be used and also the large sample confidence interval methods.

The effect which a time-dependent variable $X_b(t)$ has on survival probability at time *t* depends on the value of the said variable at the same time *t* and not on the value at an earlier or later time. This is an important assumption of the extended Cox proportional hazard model and that is to say, the hazard at time *t* depends on the value of $X_b(t)$ at the same time. (Hosmer & Lemeshow, 1999). The values of the variable $X_b(t)$ may change over time but the hazard model provides only one coefficient for each time dependent variable in the model. This statement simply says that, at a given time *t*, there is only one value of the variable $X_b(t)$ that influences the hazard, that value being measured at time *t*.

The formula for the hazard ratio as derived from the extended Cox model is as shown below:

$$\widehat{HR}(t) = \frac{\hat{h}(t, X^*(t))}{\hat{h}\big(t, X(t)\big)}, \qquad (53)$$

$$= \exp[\textstyle\sum_{i=1}^{n_1} \widehat{\beta}_i (X_i^* - X_i) + \sum_{b=1}^{n_2} \delta_b \big(X_i^*(t) - X_i(t)\big)\,], \qquad (54)$$

The assumption of the proportional hazards is not satisfied when we apply the extended Cox model. The general hazard ratio formula for the extended Cox model shown above describes the ratio of hazards at a particular time *t*, and requires the specification of two sets of predictors at time *t*. These two sets are denoted as $X^*(t)$and $\boldsymbol{X}(t)$. The two vectors of predictors, $X^*(t)$and $X(t)$, identify two categories at time *t* for the combined vector of predictors containing both time-independent and time-dependent variables (Ingabire, Mwalili, & Orwa, 2015).

3.8.3 Model assumption assessment and Goodness-Of-Fit

The proportional hazard assumption for the Cox model was performed. Cox Snell residual test was performed to test goodness of fit. The linearity assumption was checked by plotting a Martingale residual plot to check linearity for covariate age. Time-varying covariates were used when modeling the hazards to test for proportionality assumption for Cox PH and Extended Cox models. Whereas, the Schoenfeld's global test was used to test the proportional hazards assumption.

3.9 Ethical consideration

Full ethical approval was granted by College of Medicine Research Ethics Committee (COMREC) to collect data from subjects in Mangochi, Malawi. Subjects' names were not used during analysis so as to uphold confidentiality. The parent trial is registered at *ClinicalTrials.gov identifier (NCTnumber):* ***NCT02047981***

CHAPTER 4
RESULTS AND DISCUSSION

This chapter presents and discusses the results that have been obtained from the study analysis. Section 4.1 presents the exploratory data analysis, section 4.2 presents the fitted models, and section 4.3 presents model assumption assessment.

4.1 Exploratory data analysis

4.1.1 Baseline characteristics

Table 1.1, gives a summary of the baseline characteristics of the children included in the Study. It presents the descriptive statistics of the variables to be used in the model estimation. This study consists of 22,492 children aged between of 0- 18 months and the median coverage was 0.56 years, from the pool of a community-randomized trial conducted in the Malawian district of Mangochi. The continuous variables, specifically, the mean age in years of the sample is 0.56 years and the standard deviation is 0.27 years, the median weight in kilograms (kgs) is 6.99 kgs and the IQR is between 5.98 kgs and 7.97 kgs, for the dose in milliliters (mls) the mean is 3.66 mls and the standard deviation is 0.78 mls. Out of 22,492 participants 11,441 were males while 11,051 were females. Under the categorical variables, the sample contains 50.9 % male children and 49.1 % female children. The children were randomly assigned to receive either azithromycin or placebo. However, after azithromycin and placebo was administered to the children, 22,214(98.7%) were alive and 278(1.2%) were dead. The outcome which is an occurrence of death, registered 1.2 % of death in the sample and 48 percent of the sample was administered the Azithro treatment drug.

Table 1. 1: Baseline Characteristics

Continuous Variable	
Age (Years) , mean (SD)	0.56 (0.27)
Survival Time (days), SD (IQR)	0.16 (1.66,1.66)
Weight (kgs), median (IQR)	6.99(5.98,7.97)
Dose (mls), mean (SD)	3.66(0.78)
Categorical Variables	**n (%)**
Sex: Male	11,441(50.9%)
Outcome: Dead	278(1.2%)
Drug:	
Placebo	11,677(52%)
Azithro	10,815(48%)
Treatment:	
	Treated
Phase	
-6 Months	0(0%)
0 Month	7,145(90%)
6 Months	5,335(88%)
12 Months	5,224(94%)
18 Months	5,400(92%)

The study was conducted in 5 phases where in the first phase (Mordor -6) was just for recruitment and no treatment was administered (baseline). The phases were separated in a space of 6 months and at each phase there were drop-outs, new recruits (new born) and some who died.

The highest number of those who were administered with drugs was at 12 months with 94% of the participants received treatment.

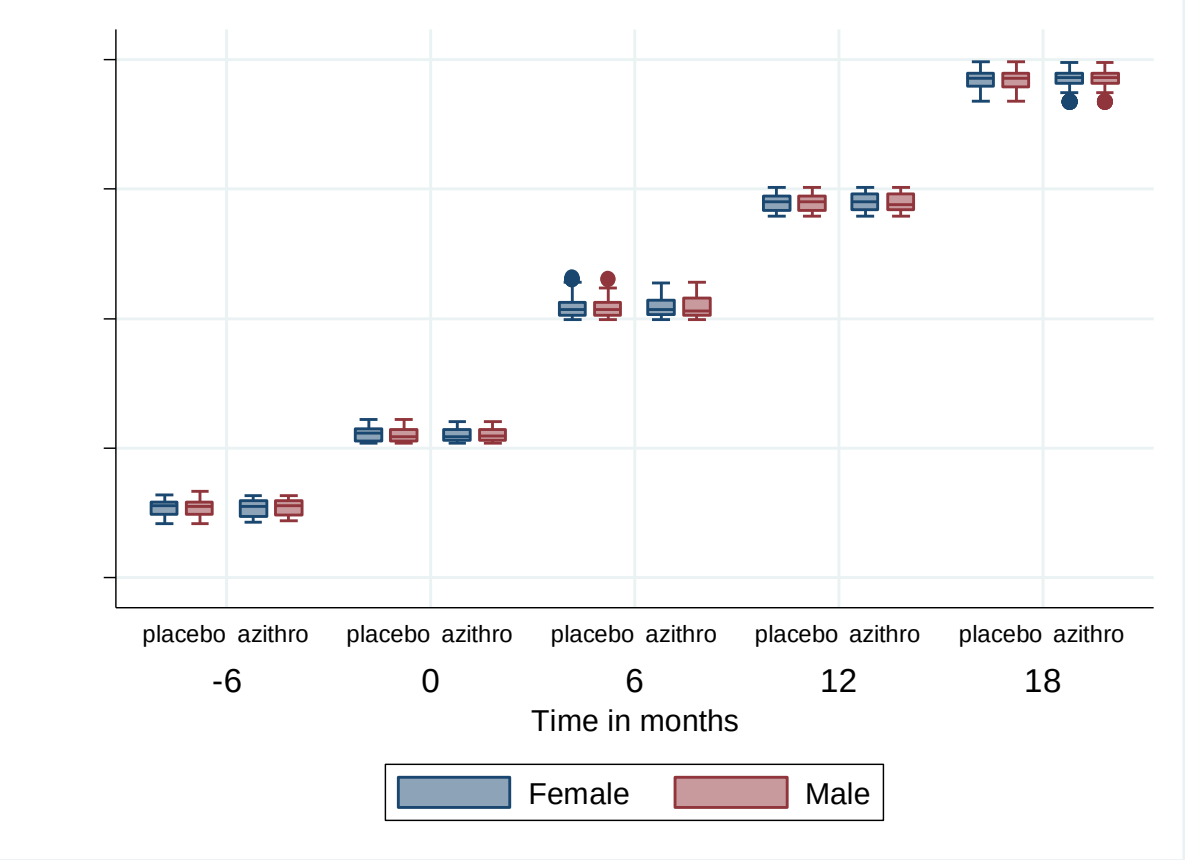

Figure 1. 1: Distribution of Survival time by Gender, Phase and Drug group

Figure 1.1 above shows how the median length is varying according to each survival time, for females with placebo and azithro, their median length stay is similar at -6 months, while males with placebo have higher median length of 20,314 days survival time than males with azithro who have the median length of 20,310 days survival time. The median length for both males and females is varying time outliers. It also shows that the distribution of survival time at -6 months was right skewed for males and females who have placebo and azithro. Figure 1.1 above shows the median length for the survival time at 0 months. Females with placebo and azithro, have similar median length of 20,417 days survival time, while males with placebo have higher median length of 20,424 days survival time than males with azithro who have the median length of 20,416 days survival time. The median length for both males and females is varying time outliers. It also shows the median length for the survival time at 6 months. Females with placebo

and azithro, have similar median length of 20,616 days survival time, while males with placebo have higher median length of 20,616 days survival time than males with azithro who have the median length of 20,615 days survival time. The median length for both males and females is varying time outliers. It also shows the median length for the survival time at 12 months. Females with placebo and azithro, have similar median length of 20,785 days survival time, while males with placebo have higher median length of 20,785 days survival time than males with azithro who have the median length of 20,775 days survival time. At 12 months the survival time is higher for placebo and for azithro. It shows the median length for the survival time at 18 months. Females with placebo and azithro, have similar median length of 20,965 days survival time, and males with placebo and azithro have similar median length of 20,965 days survival time. At 18 months the survival time is similar for both placebo and azithro.

4.2 Model Estimation Results

As already discussed in section 3.8 of the methodology chapter, the Extended Cox model (to account for time-varying covariates) and Cox-proportional hazard model were fitted separately. We then plotted Kaplan Meier curves to compare survival times between the 2 groups. Thereafter, the fitted models were compared to check if the estimates were different.

This section presents estimates for all the fitted models.

4.2.1 Kaplan- Meier survival estimates

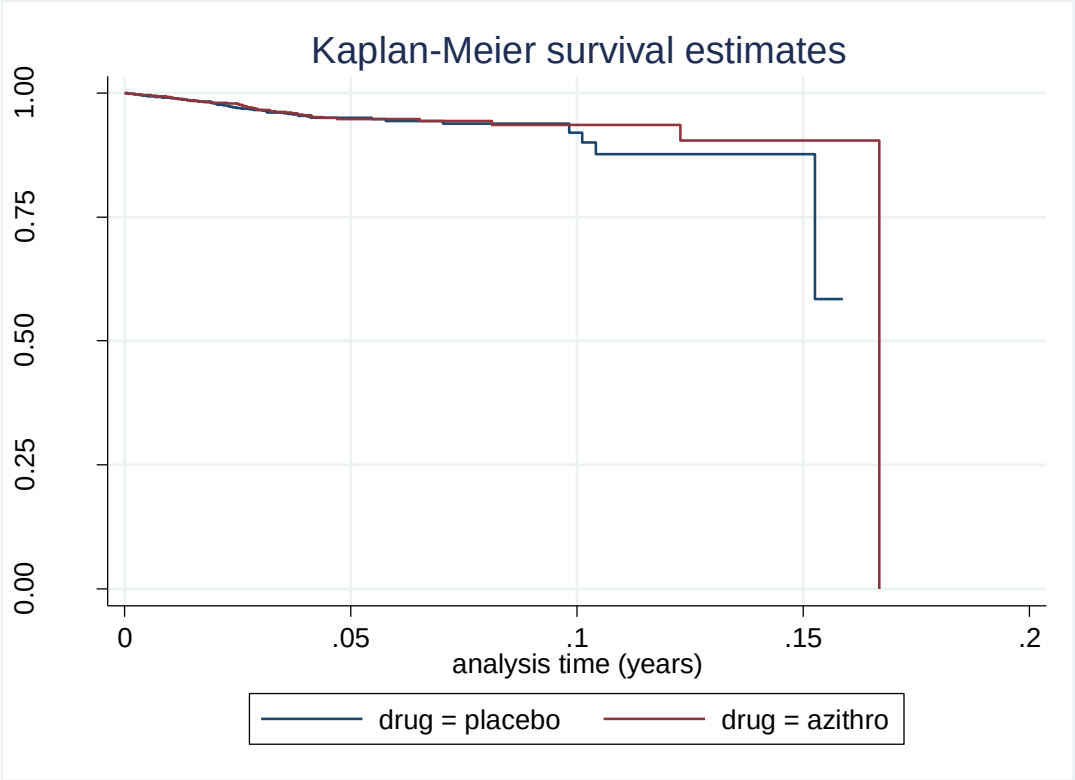

Figure 1. 2: Survival estimates in the 2 Drug groups

In Figure 1.2, The median survival is the smallest time at which the survival probability drops to 0.5 (50%) or below. If the survival curve does not drop to 0.5 or below then the median time cannot be computed. The median survival time and its 95% CI is calculated according to Brookmeyer & Crowley, 1982. Azithromycin drug group seems to be doing better than Placebo group with a median survival time 0.1though we fail to determine the median survival in placebo group as the survival curve does not drop to 0.5 or below. From the first year, the survival probability can hardly be compared between the two groups as the rate at which the survival probability drops is all the same. The Kaplan-Meier survival probability estimates at 12 months were both about 0.95 for both Azithromycin drug group and Placebo. The difference in the drop is more clearly soon after one year as the survival probability in the Placebo group drops more than in the azithromycin drug group.

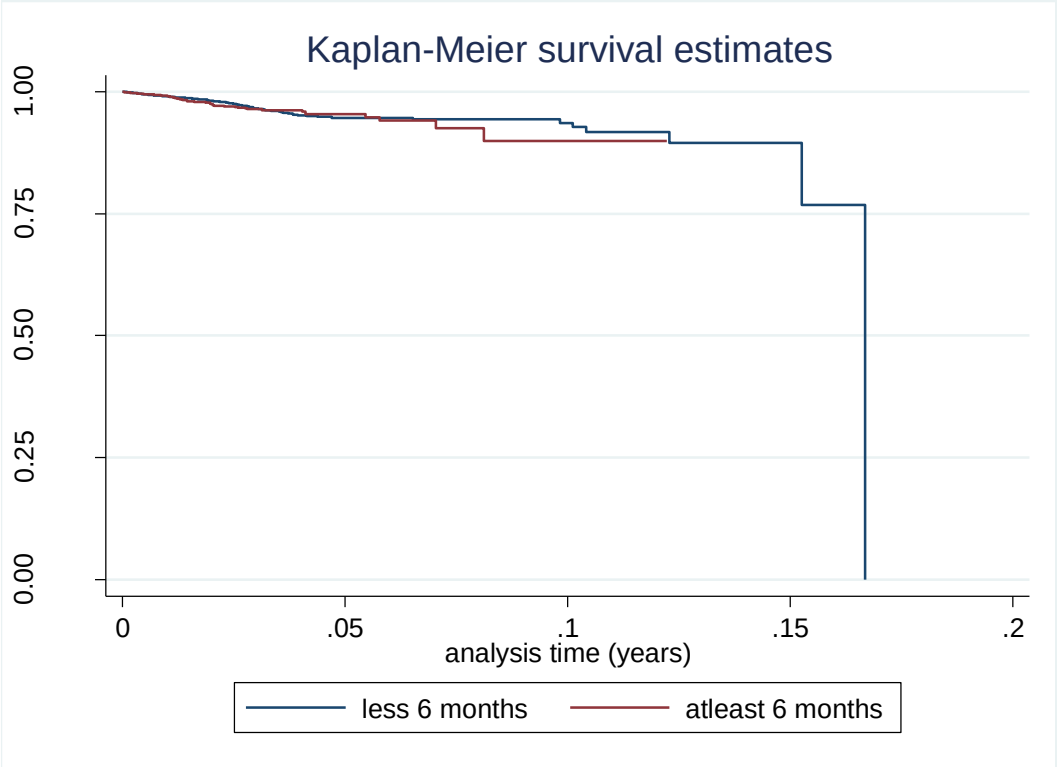

Figure 1. 3: Survival estimates in the age categories (less than 6 months old) and (atleast 6 months old)

From Figure 1.3 above, we can only compute median survival for those that are less than 6 months old as their graph drops to 0.5 and below. There are more deaths in age category 1 (those that are less than 6 months old) than in the 6 months plus age group. This shows that as you grow, survival is higher than in the infants.

4.2.2 Failure rates and rate ratios

Table 1. 2: Failure rates and rate ratios

Drug	**Deaths (D)**	**Person-in-time (years) (Y)**	**Rate (D/Y)**	**(95%CI)**
Placebo	152	0.14	1062.79	(906.6, 1245.9)
Azithro	126	0.13	933.47	(783.9, 1111.6)

The Table 1.2 above shows that the failure rate in the placebo group is higher than in the Azithromycin group. This is also illustrated by the graph in Figure 1.4 below

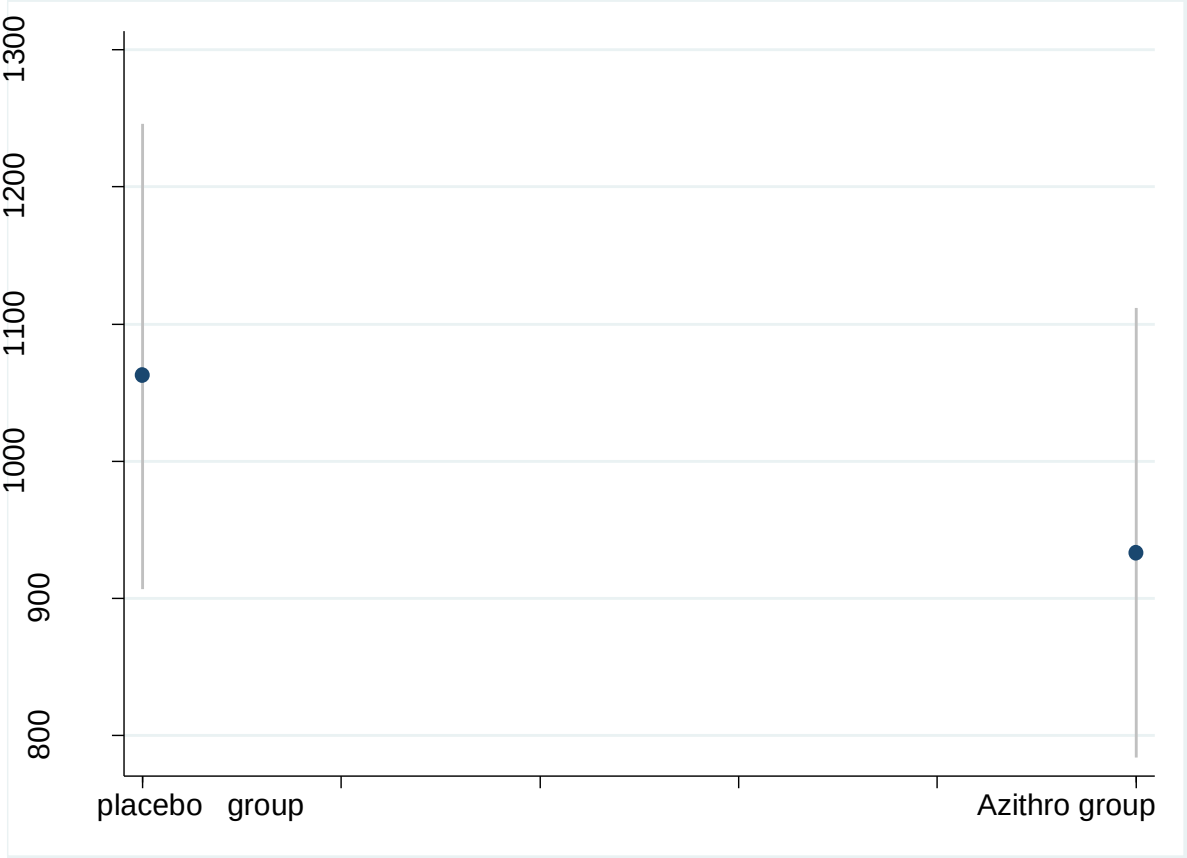

Figure 1. 4; failure rates across the 2 Drug groups

4.2.3 Logrank-test for equality of survival functions

The Table 1.3 presents the results of Logrank-test for equality of survival functions of the placebo and Azithro drug. As shown in the table, the hypothesis that survival functions are equal or the same cannot be rejected. This is evidenced by the p-value of 0.366, which indicates that there is no significant difference between survival functions of the two drug treatments.

Table 1. 3: Logrank-test for equality of survival functions in the two groups

Drug	**Events Observed**	**Events expected**
Placebo	152	144.5
Azithro	126	133.53
Total	278	278.00
	p-value = 0.82	p-value = 0.366

152 cases in Placebo group and 126 cases in the azithromycin group presented the outcome of interest. The Chi-squared statistic was 0.82 with associated P-value of greater than 0.05. The conclusion therefore is that, statistically, the two survival curves do not differ significantly, or that the grouping variable has no significant influence on survival time**.**

4.2.4 Fitting Cox Proportion hazard (PH) model

The table 1.4 below, indicates that the model fits well as it is significant at 1 percent level, this is shown by the p-value of 0.0001. From the table, only age and weight are significant at 1 percent level, where age is positively associated with occurrence of death and weight is negatively associated with occurrence of death. This implies that occurrence of death for individuals with more weight is unlikely and while for those that are aging or growing the occurrence of death is likely. The variable of interest (the drug treatment) and the other variables are however, not significant.

Table 1. 4: Unadjusted and adjusted Hazard ratios

variable	unadjusted HR (95% CI)	P-value	adjusted HR (95% CI)	P-value
age	1.07(0.66, 1.73)	0.798	15.03(5.17, 43.74)	<0.001
Treatment-received	0.95(0.73,1.2)	0.709	0.04(0.02, 11.94)	<0.001
Sex Male	1.01(0.79,1.28)	0.955	1.29(0.89, 1.86)	0.176
Dose	1.08(0.87,1.32)	0.495	1.68(0.78, 3.62)	0.182
Drug Azithro	0.87 (0.69 , 1.11)	0.269	1.08(0.75, 1.55)	0.666
Weight	0.87 (0.76 , 0.99)	0.039	0.56(0.37, 0.85)	0.001

The Table 1.4 above shows unadjusted and adjusted Hazard ratio summaries. All variables do not fit well in the model when specified without other variables. When the model was fitted with all variables in it, the 3 variables were found to be significant and were fitted separately in the model shown in Table 1.5 .

Table 1.5: Fitted Cox PH model

variable		Coef (95% CI)	P-value
Main	age	2.61(1.59, 3.63)	<0.001
	weight	-0.27 (-0.44 , -0.10)	0.001
Treatment-received		-3.33 (-4.18 , -2.48)	<0.001

The Table 1.5 indicates that the overall model is not significant, evidenced by the p-value of 0.9097. The model does not fit which means there is misspecification issue.

Note that there is a positive association between age and all-cause mortality and between weight and all-cause mortality (i.e., there is increased risk of death for older participants and for those with less weight). Again, these two parameter estimates represent the increase in the expected log of the relative hazard for each one unit increase in the predictor, holding other predictors constant. There is a 2.61 unit increase in the expected log of the relative hazard for each one-year

increase in age, holding other variables constant, and a -.27 unit decrease in expected log of the relative hazard for each gram decrease in weight, holding all variables constant.

For interpretability, we compute hazard ratios by exponentiating the parameter estimates. For age, exp(2.61(1.59, 3.63) = 13.56 (4.89, 37.64) HR. the expected hazard is 13.56 times higher in a person who is one year older than another, holding all other variables constant. Similarly, exp(-3.33 (-4.18 , -2.48) = 0.03 (0.02, 0.08). The expected hazard is 0.03 times lower for those who receive treatment compared to those who did not receive treatment, holding other variables constant.

4.2.5.Fitting extended Cox models

The Table 1.6 below indicates that the overall model fits well as it is significant at 5 percent evidenced by the p-value (0.000). The treatment received that factors in the time independent predictors is positively significant at 1 percent with a magnitude of 3.05. The treatment received for time-dependent factors is also positively significant at 1 percent with a lesser magnitude of 0.62. This basically, indicates the effectiveness of the treatment drug, that will diminish with respect to time.

Table 1. 6: Extended Cox model

variable	**Coef (95% CI)**	**p-value**
Main age	1.32(-0.63, 3.27)	0.185
Treatment-received	3.05 (1.92, 4.17)	<0.001
tvc age	0.22 (-0.17, 0.61)	0.274
Treatment- Received	0.62(0.39, 0.84)	<0.001

The Table 1.6 above shows parameter estimates from a fitted extended Cox model which takes into consideration the time varying covariates. The fitted model shows that it is significant with a p-value of <0.001 but the model shows different estimates from the same variable in this model. From the main model, there is a 1.32 unit increase in the expected log of the relative hazard for each one year increase in age, holding other variables constant, and a 3.05 unit increase in

expected log of the relative hazard for those that received treatment compared to that that did not receive, holding all variables constant.

From the tvc (time- varying covariate) model, there is a 0.22 unit increase in the expected log of the relative hazard for each one year increase in age, holding other variables constant, and a 0.62 unit increase in expected log of the relative hazard for those that received treatment compared to that that did not receive, holding all variables constant.

For interpretability, we compute hazard ratios by exponentiating the parameter estimates. For age, exp (1.32(-0.63, 3.27)) = 3.74 (0.53, 26.35) HR. The expected hazard is 3.74 times higher in a person who is one year older than another, holding all other variables constant. Similarly, exp (3.05 (1.92, 4.17)) = 21.01 (6.82, 64.72) HR. The expected hazard is 21.12 times higher for those who did not receive treatment compared to those who received treatment, holding other variables constant.

For age, exp (0.22 (-0.17, 0.61)) = 1.24 (0.84, 1.83) HR. the expected hazard is 1.24 times higher in a person who is one year older than another, holding all other variables constant. Similarly, exp (0.62(0.39, 0.84)) = 1.85 (1.48, 2.31) HR. The expected hazard is 1.85 times higher in the those who did not receive drug as compared to those who received, holding other variables constant.

Looking at the parameter estimates, the main model overestimates survival as compared to the tvc model which takes care of the time varying covariates.

4.3 Model assumption assessment and Goodness-Of-Fit

The section outlines model adequacy assessment results.

4.3.1 The Schoenfeld's global test

Table 1.8 below gives the results for the Schoenfeld's global test which assesses the assumption that the hazards in the Cox-proportional hazard model are proportional over time, i.e. testing whether effects of covariates on risk remain constant over time. Specifically, the test computes a test for each covariate i.e. testing the hypothesis of zero slopes in each of the covariates in the

model, along with a global test for the model as a whole. Thus, a non- zero slope is an indication of a violation of the proportional hazard assumption.

We observe that, at 95% confidence level (CI), all the covariates and the global test are not statistically significant (p-values > 0.05). Therefore, we fail to reject the hypothesis of zero slopes i.e. the assumption of proportional hazards is not violated. However, as with any regression model, it is recommended to look at the graphs of the regression in addition to performing the tests of non-zero slopes.

Table 1. 7: The Schoenfeld's global test

covariate	**rho**	**Chi-square**	**P-Value**
age	0.006	0.01	0.936
weight	0.089	1.66	0.196
1.treatment	-0.077	0.93	0.334
global test		2.83	0.418

Therefore, Figure 1.5 below presents the graphs for the scaled Schoenfeld residuals for each explanatory variable versus survival time. The solid line is a smoothing-spline fit to the plot, with the broken lines representing a +- 2-standard-error band around the fit. From the graphs, we also clearly observe that the fitted lines (slopes) for the scaled Schoenfeld residuals for each covariate are not significantly different from zero (i.e. no systematic departures from a horizontal line), that is confirming the test results obtained in the Schoenfeld global test.

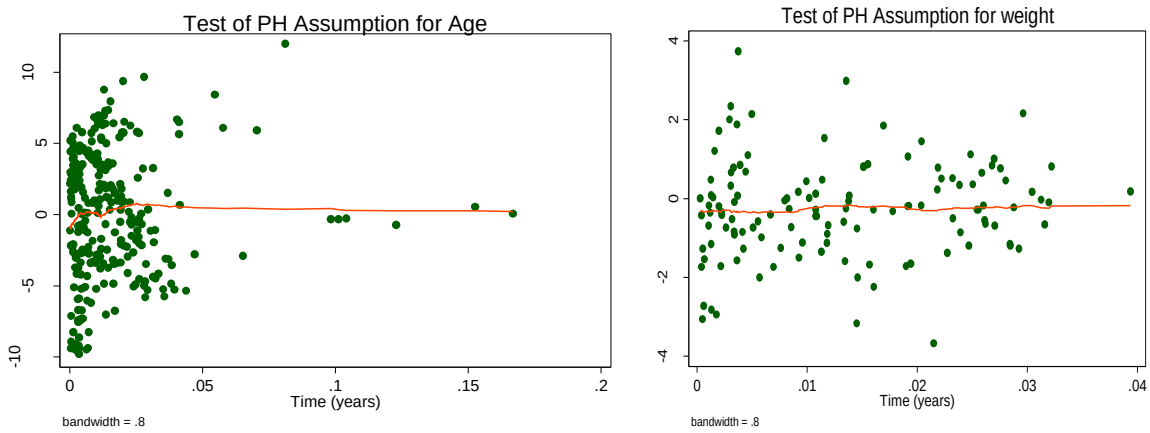

Figure 1. 5: Schoenfeld residual plots for each predictor for event death

4.3.2 Time-Varying covariates

In order to test if the Cox PH model satisfied the proportional hazard assumption, the Cox PH model was performed with age, treatment as time-varying covariates interacting with the analysis time. Table 1.8 presents results that were obtained after fitting the Cox PH and extended cox PH models for failure event Death.

Table 1. 8: Cox PH model and Extended Cox Model with time-varying covariates

Variable		HR(95% CI)	P-value
Cox PH model			
Main	age	3.55(1.37, 9.19)	0.009
	weight	0.72(0.61, 0.86)	<0.001
Treatment-received		0.37(0.17, 0.80)	0.011
Extended Cox Model with time-varying covariates			
Main	age	5.63(0.00, 7.51)	0.552
	weight	5.28(1.31, 2.11)	0.289
Treatment-received		3.02(1.72, 0.00)	0.447
tvc	age	28.45(-64.95, 121.84)	0.551
	weight	9.07(-7.75, 25.89)	0.291
Treatment- Received		-29.17(-14.05, 45.72)	0.445

The estimated hazard ratios are split into two categories in Stata, hazard ratios for variables with constant time and HR for time-varying covariates. From Table 1.8, it is observed that Age did not significantly interact with time ($p>0.05$), therefore a conclusion can be made that the PH assumption for the Fine and Gray regression is not violated**.** However, treatment variable significantly interacted with time ($p<0.05$) and we conclude that the PH assumption for the fine and Gray regression was violated. The same conclusion on the PH assumption can be made for the extended Cox model with death as the failure event, age, weight and treatment are not significant, thus failing to reject the null hypothesis of PH assumed. The PH assumption is not violated for this model ($p>0.05$).

4.3.3 Checking Linearity for Age

To check if the variable age is appropriate in a continuous form, the Martingale's residuals were plotted against age. Figure 1.6 presents the results for the analysis.

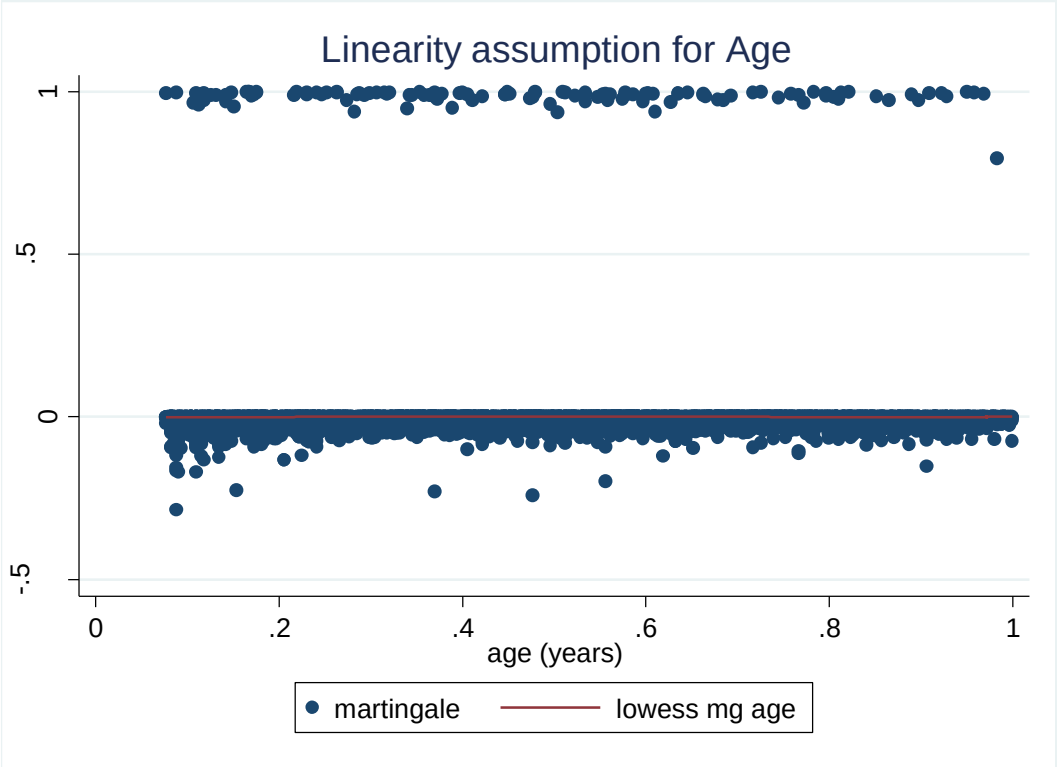

Figure 1. 6: Testing Linearity on variable age.

There was an approximate linearity in the functional form of the covariate age. This indicates the need to transform the covariate Age was not minimal. This shows that the log-hazard is slightly linear in age. Therefore, in addition to the violated PH assumption, results of age on the fitted Cox PH models were not acceptable too.

4.3.4 Goodness of Fit Test

Cox-Snell residuals were used to evaluate the fit of the model. If the model fits the data well then, the true cumulative hazard function conditional on the covariate vector has an exponential distribution with a hazard rate of one. First the Cox PH models were fitted for failure event death followed by the extended Cox model for the time-varying covariates. The Nelson-Aalen cumulative hazard functions were plotted to compare the hazard functions to the diagonal line. Goodness of fit was determined if the hazard function follows the 45 degrees line, implying that the cumulative hazard was approximately exponential with a hazard rate of one.

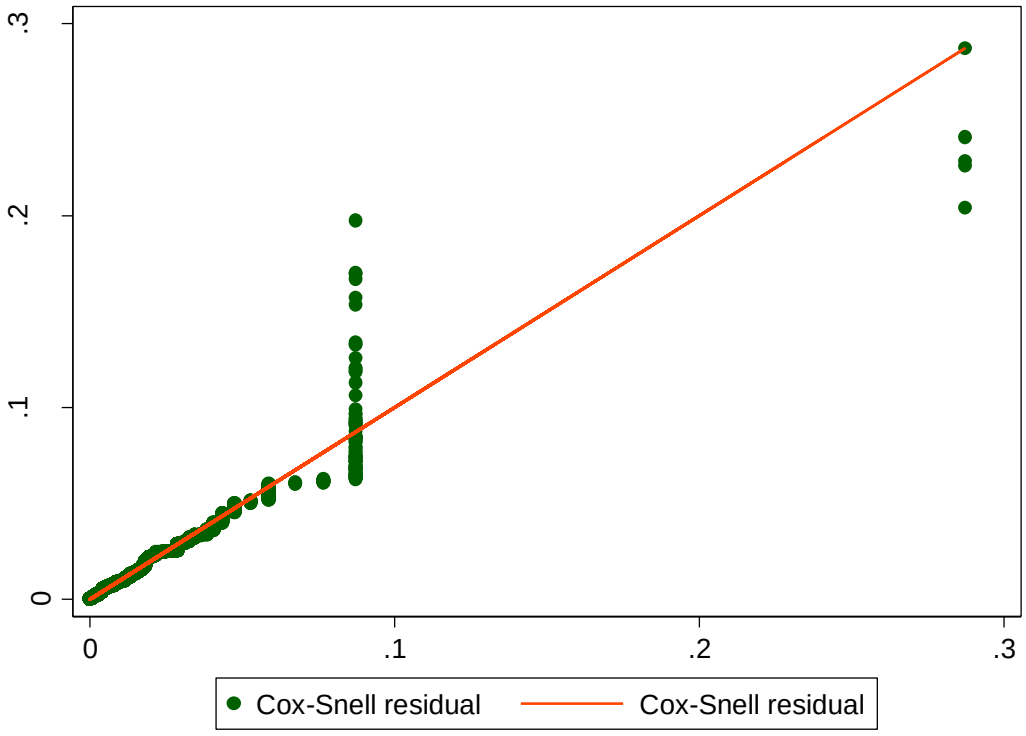

Figure 1. 7: Goodness of Fit for a Cox PH model

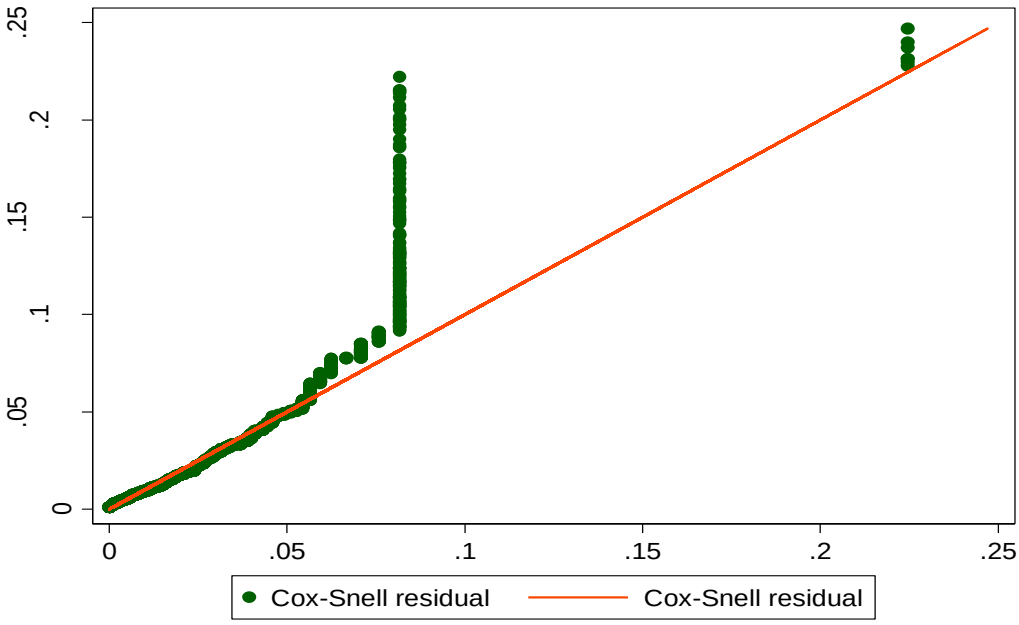

Figure 1. 8: Goodness of Fit for a Extended Cox PH model

Figure 1.7 and Figure 1.8 above shows that the hazard functions appear to follow the 45-degree line very closely at the beginning of the curve except for very large values of time especially for hazard event death. It is hard to fully conclude that the models fit the data well or that the models adequately fits or describes the data well**.**

In the preceding chapter, we have presented the results that have been obtained from the study analysis. In the following chapters, we will present the Discussion, conclusion, recommendations, and study limitations**.**

4.5 Discussion

In the MORDOR trial, the hazard for death in children aged 1–59 months in the time after distribution was significantly lower in communities randomized to azithromycin compared to placebo. This study also found that the hazards of death were lower in azithromycin group compared to placebo group through the logrank test. However, from the sample used for this thesis, the two survival curves do not differ significantly, or that the grouping variable has no significant influence on survival time**.**

The study took into account the presence of time varying covariates to estimate the effect of Age, weight, dose and receiving treatment on the hazard of death for participants in the controlled randomized trial (MORDOR). The Cox proportional hazard model overestimated the hazards of dying in the two study arms. The Cox proportional hazard model estimates treat covariates as being constant from baseline value. This leads to overestimation of hazards for the failure event. This is the case since the Cox proportional hazard model interprets the hazards of death without taking into account the presence of time varying covariates in the model. The extended Cox proportional hazard model must be used instead to estimate the hazards of survival when dealing with data that has time varying covariates. Several studies and authors by (Barnett et al., 2011) and (Ngwa et al., 2016) have pointed out that the extended Cox Model is an appropriate tool to use for estimation in the presence of time varying covariates.

In the study, we were able to determine the effect of the covariates on the hazard of death in the presence of time varying covariates by observing the estimates obtained from the both the Cox

proportional hazard model and extended Cox model. If the estimates are similar between the models, then we would say that the models are estimating the estimates well. The study showed that covariate estimates between the models Cox proportional hazard and extended Cox proportional hazard were slightly different with different confidence interval spans. Based on these results, it implies that time varying covariates affect estimation of the covariates on the event death. Therefore, it is not good to ignore the presence of time varying covariates in the models.

The results from this study agree with various authors who stated that the extended Cox model is a better model when studying data that involves time varying covariates [28]. Therefore, to have estimates that better explains the hazards in the presence of time varying covariates, the extended Cox model was a better model than the commonly used Cox proportional hazard model. In Table 1.6, Age, weight and receiving drug were identified with higher hazards in the Cox proportional hazard model and lower hazards in the extended Cox model (Table 1.7). This is similar to one study which showed that using models that do not account for time varying covariates overestimates the results [28]. Another interesting result on covariates affecting the probability of death taking into account the presence of time varying covariates was age which was significant only in Cox proportional hazard model and not in the extended Cox model. The expected hazard is 3.74 times higher in a person who is one year older than another in the Placebo group than in the Drug group, holding all other variables constant. Whereas the model that account for time varying covariates estimates states that the expected hazard is 1.24 times higher in a person who is one year older than another in the Placebo group than in the Drug group, holding all other variables constant. This is quite a high difference and would require the attention of medical researchers to take care of this issue when conducting analysis of such data to avoid overestimated conclusion.

CHAPTER 5
CONCLUSION, RECOMMENDATIONS AND LIMITATIONS

This chapter summarizes the study, outlines some recommendations for analyzing survival outcomes subject to time-varying covariates, and finally the limitations of the study.

5.1 Conclusions

A comparison of the extended Cox model and Cox proportional hazard model showed that cox proportional hazard model that do not take into account time varying covariates produced higher hazards of the failure event death unlike the extended cox model that takes into account time varying covariates. Since Cox proportional hazard model considers all covariates as fixed and calculates the probability estimates of death without taking into account the effect of the time varying covariates. It is therefore of importance to use the extended cox model to obtain the survivorship of an event of interest in the presence of time varying covariates.

The study showed that the influence of the time varying covariates on the cox proportional hazard model and on extended cox model of the event of interest gave different results. Age, weight, and treatment received were the only significant covariates in the Cox proportional hazard model and extended cox model. In all these models, age, weight, and treatment received had a significant effect on the probability estimates of event death. The estimates were high and overestimated in in the Cox proportion hazard model than in the extended cox model. The difference arises since the Cox proportional hazard model looks at the effect of covariates on the event of interest only without regards to how the covariates change over time as it treats them as fixed covariates. While this is the case for Cox proportional hazard model, the extended cox models take into account the effect of time varying covariates.

Ignoring the presence of time varying covariates in Cox regression model can lead us completely wrong results. Using a Cox regression model without ensuring that the underlying assumptions

are validated may result in negative implications on the estimates. If the assumption is violated, the extended Cox regression model is appropriate because it is more flexible to handle time dependent variables.

In our analysis, initially, the Cox regression was performed by considering that all explanatory variables are constant over time. Then, extended Cox regression models were estimated by including the time-dependent explanatory variables in the model as it was with the cox regression model. The fitted extended model results have shown that it become useful to estimate the Cox Proportional Hazards regression by also including the time-varying explanatory variables to the analysis. Both the time-independent and time-dependent variables create significant effects on the probability of survival of the time to death of the under-fives.

The study revealed that factors Age, weight and treatment-received of a participant were significant predictors in time to death for under-fives. Older participants and those with more weight were more likely to survive than the infants with less weight. The study showed a non-significant effect of gender on time to death of the under-fives. Thus, gender was considered as unimportant factor affecting time to death. More importantly our main goal was not to show the significant contributors on first death for infants in the two study arms but also how different models handle their different behavior over different time interval. We found that we had covariates that were time dependent and fitting Extended Cox regression model in such time dependent covariate situation performs better (fitted better) than traditional Cox regression model.

5.2 Recommendations

It is important to use models that handle time varying covariates in analyzing data that has time varying covariates present in the datasets. It is best to use the extended cox regression model than the traditional Cox PH model to estimate the survivorship function when modelling data with time-varying covariates. It will be statistically wrong to overlook the presence of time varying covariates in the model as we may end up with overestimated estimates.

This Extended Cox modelling approach of analyzing survival data that takes into account the presence of time varying covariates is therefore, valuable and effective because it incorporates all the available information in the data, and this suggests that the overestimation that might result from an analysis that ignores the presence of time varying covariates in the data is minimized.

5.3 Limitations

Firstly, the study sample size is very small and only include those two-year-olds and below as the primary study results were more significant in this age group, which insignificantly represents the population. This was due to a lot of inconsistencies in the dataset that the study used as most records had one follow up visit in the period of 5 rounds. It should be noted that this was secondary data and there was little that we could do with the data inconsistencies. That is, we discarded data that had dubious values which could not be verified. Therefore, only few participants had all the necessary information that the study required. As a result, the results of this study cannot significantly be inferred to the population from which the sample came from. For this reason, the results that we get in this study only signifies the greater competence that an extended Cox model might have in modelling survival data subject to non-ignorable time varying covariates.

Lastly, in the study, all participants with intermittent missing values were not considered in the study analysis. This was due to the study's focus on only time-to-death data. Therefore, the study proposes a further study that incorporates the intermittent missing values in its analysis. In this last chapter, we have summarized the study, and outlined some of the recommendations for analyzing survival outcomes in the presence of time varying covariates and put forward some of the limitations that the study faced.

REFERENCES

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716-723. https://doi.org/10.1109/TAC.1974.1100705

Allison, P. (2010). *Survival Analysis Using SAS: A Practical Guide.* Sas Inst.

Altman, D. (1991). *Practical Statistics for Medical Research.* Chapman and Hall.

Barnett, A. G., Beyersmann, J., Allignol, A., Rosenthal, V. D., Graves, N., & Wolkewitz, M. (2011). The time-dependent bias and its effect on extra length of stay due to nosocomial infection. *Value in Health*, *14*(2), 381–386. https://doi.org/10.1016/j.jval.2010.09.008

Bellara, C., MacGrogan, G., Debled, M., & Brouste, V. (2010). Variables with time-varying effects and the Cox model: Some statistical concepts illustrated with a prognostic factor study in breast cancer. *BMC Medical Research Methodology*.

Black, R. E., Cousens, S., Johnson, H. L., Lawn, J. E., Rudan, I., Bassani, D. G., Jha, P., Campbell, H., Walker, C. F., Cibulskis, R., Eisele, T., Liu, L., Mathers, C., & Child Health Epidemiology Reference Group of WHO and UNICEF (2010). Global, regional, and national causes of child mortality in 2008: a systematic analysis. *Lancet (London, England)*, *375*(9730), 1969–1987. https://doi.org/10.1016/S0140-6736(10)60549-1

Comparing of Cox model and parametric models in analysis of effective factors on event time of neuropathy in patients with type 2 diabetes. Available from: https://www.researchgate.net/publication/320731188_Comparing_of_Cox_model_and_parametric_models_in_analysis_of_effective_factors_on_event_time_of_neuropathy_in_patients_with_type_2_diabetes [accessed May 14 2019].

Collett, D. (2014). *Modelling survival data in Medical Research* (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b18041

Clark, T.G., Bradburn, M.J., Love S.B., Altman, D.G. (2003). Survival analysis part I: basic concepts and first analyses. *Br J Cancer*, 89, 232–8. [PMC free article] [PubMed] [Google Scholar]

Dekker, F. W., Mutsert, R. d., van Dijk, P. C., Zoccali, C., & Jager, K. J. (2016). Survival analysis: time-dependent effects and time-varying risk factors. *PubMed*, 994-997.

Ingabire, D., Mwalili,S.M., Orwa, G.O. & Extended Cox Modeling of Customer Retention in Mobile Telecommunication Sector of Rwanda.(2015). *American Journal of Theoretical and Applied Statistics,* 4(6), 471-479. doi: 10.11648/j.ajtas.20150406.17

Efron, B. (1988). Logistic regression, survival analysis, and the Kaplan–Meier curve. *Journal of the American Statistical Association,* 83, 414–425.

Emerson, P.M., Hooper, P.J. & Sarah, V. (2017). Progress and projections in the program to eliminate trachoma. *PLoS Negl Trop Dis* , 11, e0005402.

Etikan, I., Abubakar, S., & Alkassim, R. (2017). The Kaplan Meier estimate in survival analysis. *Biometrics & Biostatistics International Journal*, 55-59.

Fry, A. M., Jha, H. C., Lietman, T. M., Chaudhary, J. S., Bhatta, R. C., Elliott, J., Hyde, T., Schuchat, A., Gaynor, B., & Dowell, S. F. (2002). Adverse and beneficial secondary effects of mass treatment with azithromycin to eliminate blindness due to trachoma in Nepal. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, *35*(4), 395–402. https://doi.org/10.1086/341414

Giorgi, R., & Gouvernet, J. (2005). Analysis of time-dependent covariates in a regressive relative survival model. *Statistics in Medicine*, *24*(24), 3863–3870. https://doi.org/10.1002/sim.2400

Held, L., Sabanés, B. D., & Held, L. (2014). *Applied statistical inference: Likelihood and Bayes*. Springer.

Hosmer, D.W., Lemeshow, S. & May, S. (2008). Applied survival analysis: Regression modeling of time to- event data. Wiley.

In, J., & Lee, D. K. (2018). Survival analysis: Part I - analysis of time-to-event. *Korean journal of anesthesiology*, *71*(3), 182–191. https://doi.org/10.4097/kja.d.18.00067

Karabey, U., & Tutkun, N.A. (2017). Model selection criterion in survival analysis. AIP Conference Proceedings 1863, 120003 (2017) https://doi.org/10.1063/1.4992296

Kargarian- Marvasti, S., Rimaz, S., Abolghasemi, J., & Heydari, I. (2017). Comparing Cox model and parametric models in analysis of effective factors on event time of neuropathy in patients with type 2 diabetes. *Journal of Research in Medical Sciences*, 115.

Keenan, J. D., Ayele, B., Gebre, T., Zerihun, M., Zhou, Z., House, J. I., Gaynor, B. D., Porco, T. C., Emerson, P. M., & Lietman, T. M. (2011). Childhood mortality in a cohort treated with mass azithromycin for trachoma. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, *52*(7), 883–888. https://doi.org/10.1093/cid/cir069

Kleinbaum, D. G., & Mitchel, K. (1996). *Survival Analysis A Self-Learning Text.* New York.

Lalanne, C., & Mounir, M. (2016). *Biostatistics and computer bases analysis of health data using Stata*. Elsevier

Leung, K. M., Elashoff, R. M., & Afifi, A. A. (1997). Censoring issues in survival analysis. *Annual Review of Public Health*, *18*, 83–104. https://doi.org/10.1146/annurev.publhealth.18.1.83

Lopez, A.D., Mathers, C.D., Ezzati, M., Jamison, D.T. & Murray, C.J. (2006). Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. *Lancet,* 367, 1747-57.

Machin, D., Cheung, Y. B., & Parmar, M. K. (2006). *Survival Analysis- A practical Approach. John Wiley and Sons.*

McNeil, D.G. Jr. (16 July 2018). Now in sight: Success against an infection that blinds. *The New York Times*. Retrieved 3 September 2018.

Moonen, B., Cohen, J. M., Snow, R. W., Slutsker, L., Drakeley, C., Smith, D. L., Abeyasinghe, R. R., Rodriguez, M. H., Maharaj, R., Tanner, M., & Targett, G. (2010). Operational strategies to achieve and maintain malaria elimination. *Lancet (London, England)*, *376*(9752), 1592–1603. https://doi.org/10.1016/S0140-6736(10)61269-X

Ngwa, J. S., Cabral, H. J., Cheng, D. M., Pencina, M. J., Gagnon, D. R., LaValley, M. P., & Cupples, L. A. (2016). A comparison of time dependent Cox regression, pooled logistic regression and cross sectional pooling with simulations and an application to the Framingham Heart Study. *BMC Medical Research Methodology*, *16*(1), 1–12. https://doi.org/10.1186/s12874-016-0248-6

Porco, T.C., Gebre, T., Ayele, B, (2009). Effect of mass distribution of azithromycin for trachoma control on overall mortality in Ethiopian children: a randomized trial. *JAMA: the journal of the American Medical Association*, 302, 962-968.

Ponnuraj, C., & Venkatesan, P. (2010). Survival models for exploring tuberculosis clinical trial data-an empirical comparison. *Survival Models on Competing Risks Data*, 755-758.

Pourhoseingholi, M., Hajizadeh, E., Dehkordi, B., Safaee, A., Abadi, A., & Zali, M. (2007). Comparing cox regression and parametric models for survival of patients with gastric Carcinoma. *Asian Pacific Journal of Cancer Prevention*, 412-416.

Rajaratnam, J. K., Marcus, J. R., Flaxman, A. D., Wang, H., Levin-Rector, A., Dwyer, L., Costa, M., Lopez, A. D., & Murray, C. J. (2010). Neonatal, postneonatal, childhood, and under-5 mortality for 187 countries, 1970-2010: a systematic analysis of progress towards Millennium Development Goal 4. *Lancet (London, England)*, *375*(9730), 1988–2008. https://doi.org/10.1016/S0140-6736(10)60703-9

Sazawal, S., Black, R. E., & Pneumonia Case Management Trials Group (2003). Effect of pneumonia case management on mortality in neonates, infants, and preschool children: a meta-analysis of community-based trials. *The Lancet. Infectious Diseases*, *3*(9), 547–556. https://doi.org/10.1016/s1473-3099(03)00737-0

Shaikh, S., Schulze, K. J., Kurpad, A., Ali, H., Shamim, A. A., Mehra, S., Wu, L. S., Rashid, M., Labrique, A. B., Christian, P., & West, K. P. (2013). Development of bioelectrical impedance analysis-based equations for estimation of body composition in postpartum rural Bangladeshi women. *The British Journal of Nutrition*, *109*(4), 639–647. https://doi.org/10.1017/S0007114512001687

Stanley, C., Molyneux, E., & Mukaka, M. (2016). Comparison of performance of exponential, Cox proportional hazards, weibull and frailty survival models for analysis of small sample size data. *Journal of Medical Statistics and Informatics*, *4*(2).

Therneau, T.M. & Grambsch, P.M. (2000). *Statistics for Biology and Health*: *Modeling Survival Data; Extending the Cox Model.* Springer. https://doi.org/10.1007/978-1-4757-3294-8_3

Zhang, Z., Reinikainen, J., Adeleke, K., Pieterse, M., & Groothuis- Oudshoorn, C. (2018). Time-varying covariates and coefficients in Cox regression models. *National Library of Medicine*.

Appendix

STATA PROGRAMS USED TO ANALYSE THE DATA

```
*Programer: Alvin Chisambi
*Program: Biostatistics Masters Thesis- Survival data: MORDOR study
*Supervisor: Prof Mavuto Mukaka

use "C:\Users\Malawi3\Downloads\MSc thesis Alvin2019\thesis\lates Concept and Data v25022019\mordor_study_CLEANED.dta", clear

**************************************************
*keep unique values for some baseline characteristics
duplicates list masterhh_numeric
duplicates tag masterhh_numeric, generate(dup)
tab dup
drop if dup>0
ta id phase
************************************************************
*keep id, phase, weight, dose, age, treatment, died, drug, sex and studyarm
replace phase=1 if phase==-6
reshape wide treatment died age dose weight , i( id) j( phase)
table drug-sex
table drug
table sex
table studyArm
ta died1
ta died0
ta died6
ta died12
ta died18
```

reshape long treatment died age dose weight , i(id) j(phase)
****

***Reload the dataset**

encode Gender, gen(sex)
drop Gender
drop gender_numeric

***renaming participants ID**
rename masterperson id
rename Dose_bl dose
rename treatment_received_bl treatment
rename Weight weight
rename age_bl_yrs age
rename Rxlab group
rename MortalityOnly studyArm

replace weight=. if weight ==0

label define treatment 0"Did not receive" 1"Received"
label values treatment treatment

***Outcome/Failure variable (thats our survival status) 1=died 0=alive**
rename VitalStatus_numeric_fu died

***survival time (days) (Time variable)**
***Phase: -6=baseline census, 0=when analysis begin, 6=6 months, 12= 12 months, 18=18months**
***used days instead of months for phase**

rename Census_fu_incl_ltfu stime

***link the individuals using ID**
***This will increase person-years at risk and decrease the rate estimates,**
***as failures will remain the same. The rate ratios will be similar**

***Counting each observation as unique**
gen id = _n

***Declaring data to be survival data**
stset stime, id(id) failure(died==1) entry(time CensCap_numeric_bl) origin(time masterDOB_numeric_bl) scale(365.25)

***check through the database for additional varibales created**
***_st _d _t _t0**
sort id
list in 1/5
set more off

***cleaning for duplicates we will use masterpersonphase as data is in long format and id will appear multiple times**
duplicates list masterperson_phase

***rename variables**
rename Gender sex

***order variables**
order died Rxlab sex treatment_received_bl Dose_bl Weight stime MortalityOnly age_bl_yrs, before(masterperson_numeric)

***Identifying Time-Varying variables**

stdescribe

stsum

stvary

***ANALYSIS**

***1. Exploratory Data Analysis**

***1.1 baseline characteristics**

tabstat age weight dose stime , by(drug) stats(mean median SD IQR p25 p75) col(stat) long

summarize stime, detail

summarize age , detail

summarize weight , detail

summarize dose , detail

***categorical variable**

ta sex if phase==0

ta sex if phase==6

ta sex if phase==-6

ta sex if phase==12

ta sex if phase==18

table drug

table treatment

table sex

table drug treatment, by (sex)

table died, by (drug)c(freq)

table died, by (sex)c(freq)

table died drug , by (phase)c(freq)

set more off

***Outcome**

table died

***Box plots**

graph box stime , over(sex) over(drug) over(phase) asyvars graphregion(fcolor(white))

graph box stime if phase==-6, over(sex) over(drug) over(phase) asyvars graphregion(fcolor(white))

graph box stime if phase==0, over(sex) over(drug) over(phase) asyvars graphregion(fcolor(white))

graph box stime if phase==6, over(sex) over(drug) over(phase) asyvars graphregion(fcolor(white))

graph box stime if phase==12, over(sex) over(drug) over(phase) asyvars graphregion(fcolor(white))

graph box stime if phase==18, over(sex) over(drug) over(phase) asyvars graphregion(fcolor(white))

graph box stime , over(agecat) over(drug) over(phase) asyvars graphregion(fcolor(white))

***Graphs**

histogram dose, normal by(drug) by(phase) graphregion(fcolor(white))

histogram dose, normal by (drug) graphregion(fcolor(white))

histogram age, normal by (drug)

histogram weight, normal by (drug) plotregion(fcolor(white))

***Died before and after intervention**

ta died drug if phase ==-6

ta died drug if phase !=-6

ta died drug

***2. Objective 1 and 2**

***Kaplan- Meier survival estimates**

generate agecat=0

replace agecat=1 if age <0.6

replace agecat=2 if age >0.6

sts graph , by(drug) graphregion(fcolor(white))

sts graph, by(agecat) graphregion(fcolor(white))

***Failure rates and rate ratios**

strate drug, per (1000) graph graphregion(fcolor(white))

***life tables for survival**

***Logrank-test for equality of survival functions**

ltable stime died, survival by(drug) test

***Objective 3**

***3a. Fitting Cox PH model**

stcox i.drug age dose i.sex weight i.treatment, nohr

stcox age weight i.treatment, nohr

stcox i.treatment, nohr

stcox age, nohr

***3b.Fitting extended cox models**

stcox age i.treatment , tvc(age i.treatment) texp(ln(_t)) nohr

stcox i.treatment

***5.0 Test of proportional-hazards assumption**

***If p-value is greater than 0.05 then we are safe, we donot reject null hypothesis that the hazards are prorpotional**

estat phtest, detail

stphplot, by(drug) adjust(age)

***Calculating survival functions**

sts list, at(-6 0 6 12 18)

***comparing survival functions**

streg age drug, d(llog)

stcurve, survival ylabels(0 .5 1)

stcurve, hazard

stcurve, survival at1(drug=0) at2(drug=1) ylabels(0 .5 1)

stcox age drug

stcurve, survival

stcurve, survival at1(drug=0) at2(drug=1)

stcurve, hazard at1(drug=0) at2(drug=1) kernel(gauss) yscale(log)

***6 Model Diagnosis**

**********Test for PH assumption***********

```
estat phtest, detail
stphtest, plot(age)graphregion(fcolor(white))
stphtest, plot(i.drug)graphregion(fcolor(white))
stphtest, plot(sex) graphregion(fcolor(white))
stphtest, plot(dose)graphregion(fcolor(white))
stphtest, plot(weight)graphregion(fcolor(white))
stphtest, plot(i.treatment)graphregion(fcolor(white))
```

********************Goodness of Fit test*******************

```
*for cox PH*
stset stime, failure(died==1)
xi: stcox age weight i.treatment, mgale(mg)
predict coxsn, csnell
stset coxsn, failure(died==0)
sts generate H=na
twoway (scatter coxsn H) (line coxsn coxsn)
stset, clear
drop mg coxsn H

*for extended cox*
drop mg
drop coxsn
drop H
stset stime, failure(died==1)
streset, id(id)
stsplit, at(failures)
generate agetvc = age*(_t^2)
```

generate treatmenttvc = treatment*(_t^2)

stcox agetvc treatmenttvc, nohr mgale(mg)

***xi: stcox age i.treatment , tvc(age i.treatment) texp(ln(_t)) nohr mgale(mg)**

predict coxsn, csnell

stset coxsn, failure(died==1)

sts generate H=na

twoway (scatter coxsn H) (line coxsn coxsn)

stset, clear

**********Checking Linearity for Age*********

stset stime, fail(died==1)

xi: stcox age weight i.treatment, mgale(mg)

twoway (scatter mg age) (lowess mg age)

stset, clear